PROFILING – CONCEPTS AND APPLICATIONS

Von Der Wirtschaftswissenschaftlichen Fakultät

der Universität Leipzig

genehmigte

HABILITATIONSSCHRIFT

zur Erlangung des akademischen Grades
Doktor habil. nauk ekonomicznych

vorgelegt

von Dr. Agata Filipowska

geboren am 05.02.1980 in Koszalin (Polen)

Tag der Verleihung: 08.07.2020

Gutachter:

Prof. Dr. Rainer Alt

Prof. Dr.-Ing. Bogdan Franczyk

Prof. Dr. Sören Auer

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Profiling is an approach to put a label or a set of labels on a subject, taking into account the characteristics of this subject. The New Oxford American Dictionary defines profiling as: *recording and analysis of a person's psychological and behavioural characteristics, so as to assess or predict his/her capabilities in a certain sphere or to assist in identifying a particular subgroup of people* [151]. Such understanding of profiling, targeting at construction and application of artefacts (profiles) generated automatically or semi-automatically from collected data in order to describe users, is consistent within the domain of information systems. According to [30] a user profile is: *a set of information representing a user via user related rules, settings, needs, interests, behaviours and preferences* and profiling means creation of such description of a user.

This work extends this definition over a different type of subjects, namely things. It is demonstrated that techniques applied for profiling may be similar in both cases (people and things) as well as the number or application scenarios is greatly extended with such an approach. This assumption is in particular valid, when it comes to personalisation of behaviour of a thing to meet requirements of a user or while proposing a service on top of a thing and its behaviour e.g. turning off a thing in case of e.g. increased need for the electric energy.

However, while studying the body of knowledge, it is clear that both profile and profiling are complex notions, defined differently by diverse groups of researchers. The profiling may be understood as:

- collecting data describing a user, client, agent, thing, etc. e.g. [129],

- deriving stereotypes of a person (profiles generalised over a group of people and shared by members of a group) e.g. [25],
- assigning a person (a thing) to a group of people (things) with similar characteristics (clustering) e.g. [163],
- describing behavioural / psychological characteristics of a person e.g. [145],
- collecting history of previous transactions / actions of a user (a thing) e.g. [98].

Depending on the type of entities, profiling also focuses on their different characteristics. This set of characteristics may include an arbitrarily long list of features characterising certain aspects of a particular person or a thing. However, from the system's point of view usually a relatively narrow range of information is required, focusing on a domain that is addressed and the application scenario in which the profile is to be applied. Values of features included in the profile may be derived directly or indirectly and involve:

- attribute-based data (read from attributes of a person or a thing),
- declarative attributes (received from a person in a response to a question or manifested by a thing e.g. ship name included in an AIS message),
- attributes based on behaviour e.g. list of locations visited, home/work location of a user of telecommunication services,
- contextual data (adding additional layer of understanding over other attributes).

It is worth to note, that these characteristics differ also taking into account the cost of data acquisition. The scope of data processed while profiling a person or a thing should depend on the scenario and the outcome envisioned. Therefore, we may distinguish between three different types of profiles [94]:

- domain-knowledge profile targeting at description of an object for a particular domain,
- domain-independent profile that describes general information related to an object of profiling,
- cognitive model describing object's preferences and features.

Additional challenges for profiling emerge also from the recent General Data Protection Regulation (GDPR) [28], however as profiling in its substance should concern improvement of user experience (quality of service) this issue is not discussed in detail in this document. In this research it assumed that a user is willing to be profiled as he/she gets improved quality of service or service that was not previously available.

The research outcomes presented in the document concern:

- content of profiles – characteristics that should be taken into account while profiling people or things,
- profiling process – starting with selection of data (sources, records, attributes) and finishing with validation of methods delivered,
- usage of profiles – presented for chosen scenarios from domains such as social networks, telecommunications, authentication and commerce.

1.1.1 Definition of Profiling

This section extends the definition of the profile, showing insights into diverse understanding of a profile and the profiling process by various researchers and business practitioners.

Understanding a Profile: Selected Definitions

Table 1.1 presents different definitions of a profile that may be found in research papers from such domains as lifelong learning, information retrieval, Web development or robotics. Albeit different, all definitions focus on user features that should be identified to allow for implementation of functionalities improving user experience or providing a user with unique functionalities. This list could be further extended, but because of the page limit was compressed to demonstrate main directions which can lead to different understanding of this concept.

Table 1.1: Different understanding of a profile: selected definitions.

Definition	**Features of a profile**	**Source**
"A user model is a knowledge source in a natural-language dialogue system which contains explicit assumptions on all aspects of the user that may be relevant to the dialogue behaviour of the system."	The model of the user is an important basis for the intelligent dialogue behaviour with the system. The model should encompass such features as e.g. user goals, plans, background knowledge and (false) beliefs.	[167]
The paper does not propose an explicit definition of a profile, however it provides a metamodel of a user.	The metamodel includes such elements as: user, user ID, domain, user attributes (dependent or independent from components existing in the domain) and attribute values.	[94]
User modelling is defined as *"the process of acquiring knowledge about a user in order to provide services or information adapted to their specific requirements".*	A user is described by: specific goals, interactions with the system to realise these goals, characteristics such as general interests, cultural information and contextual information.	[159]
User profile includes *"information related to age, gender, skills, education, experience, and cultural level".*	Features of a user regarding his demographics and experience.	[88]
A user profile is *"a set of beliefs that the machine holds about the user".*	The profile includes not only a description of user features, but also enables to specify *"which parts of the model are stored on which devices, and which parts of the model should be shared with particular applications and people".*	[90]
"A user profile is a set of information representing a user via user related rules, settings, needs, interests, behaviours and preferences."	The user profile includes two types of data: static (e.g. country) and dynamic (e.g. needs). The content of a profile may vary depending on the application area.	[30]

The user profile may be diversely understood. Explicit user profile concerns: *”user static and predictable characteristics”*, whereas the implicit user profile describes a situation when a system learns a user by studying his behaviour. A user profile may be also hybrid: combining user data with user behaviour.	Features of the profile are mainly described using data from surveys. The data regarding behaviour needs to be processed depending on its type to describe user features.	[89]
Profiling a customer means describing him/her in a way that the most suitable products/services are marketed to the most appropriate individual given a stream of customer service consumption data generated in real-time.	The features of the profile relate to the preferences of a user towards a certain product/service. The profile should also include features enabling user segmentation.	[39]
User profile is understood as a way of defining user preferences that *”can be explicitly determined by getting users’ response/feedback to information/questionnaire”* or *”obtained by looking over the user’s shoulder”*.	A complete model of the user should include his/her cognitive state e.g. intentions behind the interaction or his/her internal state, and his/her preferences regarding social interaction characteristics.	[137]
Authors distinguish a short and a long-term user profile. A short-term user profile is built from recent interactions of a user with a system and is useful to predict the intents in the current task involving a user and the system. A long-term user profile describes long-standing user characteristics and is less sparse than a short-term profile.	The profile is built on top of data describing interactions of a user with the system, such as a user identifier, a query, a session identifier, query issued time, the top 20 URLs retrieved by the search engine, clicks and the dwelling time.	[47]

Source: own study

Profile, Persona and Stereotype

Recently, in the field of personalisation, new concepts have appeared. In addition to a profile, such notions as a stereotype and persona gain on attention. Therefore, it is worth to describe differences between these concepts.

The profile, as it was already mentioned, concerns individual description of a person that may change in time (along with acquiring new data). On the contrary, the stereotype:

- may be defined as characteristics suitable for more than one person, e.g. a working pensioner, describing features shared by more than one person;
- can be used to solve the problem of ”cold start” when we do not have attributes or behaviour of a certain person, but based e.g. on location or time of a day, this person may be assigned an initial profile derived from a stereotype,
- allows grouping people (clients, agents) and addressing their needs jointly (it is a classification expression),
- evolves slower than a profile, reflecting changes in the society to be uniform for a group of people.

On the other hand, in marketing and also in software design, personas seem to be a useful tool while designing a product or a service for a user. According to Cooper, Reimann and Dubberly [27]: *personas are a gathering of realistic representative information which can include fictitious details destined to a more accurate characterization.* Personas are:

- the most complete outline of the representative of the target group,
- description of the needs of the target group,
- description of the needs and priorities of a model client.

The purpose of creating personas is understanding the needs of a client by visualising the client. Examples of personas are presented in Figure 1.1.

Recently, we also refer to a user profile as to his personality [137]. Such an understanding came from psychology and sociology, and relates to studying user behaviour. According to [137] the personality *”(1) shows behaviours that are relatively pervasive in the person's lifestyle in that they show some consistency across situations; (2) shows behaviours that are relatively stable in the person's lifestyle across time, and (3) is indicative of the uniqueness of the person.”*

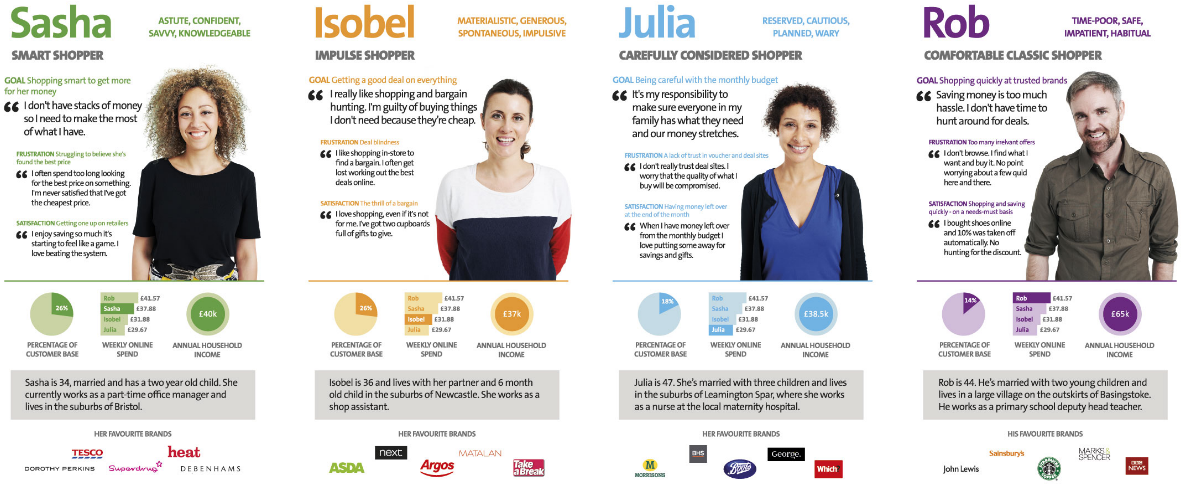

Sasha
ASTUTE, CONFIDENT, SAVVY, KNOWLEDGEABLE
SMART SHOPPER
GOAL Shopping smart to get more for her money
“ I don't have stacks of money so I need to make the most of what I have.
FRUSTRATION Struggling to believe she's found the best price
“ I often spend too long looking for the best price on something. I'm never satisfied that I've got the cheapest price.
SATISFACTION Getting one up on retailers
“ I enjoy saving so much it's starting to feel like a game. I love beating the system.
26%
PERCENTAGE OF CUSTOMER BASE
Rob £41.57
Sasha £37.88
Isobel £31.88
Julia £29.67
WEEKLY ONLINE SPEND
£40k
ANNUAL HOUSEHOLD INCOME
Sasha is 34, married and has a two year old child. She currently works as a part-time office manager and lives in the suburbs of Bristol.
HER FAVOURITE BRANDS
TESCO
heat
DOROTHY PERKINS
Superdrug
DEBENHAMS
Isobel
MATERIALISTIC, GENEROUS, SPONTANEOUS, IMPULSIVE
IMPULSE SHOPPER
GOAL Getting a good deal on everything
“ I really like shopping and bargain hunting. I'm guilty of buying things I don't need because they're cheap.
FRUSTRATION Deal blindness
“ I like shopping in-store to find a bargain. I often get lost working out the best deals online.
SATISFACTION The thrill of a bargain
“ I love shopping, even if it's not for me. I've got two cupboards full of gifts to give.
26%
PERCENTAGE OF CUSTOMER BASE
Rob £41.57
Sasha £37.88
Isobel £31.88
Julia £29.67
WEEKLY ONLINE SPEND
£37k
ANNUAL HOUSEHOLD INCOME
Isobel is 36 and lives with her partner and 6 month old child in the suburbs of Newcastle. She works as a shop assistant.
HER FAVOURITE BRANDS
next
ASDA
Argos
MATALAN
Take a Break
Julia
RESERVED, CAUTIOUS, PLANNED, WARY
CAREFULLY CONSIDERED SHOPPER
GOAL Being careful with the monthly budget
“ It's my responsibility to make sure everyone in my family has what they need and our money stretches.
FRUSTRATION A lack of trust in voucher and deal sites
“ I don't really trust deal sites. I worry that the quality of what I buy will be compromised.
SATISFACTION Having money left over at the end of the month
“ When I have money left over from the monthly budget I love putting some away for savings and gifts.
18%
PERCENTAGE OF CUSTOMER BASE
Rob £41.57
Sasha £37.88
Isobel £31.88
Julia £29.67
WEEKLY ONLINE SPEND
£38.5k
ANNUAL HOUSEHOLD INCOME
Julia is 47. She's married with three children and lives in the suburbs of Leamington Spar, where she works as a nurse at the local maternity hospital.
HER FAVOURITE BRANDS
BHS
M MORRISONS
Boots
George.
Which?
Rob
TIME-POOR, SAFE, IMPATIENT, HABITUAL
COMFORTABLE CLASSIC SHOPPER
GOAL Shopping quickly at trusted brands
“ Saving money is too much hassle. I don't have time to hunt around for deals.
FRUSTRATION Too many irrelvant offers
“ I don't browse. I find what I want and buy it. No point worrying about a few quid here and there.
SATISFACTION Shopping and saving quickly - on a needs-must basis
“ I bought shoes online and 10% was taken off automatically. No hunting for the discount.
14%
PERCENTAGE OF CUSTOMER BASE
Rob £41.57
Sasha £37.88
Isobel £31.88
Julia £29.67
WEEKLY ONLINE SPEND
£65k
ANNUAL HOUSEHOLD INCOME
Rob is 44. He's married with two young children and lives in a large village on the outskirts of Basingstoke. He works as a primary school deputy head teacher.
HIS FAVOURITE BRANDS
Sainsbury's
MARKS & SPENCER
John Lewis
STARBUCKS COFFEE
BBC NEWS

Figure 1.1: Examples of personas. Source: [99]

Profiling

The process of development of a user profile is called profiling. A profile may be constructed in two different ways: based on studying user features (bottom-up) and based on specification of stereotypes (top-down). First, feature-based approach concerns a situation where we build a profile for each user separately. These profiles may evolve in time and therefore reflect changes concerning a user [22].

Profiling based on stereotypes is carried out in two steps. Firstly, generation of stereotypes that describe typical users (groups that may be identified based on similar features) is performed [90]. In the second step a user is assigned with the best matching stereotype. Such profiles include typical hypotheses for all community members. For example, we may assume that all employees of a given organisation will have similar information needs or customers interested in a particular group of products would like to be notified on all actions concerning these products [1, 110]. Such stereotypes may be constructed by qualified designers or by linking (aggregating) profiles already existing in the system.

1.1.2 Representation of the Profile: Data Structures

While analysing a profile from a technical perspective, we translate characteristics of a user or a thing into a set of attributes, relations between these attributes, weights concerning attributes, their parts or relations, etc. We may divide these elements into the following categories:

- attributes (parameters) that may be constant, not depending on time, or variable (changing in time or on event-based mode),
- relations between attributes being e.g. rules connecting attributes together under a condition or indicating proximity of attributes,
- classes (types) of a profile emerging e.g. from stereotypes.

The data structures used to represent a profile differ taking into account not only the amount of information that may be stored to describe a user or a thing, but also w.r.t. reasoning possibilities (if there is no connection between diverse attributes within a profile, no reasoning is possible) and impact on the computational complexity of methods used to process the data. Therefore, the data structures that are used to describe a profile may be divided into the following groups:

- models encompassing a set of independent features describing a user or a thing,
- models describing relations between different attributes of a profile,
- models in which we indicate classes that define various elements of a user profile.

Table 1.2 presents details for diverse data structures that may be used while profiling a person or a thing.

Table 1.2: Data structures for representation of a profile.

Data structure	**Features of the data structure**	**Examples**
Bag of words	A set of words describing a profile e.g. indicating user information needs or interests. Words may come e.g. from content read by a user or may be provided explicitly by him in a survey. In case of thing, they may be provided by a producer or learned automatically while functioning of a thing.	[46, 161]
Vector of values	Each element of the vector shows one dimension of a profile (otherwise referred to as feature). Values within the vector may include: weights, numbers, words, etc. Vectors are easy to use as comparison of two profiles involves application e.g. of cosinus or Euclidean distance measures.	[30, 35, 139]

Collection of attribute-value pairs	Templates for profiles include attributes that need to be filled in with values while instantiating a profile. Such approach is used when a finite set of attributes may be defined.	Description of books, things, etc.
Semantic network	Extension over a vector model, indicating relations between attributes in the profile. Represented as a graph of different notions where nodes indicate terms and edges relations between these terms. Both edges and nodes may be annotated with weights describing importance of these elements. Edges may have different types indicating different types of relations e.g. hyponym/hypernym, antonym, etc. Such structure enables reasoning over the profile.	[46], WordNet
Ontology	Data structure similar to the semantic network, however more complex as it includes not only notions and relations, but introduces new types of elements, such as: classes (general entities), instances (exemplification of these classes), relations between entities, attributes and functions, processes, axioms concerning entities, etc. [31].	[148]
Bayesian network	Approach based on indication of probability. The profile is represented as an acyclic directed graph, in which nodes represent variables and edges indicate relations between these variables. Each variable in the model has its state or a set of states with indicated probabilities. If two nodes are not connected with an edge, these variables should be treated as independent, if connected - there is an influence of one variable on the other (correlation).	[76, 116, 155]

Rule-based profile	Relations between various terms in a profile are represented in a form of rules, that concern e.g. user behaviour. The rules are derived after an analysis of large amounts of data, based on which also support of these rules within the dataset is calculated.	[126]
Neural Networks	Often used while automatically learning user behaviour. Nodes within the neural network may represent terms important for a user and edges representing strength of associations between these terms. Though, the structure of the neural network depends on the initial data set and the classification scenario.	[30]

Source: own study

1.2 Research Goal

Nowadays, even in the world after General Data Protection Regulation [28], profiling is a point of interest of many different entities, as personalisation of content, products, etc. is a way of increasing sales or the customers' base and thus the company value. It may be easily noticed that there are no significant limitations regarding profiling (from the perspective of a concept, however not mentioning the computational complexity) and therefore while designing personalisation or recommendation methods, one should not assume methods or models, but rather user requirements for targeted application scenarios. It should be also explicitly written, that the problem of profiling in general is not new, but when studying diverse application scenarios, one may observe a number of research challenges that should be addressed.

The focus of this research concerns proposing methods for discovery of profiles of users and things with application of Data Science methods. The profiles are utilised in vertical and horizontal scenarios, and concern such domains as smart grid and telecommunication (vertical scenarios), and support provided both for authorisation and personalisation (horizontal perspective). Therefore, the research questions addressed are as follows:

1. What is a profile of a user or a thing? How to define a profile? How a profile of a person or a thing may be represented to enable for diverse application scenarios?

2. How to model a profile? How to derive a profile from data provided by a user? How to describe relations between entities?

3. How to profile a user or a thing for the needs of solutions supporting management of electric energy production and consumption in the smart grid?

4. How to profile a user for the needs of telecommunication services, especially taking into account the issues of privacy and trust?

5. Is it possible to use a user profile for authentication in various services? How detailed the profile should be to enable for safe authentication? What level of granularity of data is needed to describe a user for the needs of authentication?

6. How to analyse a user and create a profile that may be used for personalisation? How to personalise taking into account user experience and computational complexity?

To address these research questions six main goals were defined, each addressing one of the aforementioned questions. Table 1.3 presents the main goals addressed in the thesis, divided into three perspectives: background and definition, vertical and horizontal applications. Each of goals is addressed by one of sections in the document (starting from Section 2).

Table 1.3: Main goals addressed in the thesis.

Perspective	**Goals**
Background and definition	The goal is to define a profile of a person or a thing and identify features of a person or a thing that may be represented in a profile and are usable for diverse application scenarios.
	The goal is to analyse profiling methods that enable for describing a user/a thing or relations between users.
Vertical applications	The goal is to create a profile of a user or a thing that will be applicable for solutions enabling management of production and consumption of the electric energy in the smart grid.
	The goal is to develop a profile of a customer/subscriber to telecommunication services, enabling for personalisation and taking into account issues of privacy and trust.

Horizontal applications	The goal is to develop a method enabling for authentication of a user based on Call Detail Record data.
	The goal is to propose profiling methods that enable to describe a user in a way supporting personalisation of content or a service.

Source: own study

To address achieving the main goals, each of them is accompanied by at least two detailed objectives, as presented in Table 1.4.

Table 1.4: Supplementary goals to be addressed in the thesis.

Goal	**Supplementary goals**
The goal is to define a profile of a person or a thing and identify features of a person or a thing that may be represented in the profile and are usable for diverse application scenarios.	The goal is to analyse different approaches for describing a profile or an identity of a user or a thing that are applied in different classes of systems e.g. identity management systems.
	The goal is to analyse how to instantiate a profile of a person or a thing using semantic approaches enabling for diverse application scenarios.
The goal is to analyse profiling methods that enable for describing a user / a thing or relations between users.	The goal is to describe a user profile w.r.t. user personality and his/her colour preferences. The supplementary goal is to study relations between personality traits and user colour preferences using different methods of analysis.
	The goal is to create a method for describing relations between users focusing on quantitative and qualitative aspects of a relation on the example of a social network.

The goal is to create a profile of a user or a thing that will be applicable for solutions enabling management of production and consumption of the electric energy in the smart grid.	The goal is to propose an architecture of a system for monitoring energy production and consumption in the smart grid, taking into account a profile of an individual prosumer.
	The goal is to create a profile of a user for the needs of electric energy supply: monitoring and describing demand for the electric energy to be used by the system enabling management of the production and consumption of the electric energy in the smart grid.
The goal is to develop a profile of a customer/subscriber to telecommunication services, enabling for personalisation and taking into account issues of privacy and trust.	The goal is to define a solution that enables to manage personal information in telecommunication.
	The goal is to propose methods enabling for user profiling based on Call Detail Records data.
The goal is to develop a method enabling for authentication of a user based on Call Detail Record data.	The goal is to verify if the Call Detail Record data is sufficient for detecting anomalies in the behavioural user profile and therefore enable for applying the CDR-based profile in authentication scenarios.
	The goal is to research how much data describing a user is needed to provide an efficient authentication solution.

	The goal is develop and validate a method for description of a user for the needs of authentication. The supplementary goal is to develop a methodology for testing the behaviour-based approaches based on Call Detail Records data.
The goal is to propose profiling methods that enable to describe a user in a way supporting personalisation of content or service.	The goal is to propose a personalisation method that improves the user experience of a solution, but does not impact the efficiency of the solution.
	The goal is to provide a method for user profiling that allows obtaining valuable insights from data that was not collected for a specific purpose.

Source: own study

1.3 Research Methodology

1.3.1 Introduction

The research at the intersection of information systems and computer science, described within this document, in majority methodologically follows the design-oriented information systems research as understood by [57, 64, 120, 165]. However, assigning this research entirely to one research paradigm, is challenging. This is due to the fact, that profiling of people, things, processes, etc. aligns well with both research approaches known for the domain of information systems, namely behaviouristic and the design-oriented paradigm. The behavioural science approach focuses on observing behaviour of entities, and then focusing on describing, explaining and predicting their behaviour [64]. This approach is especially valid while analysing Call Detail Records to create mobility profiles that may be used for various purposes e.g. authentication or marketing.

On the other hand, design science concerns problem solving. It targets innovations regarding ideas, practices, technical aspects through which information systems may be effectively accomplished [64]. Profiling in its essence is about designing new approaches, innovating methods, proposing new application scenarios, and therefore aligns well also with the design-science paradigm. For example, in case of profiling for the smart grid to support management of production and consumption of electric energy, the goal is to define a problem and propose a solution that enables to target this problem.

On top of these two approaches, one may distinguish also the consortium research [123]. The information systems research includes advances in the community of practitioners and therefore the consortium research addresses the issue of getting access and exchanging knowledge with also potential users or entities acting in the domain (including application of research results achieved). Consortium research supports development of artefacts in collaboration between the university and its partners in all stages of the design-oriented research process. The goal is to deliver results of practical relevance, what in case of this thesis concerned e.g. research in the field of telecommunication carried out in collaboration in Orange[1].

Summarising, the methodology applied for specific parts of this work is diverse, sometimes focusing on explanation of behaviour e.g. Section 7.3, but mainly providing new artefacts and solutions as proposed by the design-oriented research. In some cases, the research was performed in collaboration with practitioners and the results are of practical relevance, so the consortium research guidelines were followed. [63] also underline that different methodologies impact one another as technology and behaviour are not dichotomous in an information system. The detailed research methodology followed within specific parts of this work is briefly described within appropriate sections (when discussing research results).

The detailed research guidelines applied by the author are described below, following points suggested by [120, 123].

1.3.2 Research Guidelines

While developing methods and artefacts related to the field of profiling, a number of challenges needs to be taken into account that influence results and their quality. These challenges include inter alia [63]:

- unstable requirements and constraints based on a context of a solution,
- complex interactions that need to be modelled to apply the solution,
- research process that is influenced by participants, outcomes and context,

[1] https://www.orange.com/en/home

- dependence on human abilities to produce effective solutions.

Most of these challenges relate to the fact that information systems research involves people in all steps of the process. The requirements come from people, they influence the context of the solution, research process involves humans, who can also hinder communication of results. To address these challenges, [64, 120] propose to follow a methodology that enables to deal with these issues. The research guidelines described below address the challenges mentioned. They include [64, 120]:

- delivering results developed within the scientific rigour,
- collaborating with practitioners,
- developing results (artefacts) that are applicable to important and relevant business problems and therefore contributing to practice,
- each artefact must be justified and validation needs to be possible, in particular utility, quality and efficacy of an artefact must be rigorously demonstrated by application of well-executed evaluation methods,
- having impact on both: research and business practice with communication of results effective to both technology-oriented and management-oriented audiences.

As it was previously mentioned, most of this work follow the Design Science Research Cycle presented in Figure 1.2. Figure demonstrates three different research cycles included in the research process. The relevance cycle is to bridge the context of the artefact with the design science activities. The context means requirements as well as acceptance criteria (enabling validation) to evaluate the results achieved. The rigour cycle relates the research activities with the knowledge base describing scientific foundations, experience and expertise. Knowledge means here not only existing artefacts and processes that may be found in the application domain, but also experience and expertise that define state of the art in the application domain of research. The design cycle, being the core of every research project, iterates between building of artefacts and evaluating them. To properly execute a research project, these three cycles need to take place.

Figure 1.3 extends the idea depicted in Figure 1.2 and presents details of the consortium research built on top of the design science research guidelines. In this approach, four main research phases were distinguished, namely: analysis, design, evaluation and diffusion. The central part

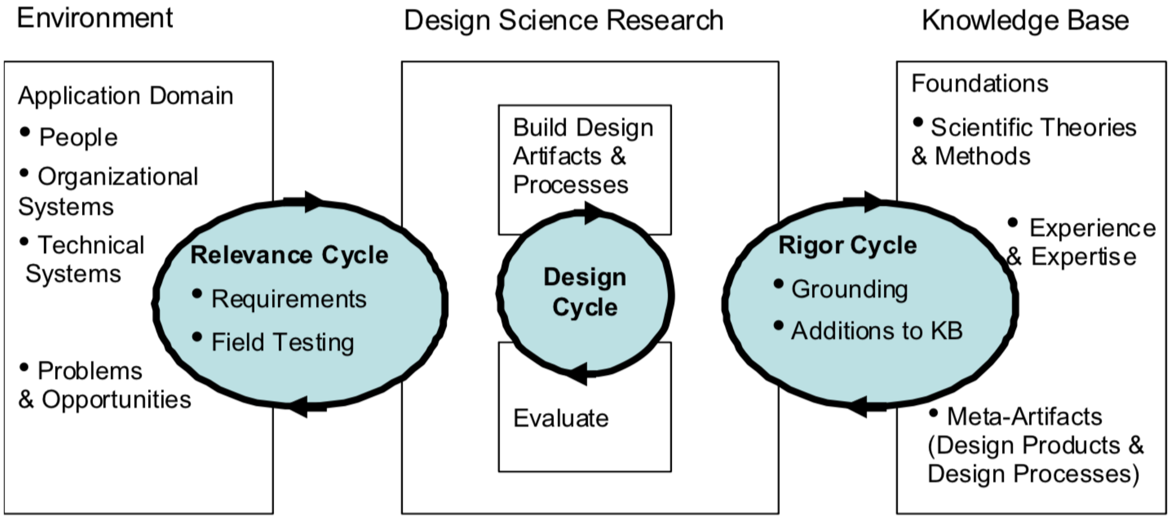


Figure 1.2: Design Science Research Cycle. Source: [65]

of the picture shows a diverse nature of the domain, which is the source of a research problem as well as the place where results are applied. The conditions defined for the consortium research emphasize collaboration between researchers and practitioners, highlighting knowledge exchange in both directions.

1.3.3 Research Contribution

An important part of every research methodology are research goals i.e. results to be achieved. [63, 64] distinguish three types of contributions of a research project including: artefact (solution to a problem), foundations (modelling formalisms, ontologies, problem and solutions representations, design algorithms, innovative information systems) and methodologies e.g. new evaluation methods or experiences gained from performing the iterative design cycles and field testing the artefacts in the application environment. Taking product-related perspective into account, these may be summarised as [64]:

- theories seeking to predict or explain phenomena that occur while using an artefact,
- constructs (vocabulary and symbols),
- models (abstractions and representations),
- methods (algorithms and practices),

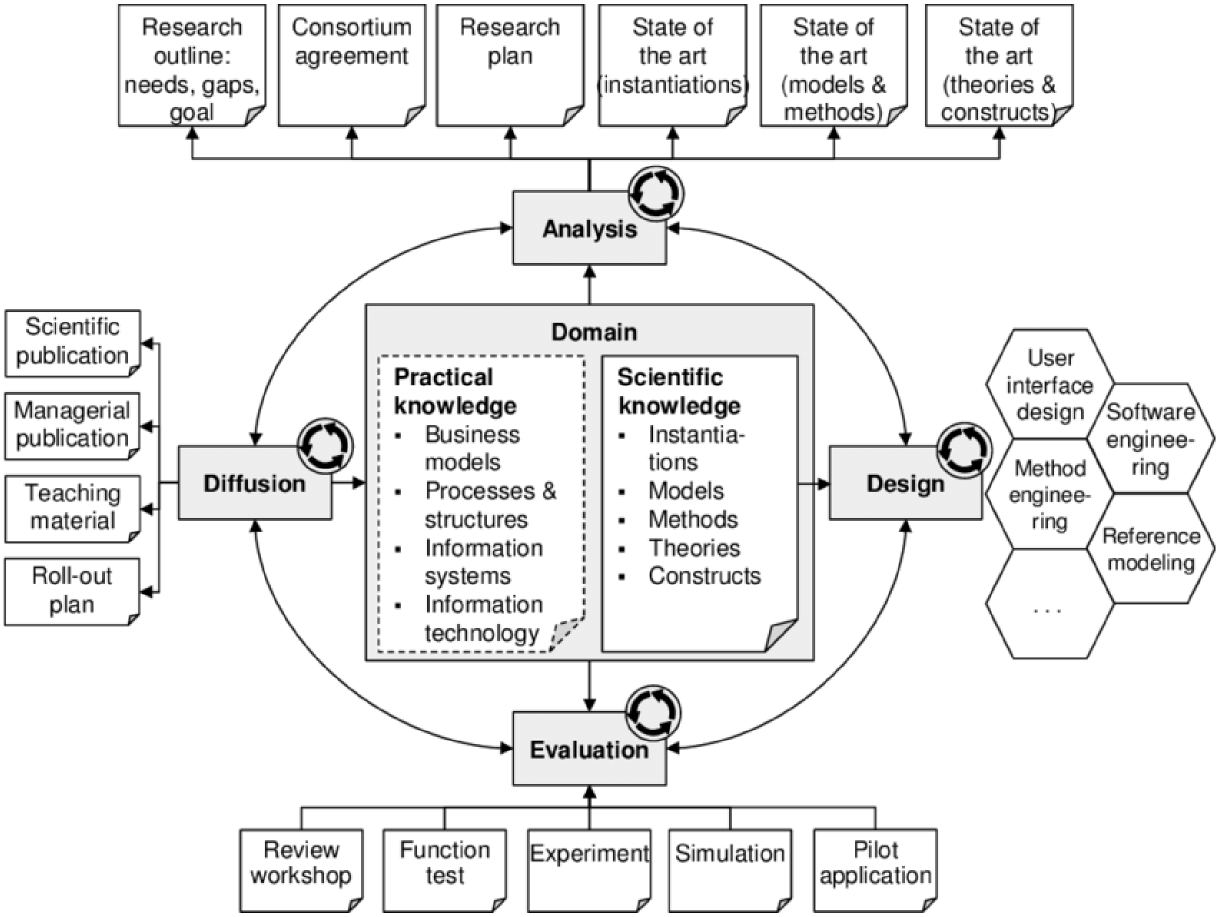


Figure 1.3: Consortium Research. Source: [122]

- instantiations (implemented and prototype systems).

On the other hand, [57] identify three levels of contribution types: from more specific and less mature to more abstract and mature knowledge, defining:

- Level 1: situated implementation of artefact: instantiations (software products or implemented processes).
- Level 2: nascent design theory-knowledge as operational principles/architecture e.g. constructs, methods, models, design principles and technological rules.
- Level 3: well-developed design theory about embedded phenomena: design theories.

[57] also provide classification of research contribution depending on the research problem targeted and the quality of the research result. Figure 1.4 presents relation of the project contexts and potential design science research contributions. Application domain maturity indicates the maturity of the problem context, whereas solution maturity represents the level of detail of the proposed artefact being a solution to identified problems. It is important to note that the framework focuses on the on the knowledge start-points to support understanding of goals and contributions to be achieved. These may be summarised as: new solutions to new problems (radical breakthrough), new solutions to known problems (create better, more efficient and effective artifices), known solutions from other domains extended to known problems from another one (innovation) and known solutions for known problems (such situations however rarely demand to apply research methods to solve a problem).

1.3.4 Research Process

The research process may be initiated both by a researcher and a practitioner and consists of four phases, including analysis, design, evaluation and diffusion. The following sections briefly describe activities undertaken in each of these phases.

Analysis

Analysis deals with identification of a research problem and development of a research plan. In this phase, also importance of results either for research or for practice shall be studied. Following the research problem, research questions and objectives should be specified. This should be performed in line with studying the knowledge base i.e. analysing the state of the art

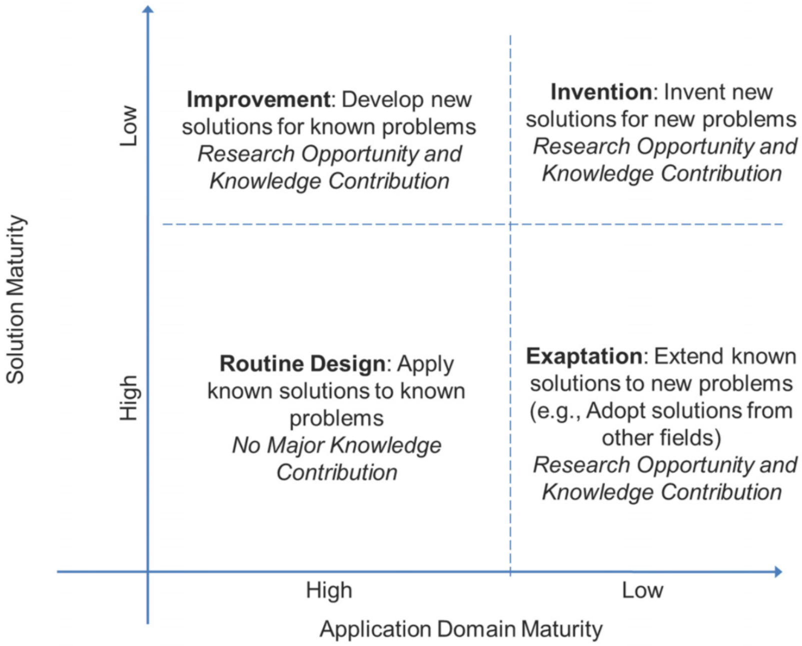

Improvement: Develop new solutions for known problems
Research Opportunity and Knowledge Contribution
Invention: Invent new solutions for new problems
Research Opportunity and Knowledge Contribution
Routine Design: Apply known solutions to known problems
No Major Knowledge Contribution
Exaptation: Extend known solutions to new problems (e.g., Adopt solutions from other fields)
Research Opportunity and Knowledge Contribution
Low
High
Solution Maturity
High
Low
Application Domain Maturity

Figure 1.4: Research contribution. Source: [57]

in the domain in question. Finally, the research plan needs to be created indicating also research methods that shall be used to develop specific artefacts.

Design

The artefacts should be designed and created using generally accepted methods and these methods shall be explicitly given. The research methods may be different depending on the type of the research problem targeted and also on the artefact that is to be developed. For example, one may apply either exploration research methods such as surveys, case studies, expert interviews and IS analysis or artefact design approaches such as demonstration of prototype construction, method engineering and reference modelling.

Evaluation

Following design of an artefact, evaluation needs to take place. Evaluation requires to rigorously demonstrate the utility, quality and efficacy of a designed artefact using well executed evaluation methods [64, 165].

The evaluation may be [165]:

- formative (focusing on consequences and supporting decisions aiming at improving the evaluand) or summative (producing empirically based interpretations that provide a basis for creating shared meanings about the evaluand in the face of different contexts);

- ex ante or ex post depending on the moment in the research process when the evaluation is executed. Ex ante evaluation concerns the predictive evaluation performed to estimate the impact of future situations and happens before design and construction begin. Ex post evaluation concerns an assessment of the value of the solution based on financial and non-financial measures.

Regardless the type of the evaluation, its goals need to be specified before starting execution of the evaluation and validation activities. These goals may be diverse and include [165]:

- studying how well an artefact achieves its environmental utility,
- providing evidence that theory leads to a developed artefact that is useful for solving a problem,
- comparison with other artefacts,
- considering utility being a complex issue and influencing complexity of evaluation,
- evaluation of artefact for side effects,
- debating why an artefact works or not, taking into account the existing body of knowledge.

Evaluation may be carried out inter alia within laboratory experiments, pilot applications (instantiation of prototypes), simulation procedures, expert interviews, field experiments (instantiations in a number of organisations. [64] indicate the following evaluation methods:

1. observational, including case study (application of an artefact in a business environment) and field study (monitoring usage of an artefact in multiple projects);
2. analytical: static analysis (examining structure of an artefact), architecture analysis (fitting an artefact into technical IS architecture), optimisation (demonstrating optimal properties of an artefact or providing optimality bounds on its behaviour), dynamic analysis studying an artefact in use for dynamic qualities;
3. experimental in controlled experiment or simulation studying artefact with artificial data;

4. testing covering white and black box testing, with emphasis either on discovering failures and defects (black box testing) or performing coverage testing of some metric in the artefact implementation (white box tests);
5. descriptive: informed argument based on the existing knowledge base or scenarios showing utility of an artefact.

These methods influence the procedure of how the evaluation should be performed. [165] propose four steps within this phase:

- define goals of the evaluation: meaning addressing the rigour (in terms of efficacy and effectiveness of the artefact), uncertainty and risk reduction, ethics and efficiency of the evaluation process.
- choose the evaluation strategy: why, when and how to evaluate,
- determine properties to evaluate: what to evaluate,
- design the individual evaluation episodes: actual evaluation.

The evaluation phase in design science research process should appear more than once. This is due to the fact that design science research is an incremental, iterative process with multiple evaluations.

Diffusion

Finally, the diffusion of the research results should be carried out. The diffusion (otherwise referred to as communication of results) may be executed with the use of the following means: papers, oral presentations, theses, books, etc. It needs to be underlined that the communication tool chosen should take into account the preferences of the audience and the language of communication towards this audience.

Summary

Table 1.5 presents the research guidelines elaborated and explains how they were targeted while carrying out the research work regarding the thesis.

Table 1.5: Research guidelines being addressed in the thesis.

Methodological challenge	**Explanation**
Delivering results developed within the scientific rigour.	All research results developed in the framework of this research followed a well defined methodology. In all cases relation to state of the art is specified and contribution is discussed.
Collaborating with practitioners.	Part of the research followed the consortium research methodology including communities of practitioners. This is especially valid for vertical research goals, addressing domains of telecommunications and smart grid. The collaboration enabled to define and validate research objectives, assess the results and their impact. This challenge is also addressed by developing of joint research projects.
Developing results (artefacts) that are applicable to important and relevant business problems and therefore contributing to practice.	Results achieved within the scope of this research are of twofold value. Firstly, they provide research contribution what is confirmed by acceptance of these results after peer review processes to conferences and journals. Second, as outcomes of joint academia-business projects, they have business value to be exploited in business scenarios.
Each artefact must be justified and validation needs to be possible, in particular utility, quality and efficacy of an artefact must be rigorously demonstrated by an application of well-executed evaluation methods.	All artefacts developed were evaluated following methods applied in the research domain for similar classes of solutions. Therefore, not only an artefact was evaluated w.r.t. to previously defined objectives, but also in comparison to a baseline in a domain.

Having impact on both: research and business practice, communicating results to both technology-oriented and management-oriented audiences.	Research results are accessible to the public as research papers or accessible resources. Companies test artefacts in their business settings and implement the methods in their final products.

Source: own study

1.4 Structure of the Document

The thesis consists of eight chapters including an introduction and a summary. First chapter describes motivation for work that was carried out for the last 8 years together with discussion on its importance both for research and business practice. The motivation for this work is much broader and emerges also from business importance of profiling and personalisation. The introduction summarises major research directions, provides research questions, goals and supplementary objectives addressed in the thesis. Research methodology is also described, showing impact of methodological aspects on the work undertaken.

Chapter 2 provides introduction to the notion of profiling. The definition of profiling is introduced. Here, also a relation of a user profile to an identity is discussed. The papers included in this chapter show not only how broadly a profile may be understood, but also how a profile may be constructed taking into account different data sources.

Profiling methods are introduced in Chapter 3. This chapter refers to the notion of a profile developed using the BFI-44 personality test and outcomes of a survey related to colour preferences of people with a specific personality. Moreover, insights into profiling of relations between people are provided, with a focus on quality of a relation emerging from contacts between two entities.

Chapters from 4 to 7 present different scenarios that benefit from application of profiling methods. Chapter 4 starts with introducing the notion of a public utility company that in the thesis is discussed using examples from smart grid and telecommunication. Then, in chapter 4 follows a description of research results regarding profiling for the smart grid, focusing on a profile of a prosumer and forecasting demand and production of the electric energy in the smart grid what can be influenced e.g. by weather or profiles of appliances.

Chapter 5 presents application of profiling techniques in the field of telecommunication. Besides presenting profiling methods based on telecommunication data, in particular on Call Detail Records, also scenarios and issues related to privacy and trust are addressed.

Chapter 6 and Chapter 7 target at horizontal applications of profiling that may be of benefit for multiple domains. Chapter 6 concerns profiling for authentication using un-typical data sources such as Call Detail Records or data from a mobile phone describing the user behaviour. Besides proposing methods, also limitations are discussed. In addition, as a side research effect a methodology for evaluation of authentication methods is proposed.

Chapter 7 concerns personalisation and consists of two diverse parts. Firstly, behavioural profiles to change interface and behaviour of the system are proposed and applied. The performance of solutions personalising content either locally or on the server is studied. Then, profiles of customers of shopping centres are created based on paths identified using Call Detail Records. The analysis demonstrates that the data that is collected for one purpose, may significantly influence other business scenarios.

Chapter 8 summarises the research results achieved by the author of this document. It presents contribution over state of the art as well as some insights into the future work planned.

Chapter 2

Introduction to Profiling

2.1 Introduction

2.1.1 Motivation

This chapter addresses the first goal from the background and definition perspective of this thesis. The goal is to **define a profile of a person or a thing and identify features of a person or a thing that may be represented in a profile and may be usable for diverse application scenarios**.

To achieve this goal, two secondary goals were defined and concern:

G1.1 Analysing different approaches for describing a profile or an identity of a user or a thing that are applied in different classes of systems e.g. identity management systems.

G1.2 Analysing how to instantiate a profile of a person or a thing using semantic approaches enabling for diverse application scenarios.

These goals are aligned with specific sections of this chapter, namely goal G1.1 with sections: 2.2 and 2.3, and goal G1.2 with section 2.4. Each of these sections includes a paper published after a peer review process (detailed bibliographic references are provided).

2.1.2 Structure of the Chapter

The chapter consists of five sections including an introduction presenting the relation to the goals of the thesis and a summary that presents results that were achieved in relation to these goals. Section 2.2 presents the concept of an automatic creation of user identities. Section 2.3 studies the state of the art in the area of profiling and identity management. Section 2.4 focuses on the process of building a user profile/an identity applying available data models that enable for reasoning.

2.2 Ego – Where User Modelling Meets Identity Management

The goal of this section and therefore the paper included, concerns proposing a system that is to bridge the gap between the identity management and user modelling systems. The developed solution is to enable users to automatically create their identities and manage these identities for the needs of different services. Sharing should allow not only authorisation, but also personalisation of content displayed to a user.

The paper was published in a Journal of Wrocław University of Economics. Detailed bibliographic reference is as follows: Abramowicz, W., Filipowska, A., Kalisz, P., Werno, Ł., Dzikowski, J., Małyszko, J., 2010, Ego - where user modeling meets identity management., Prace Naukowe Uniwersytetu Ekonomicznego (AE) we Wrocławiu, 119, pages 11-20.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 119
Informatyka Ekonomiczna 18 2010

Witold Abramowicz, Agata Filipowska, Jakub Dzikowski, Paweł Kalisz, Jacek Małyszko, Łukasz Werno
Poznan University of Economics

EGO – WHERE USER MODELING MEETS IDENTITY MANAGEMENT

Summary: Current identity management and user modeling systems suffer from certain limitations. Although, theoretically, they have similar goals: to represent a user and to support him while carrying out certain activities, there is a clear gap between them in means of aspects of users' activity, which they represent. This paper addresses this issue, proposing a system, which aims at incorporating strengths of both types of systems to provide a single, unified method of representation of users in the Web. Description of the system includes functional analysis, system architecture, identity life cycle scheme and more.

Key words: identity management, user modeling, adaptive systems, personalization.

1. Introduction

The Web is nowadays becoming an integral part of both industry and people's everyday life [Mahonen 2006]. As a result, more and more real-world activities are, at least partially, transferred to the Web. These activities include for example shopping [Sangwan, Siguaw, Guan 2009], communication with friends [Preece 2000], fulfilling one's information needs [Abramowicz 2008]. During these activities Internet users often encounter some obstacles.

Information overload [Grise, Gallupe 2000] makes it often impossible for an individual to process the information delivered as a response to a query and find the information he really needs [Ho, Tang 2001]. The information systems could help users in searching the information they need, but first they have to 'understand' users' needs [Brusilovsky, Millán 2007].

Other problems users may encounter in the area of virtual communities [Rheingold 1993]. Users create communities by gathering around specific topics, related to their interests, in online forums, chats, massively multi-player online games, etc. [Nabeth 2006]. A problem, that a user might come across in this situation, is an obligation to create a new user account and perform a registration process each time, when he wants to become a member of a new portal and new community [Grohol 2006]. Additionally, accounts on these portals are separate entities, and user cannot easily build his single, unified online identity and update it, if needed [Jaquet-Chiffelle 2008].

In the past there have been numerous initiatives aimed at solving such problems. Research on user modeling, adaptive systems and identity management systems are the most obvious examples [Brusilovsky, Millán 2007; Meints 2009, Zwingelberg 2009]. Nevertheless, the current solutions suffer from certain limitations. As section 2 of this paper shows, there is a lack of a single method of representing users, which would be adequate to a broad set of applications, required by the rapidly growing users' involvement in the Web.

The presented paper proposes a system addressing the current limitations. It introduces Ego – a project aiming at creation of a system based on virtual identities, that will enable Internet users to easily present their needs in a broad set of applications. Our approach focuses on giving the control over the identity to the user himself and providing a framework for exploiting information stored within identities on a wide range of applications. Such system will improve the overall accuracy of an information flow within the Web and help users in their everyday Internet activities.

This paper is structured as follows: section 2 presents the current state of the art in the domains of user modeling and identity management. It raises issues of adaptive systems and identity management. Third section presents the proposed approach, including the functional analysis, the Ego system architecture and the identity life cycle scheme. The article concludes with a summary including also directions of the further research.

2. Related work

2.1. User identity in the information society

The notion of identity in the information society is gaining recently much attention. It is expected, that the popularity of the Internet will result in new forms of online identities, merging the real world and life of individuals with the digital ones [Mahonen 2006]. FIDIS project aims at pointing out, that proper identity management is a key in the modern information society development [Rannenberg, Royer, Deuker 2009]. The project, among others, raise a problem of developing, organizing and standardizing concepts connected with identities, such as user modeling and identity management [Hedbom, Van Alsenoy 2009]. We will focus on these two concepts.

2.2. User modeling and adaptive systems

The issues of understanding the user needs and adaptation to them have been topics of interest of various researches for many years, since the introduction of the classic papers on the subject [Goldberg, Nichols, Oki 1992; Rich 1979]. The understanding of users' needs requires building user models [Brusilovsky, Millán 2007]. A user model is a source of knowledge about the user, containing assumptions covering all users' activity aspects that can be used by the particular system [Chen, Magoulas 2005]. Information systems that adapt their contents or behavior to the user needs are called adaptive systems [Brusilovsky, Millán 2007].

There is a wide range of applications of adaptive systems, ranging from recommendation systems and information filters, intelligent e-learning, to various systems adjusting interaction patterns to users' preferences [Kay 2001]. One of the most successful applications of this kind of systems is a recommender system embedded in the Amazon.com online shop [Linden, Smith, York 2003]. It builds users' models based on their historical shopping behaviors and on marks, that users have given to in-stock items [Gregory, Jacobi, Benson 2001]. These models are then exploited by the recommender system that as a result provides a personalized offer to each client [Gregory, Jacobi, Benson 2001].

Another example of an adaptive system is the Google search engine. User models are built here upon information about users' search and browsing history for users logged-in to their Google account [Sullivan 2007] or accept anonymous cookie file [Horling 2009]. These models are used to personalize search results in Google Search Engine [Horling 2009; Sullivan 2007].

Both systems share some limitations. Although they gather a lot of information about users, the user models are inaccessible from the outside of the system. Apart from trust and privacy issues, the user models, although potentially very useful for the users, are stored beyond their control and cannot be transferred from one system to another. As a result, every time, when one registers to a new system, the process of learning the user's needs by the new system has to be started from scratch.

2.3. Identity management systems

Another interesting type of systems from the point of view of this paper are the identity management systems. Identity management systems deliver a wide range of functionalities, from implementing authentication, authorization, and accounting to user-controlled context-dependent role and pseudonym management [Meints, Zwingelberg 2009]. These systems work both in corporate environment and, in a wider context, on a scale of the whole Web (Global Identity Management systems) [Meints, Zwingelberg 2009].

W3C [*Requirements…* 2001] has defined requirements for a global Identity Management service. The list of requirements for this service, among others, includes interoperability, extensibility and negotiated privacy and security.

A popular example of a global identity management system is OpenID[1]. OpenID eases registration and logging-in processes by providing a single login and password on many different systems and allows reusing of some basic information

[1] http://www.openid.net.

about the user across these systems [Becker 2006; Recordon]. The OpenID system is based on a decentralized architecture, what means, that there is no central repository of identities. Each user can choose where to store his identity – even at his own local server [Becker 2006].

The classical identity management systems focus on user authentication, and amount of other information about users managed by such systems is limited [Meints, Zwingelberg 2009]. In the case of OpenID, identity contains only basic information about the user, such as his name, pseudonym and the e-mail address [Becker 2006]. Little information makes it impossible to process such identity in order to help user in fulfilling his information needs.

Based on the above analysis, one may note, that the identity management and user modeling are currently heavy researched, but the current systems still suffer from some limitations. There is a need for providing a more universal approach to the identity management, if the Web is to meet growing requirements of more and more active users. The next section brings a proposal of a solution to these issues.

3. Proposed approach

The proposed approach aims at creating a model of identity, which will allow reusing of users' representations to improve addressing and flow of information throughout the Web. We assume that the system should have the following features:

- versatility (extensibility and interoperability),
- decentralization,
- full control of a user over his identity in means of access control issues.

In this section we discuss certain features, which future identity systems should incorporate. We also propose a schema for an identity life cycle.

3.1. Usage scenario

The basic idea of the Ego system is to create a versatile application, which will automatically build universal user model and manage its evolution. Enabling user to create and manage their representation should enhance Internet browsing and allow delivering more user-oriented content. The process starts when a user creates an account and provides some basic data about him or her. From this time on, a special component of the Ego system starts learning user's information needs and building (and updating) his or her model (called in the Ego project user's virtual identity). The creation will be based upon the analysis of user's online activities (for example extended analysis of web pages the user visits).

In a typical scenario (Fig. 1), a user (having logged into the Ego system in a form of a plug-in toggle activation) browses a particular web page. A certain component (scrobbler) retrieves information about the content of the website and browsing context (date and time, localization of the user, hardware and software platform). Acquired information is sent to identity provider (identity server), which – using the information – updates user's virtual identity. During browsing the Internet by the user, service providers (websites, recommender systems) can query user's virtual identity and personalize presented content utilizing Ego response.

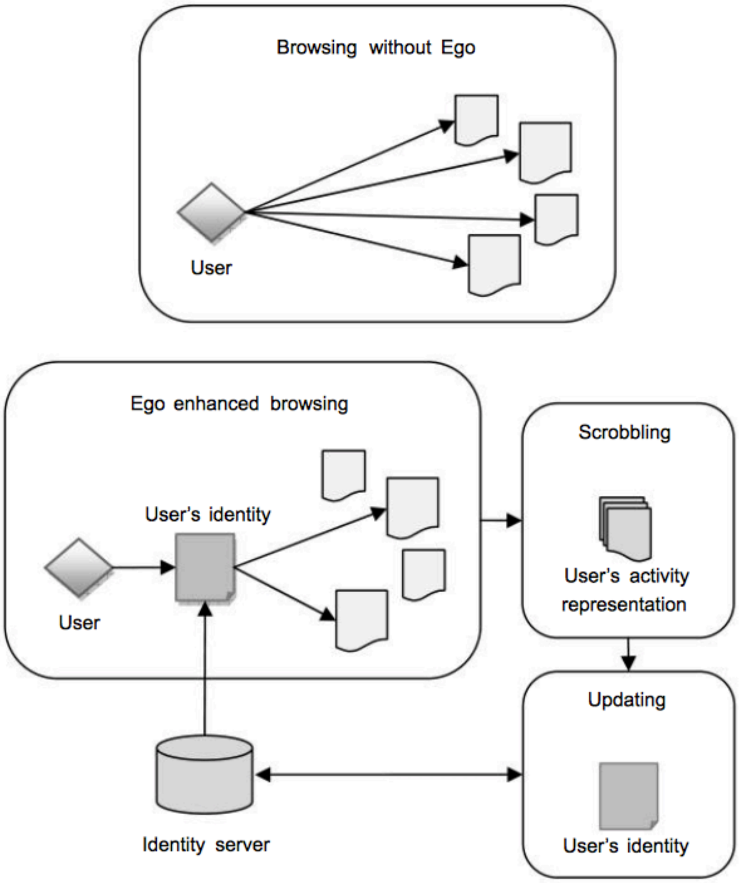


Fig. 1. Browsing with and without Ego system

Source: own elaboration.

3.2. Functional requirements

To address the main objective of the system, it should support the following functional requirements:

1. The system must gather various information about the user from heterogeneous and mostly unstructured sources.

2. The process of information gathering must not be bothersome to the user.

3. The user needs to have control over information stored in his or her identity and over access rights to the information.

4. User's virtual identity should comprehensively represent user's information needs and their evolution.

5. The external applications should be able to query user's virtual identity in order to receive relevant information for the purposes of personalization.

3.3. System architecture

There are two principles we are following as far as system architecture is considered and these are versatility and decentralization. By the versatility we understand the feasibility to be implemented in the scope of a single adaptive system as well at a scale of the whole Internet. The decentralization we understand similarly as in the OpenID case, which means that anyone can use the system or be an Ego provider, without being registered or approved by any central organization. In this section we present the architecture that allows gaining these two features.

The basic view on the architecture of the proposed system is presented in the Fig. 2. The architecture comprises several elements: scrobblers, identity servers, identity catalogues and client services.

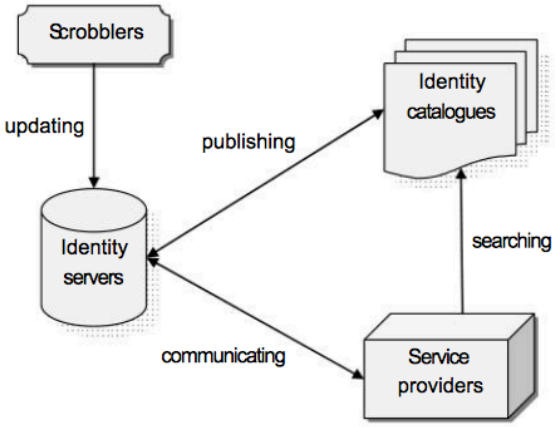


Fig. 2. The components of the Ego architecture and flows between them

Source: own elaboration.

Scrobblers are applications that examine users' activities within the system and provide the identity servers with information about users actions. Scrobblers can be for example:

- plug-ins to a Web browser, which extract contents of web pages read by the user,
- server-side applications, which gather information about resources accessed by users,
- plug-ins to media players, which send information about, for example, songs that were listened by the user.

These programs can analyze resource content, extract the essence of accessed documents and perform various operations (tokenization, lemmatization or stemming, morphological analysis and named entity recognition). For the purpose of user modeling in Ego system, any scrobbler needs to return a fixed representation of current user's activity (surrogate of information resource, the source and the user's identifiers, timestamp etc.).

Acquired information in a form of an RDF document is sent to a chosen **identity server** via the HTTP protocol. The identity server's main functionality is to store and update the user's virtual identity to make it better reflect user activities, interests and knowledge. This issue is described in the identity lifecycle section.

The identity servers provide users with a control panels for managing access rights for information stored within identities. It allows an identity owner to decide, which services, Web agents or websites can access which part of his or her identity information.

According to the decentralization principle, the identity sever can be deployed by virtually anyone. To enable versatility, such a server can work both in small networks, for example in a company environment, and at a scale of a whole Internet. General-purpose user's virtual identity stored on the Internet-level server is the global identity. Every other identity will be referred to as a local identity. The user can have a global identity and simultaneously the local one in his company, which stores only information about his in-work activities. It is possible to connect local identities with the global one. Such assignment implies that information stored in the local identities will be used to update the global identity.

The **identity catalogues** publish identities to other services or users within the Web. The client services can search for certain identities and information, according to permissions assigned by identity owners.

The last link of the Ego identity system are **client services**. Using purpose-built query protocols, and both authentication (that particular service) and authorization (user has agreed to share his data) methods, client services retrieve information about the user from his identity and can exploit it for many different purposes.

3.4. Identity life cycle

During its lifecycle, an identity is subject to many different processes. These are creation, updating, merging, user interference, access control and querying (Fig. 3).

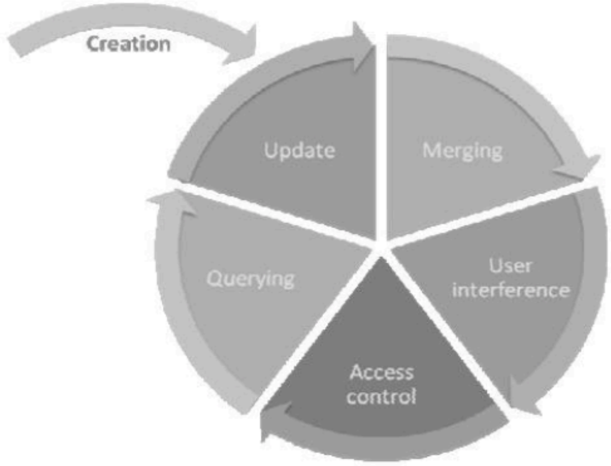


Fig. 3. The identity life cycle

Source: own elaboration.

Creation. To assure the usefulness of an identity from the very beginning of its existence, the system enables using start-up identities. A start-up identity is a predefined user stereotype, which includes expected user characteristics. For example, a newly hired employee in a company can be fitted with a predefined identity based on a role he plays in an organization. The stereotype in this case will be built by merging identities of other users, that play the same role in the organization.

Updating is a process of incorporating into the identity the structured data retrieved by scrobblers from recent activities of the user. It is crucial at this point to establish such method of updating, which assures gradual evolution of an identity, and not its rapid changes. The identity must reflect the current characteristics of a user, but also be aware of its history.

Merging. It is possible for the user to have many identities. Therefore, it is necessary to provide an efficient method of merging several identities, so that querying would concern a broader view of user's information needs and interests. This functionality potentially can be used also for other purposes, for example in already mentioned generating of start-up identities by merging individuals' identities.

User interference. The user should have a possibility to control the information stored in his identity. For example, he or she should be able to determine what kind of information about him or her should be gathered, or even to delete certain information from identity. Still, this interference should be limited to ensure objectivity of information stored in the identity. As a result, user interference should improve his or her virtual identity.

Access control. Another dimension of control of the user over his identity is a possibility to assign permissions to access some parts of the identity by certain services or groups of services. It is also important to make this activity simple and not bothering to the user.

Querying is the most important phase in means of utilization of information stored in identity. Based on permissions assigned, services can send certain queries to the identity. As a result, services can gain knowledge about a user, which can be exploited, for example, for adaptation to user needs.

4. Conclusions and further work

The current state of the Internet development faces many issues that can and need to be solved. There is a need to create a system which combines advantages of identity management and user modeling so as to provide more accurate information about Web user to wide range of adaptive systems.

Ego system will cover requirements of versatility and distribution, and will allow users to improve or adjust their identities. The system will provide new methods of learning about users' interests and pursues and deliver more direct content.

Further work will focus on the modularization of the system and on considering the possibilities of integration with existing solutions, such as OpenID. Additional methods of identity utilization will be considered, such as, for example, it's presentation on social networking sites.

Literature

Abramowicz W., *Filtrowanie informacji*, AE, Poznań 2008.

Becker P., *The case for OpenID*. Retrieved February 19, 2010, from http://blogs.zdnet.com/digitalID/?p=78 (Published 2006, December 4).

Brusilovsky P., Millán E., User models for adaptive hypermedia and adaptive educational systems, [w:] P. Brusilovsky, A. Kobsa, W. Nejdl, *The Adaptive Web: methods and strategies of web personalization*, Springer, Berlin 2007, s. 3-53.

Chen S., Magoulas G., *Adaptable and Adaptive Hypermedia Systems*, IRM Press 2005.

Goldberg D., Nichols D., Oki B.T., *Using collaborative filtering to weave an information tapestry*, Communications of the ACM 1992, 35 (12), s. 61-70.

Gregory D., Jacobi J.A., Benson E.A., *Collaborative recommendations using item-to-item similarity mappings*, 2001, United States Patent 6266649.

Grise M., Gallupe R., *Information overload: Addressing the productivity paradox in face-to-face electronic meetings*, "Journal of Management Information Systems" 2000 (16).

Grohol J., *Anonymity and online community: Identity matters*, Retrieved February 19, 2010, from http://www.alistapart.com/articles/identitymatters/(2006, April 4).

Hedbom H., Van Alsenoy B., *D19.3: Standardisation report*, FIDIS (Future of Identity in the Information Society) 2009.

Ho J., Tang R. *Towards an optimal resolution to information overload: an infomediary approach*, Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work 2001.

Horling B.K. *Personalized Search for everyone.* Retrieved February 19, 2010, from The Official Google Blog: http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html (2009, December 4).

Jaquet-Chiffelle D., *D2.13 virtual persons and identities*, FIDIS (Future of Identity in the Information Society) 2008.

Kay J., *User Modeling for Adaptation*, [w:] C. Stephanidis, *User Interfaces for All*: *Concepts Methods, and Tools*, Lawrence Erlbaum Associates 2001.

Kobsa A., Wahlster W., *User Models in Dialog Systems*, [w:] A. Kobsa, W. Wahlster, *User Models in Dialog Systems*, Springer, Berlin 1989.

Linden G., Smith B., York J., *Amazon.com recommendations: Item-to-item collaborative filtering,* IEEE Internet Computing 2003, 7 (1), s. 76-80.

Mahonen P., *The Future Networked Society*, EIFFEL Think Tank 2006.

Meints M., Zwingelberg H., *D3.17: Identity management systems – recent developments*, FIDIS (Future of Identity in the Information Society) 2009.

Nabeth T., D2.2: *Understanding the identity concept in the context of digital social environments*, *FIDIS* (Future of Identity in the Information Society) 2006.

Nabeth T., *D2.3: Models*, FIDIS (Future of Identity in the Information Society) 2005.

Preece J., *Online Communities: Designing Usability and Supporting Sociability*, John Wiley & Sons, New York 2000.

Rannenberg K., Royer D., Deuker A., *The Future of Identity in the Information Society*: *Challenges and Opportunities*, Springer, Heidelberg 2009.

Recordon D., *ACM Workshop on Digital Identity Management*, New York, s. 11-16.

Requirements for a Global Identity Management Service, W3C Workshop on Web Services, San Jose, CA USA 2001.

Rheingold H., *The Virtual Community*: *Homesteading on the Electronic Frontier*, Perseus Books, 1993.

Rich E., *User modeling via stereotypes*, "Cognitive Science" 1979, 3 (4), S. 329-354.

Sangwan S., Siguaw J. A., Guan C., *A comparative study of motivational differences for online shopping*. ACM SIGMIS Database 2009, 40 (4), s. 28-42.

Sullivan D., *Google Search History Expands, Becomes Web History*, Retrieved February 19, 2010, from http://searchengineland.com/google-search-history-expands-becomes-web-history-11016 (2007, April 19).

EGO – NA POGRANICZU MODELOWANIA UŻYTKOWNIKA I ZARZĄDZANIA TOŻSAMOŚCIĄ

Streszczenie: Obecne systemy zarządzania tożsamością oraz systemy modelujące użytkownika, pomimo podobnego celu – reprezentacji użytkownika oraz wspierania go w określonych czynnościach – obejmują całkowicie różne aspekty aktywności użytkownika. W artykule przedstawiono wymagania dla systemu łączącego zalety zarządzania tożsamością oraz modelowania użytkownika, tworzącego ewoluujące w czasie, uniwersalne modele użytkownika, ułatwiające personalizację w sieci. Opisano wymagania funkcjonalne dla takiego systemu, jego architekturę oraz cykl życia wirtualnej tożsamości użytkownika.

2.3 A Survey of Empirical Research on the Digital Identity

The goal of this section is to study functionality and use cases for identity management systems available on the Web. The focus of the work is to describe trends, ideas an shortcomings of existing solutions, identifying potential extensions. This section contributes to achieving the following secondary goal of the thesis: ”Analysing different approaches for describing a profile or an identity of a user or a thing that are applied in different classes of systems e.g. identity management systems”.

The paper was submitted to: Business & Information Systems Engineering[1] and is currently in the review process. The submission identifier is: BUIS-D-18-00231.

[1] `http://www.bise-journal.com`

Business & Information Systems Engineering manuscript No.
(will be inserted by the editor)

A survey of empirical research on the Digital Identity

Agata Filipowska · Milena Stróżyna · Jacek Małyszko · Piotr Kałużny · Dawid Grzegorz Węckowski · Witold Abramowicz

Received: date / Accepted: date

Abstract Nowadays, more and more of day-to-day activities are performed on the Internet. The information about people, that is manifested by them on the Web, can be collected to build representations of individuals expressed in the form of digital identities. Providing a platform, that enables secure exchange of parts of an identity compliant with the General Data Protection Regulation (GDPR) allows for a wide range of services to be created. Within this paper, the digital identity definition, as well as functionality of user-oriented Identity Management Systems (IMS) are discussed. We also study the state-of-the-art of identity-related solutions and address the research gap in this area.

Keywords Digital identity · Personalization · Identity Management Systems

1 Introduction

Nowadays, the Internet becomes more and more important area of our day-to-day activity. It cannot be treated any more as a space, in which users are anonymous. Most users leave a significant amount of information about them on the Web. They often abandon their anonymity freely, to stay connected with friends on social networking sites, communicate with government or build their reputation [Leenes et al., 2008, Bernstein et al., 2011]. The information left by users is collected by different service providers, who use profiling mechanisms to analyse users' behaviour in order to personalize their services. As a result, a problem, of how users should manage their personal data spread across the

Department of Information Systems, Faculty of Informatics and Electronic Economy, Poznań University of Economics and Business
al. Niepodległości 10, 61-875 Poznań, Poland
Tel.: +48 61 8543381
E-mail: {agata.filipowska, milena.strozyna, jacek.malyszko, piotr.kaluzny}@ue.poznan.pl, dawid.weckowski@gmail.com, witold@abramowicz.pl

Web as many partial identities and profiles, emerged. This problem is especially valid after introduction of the General Data Protection Regulation[1].

This issue is being researched for many years now and a number of solutions in the area of user-oriented identity management already exists [Hildebrandt & Backhouse, 2005, Barisch, 2009, Ahn et al., 2007, Chen, 2007, Koch, 2002]. Although the literature on the topic of Identity Management is broad, and includes theoretical principles and underlying security issues (with the most broad example in form of the book about the digital identity management [Laurent & Bouzefrane, 2015]), there is still scarcity of papers that would focus particularly on the comparison of user-controlled systems. This gap is addressed by this paper.

Compared to our work, similar surveys on profiling and Identity Management Systems (IMS) have been conducted by [Ferdous & Poet, 2012, Manso et al., 2014, Torres et al., 2013, Banihashemi et al., 2016]. The most recent research is a survey that focuses on the desirable qualities that should identify Identity Management systems, providing a taxonomy of criteria, and evaluation of those systems' traits in aspects of: security, user control, system capabilities and cost-effectiveness [Banihashemi et al., 2016]. Other survey focuses on more technical details and data flow in popular protocols used in Web-based IMS [Manso et al., 2014]. [Ferdous & Poet, 2012, Torres et al., 2013] present a comparison of IMS from the point of view of system requirements, such as privacy, security, affordability, trustworthiness, law enforcement, interoperability and functionality. [Torres et al., 2013] also focus on systems' architectures and used components. [Ferdous & Poet, 2012] in turn, compare only 6 leading Identity Systems, without information on other solutions, which are less popular or are still under-development. The main shortcoming of these reviews is that the comparison between various solutions (projects, protocols, alliances) is made jointly, using the same requirements.

This paper focuses on comparing functionality and use cases of diverse IMS available on the Internet. Our aim is to analyse, what are the current trends and ideas concerning the identity management, what problems in this area have already been solved, and what are shortcomings of the existing solutions. In comparison to other surveys on identity systems, our analysis is conducted from the end user's point of view (user-centric approach).

The paper is organized as follows. Section 2 introduces the notion of a digital identity and is a theoretical foundation of the paper. Section 3 presents the research methodology applied. Sect. 4 introduces the identity management systems and defines the criteria used while carrying out the analysis. Then, selected identity-related protocols, projects and initiatives are compared, along with indication of open issues and related challenges. In Sect. 5 we summarise the findings and show future research directions.

[1] https://ec.europa.eu/info/law/law-topic/data-protection_en

2 Virtual Identity - Fundamentals

The concept of a digital identity has been adopted by the information science as a formal representation of knowledge or a set of claims, made by one party about itself or about another entity [Cameron, 2005]. Such an entity can be anyone or anything that can be characterized through measurement of attributes [Kubicek, 2010]. Thus, an identity can be assigned to a wide range of subjects: a person (user, citizen, customer, employee etc.), an abstract subject (group of people, organization, legal entity), a physical good and an object (appliance, device, vehicle, computer), a process or a virtual object, which is a virtual representation of capabilities/resources provided by real-world objects [Miorandi et al., 2012, Rundle et al., 2010, Kubicek, 2010]. The review of working definitions of key terms that underpin the studies on the identity in the information science is presented by [Clarke, 2008, Nabeth, 2009].

In the concept of the Internet of Things, each object or a thing that connects and communicates with other objects, can have an identity or a virtual personality [Miorandi et al., 2012]. Also a process can have an identity, that defines e.g. roles involved in this process, people assigned to these roles as well as access rights of the process to various resources. However, so far there is no widely-recognised identity framework, that would enable to recognize these identities. Therefore, existing objects' identities are managed with proprietary solutions or niche standards that have not been mapped to each other nor interoperate [Bassi & Horn, 2008]. Within this article, the understanding of the concept of the identity (or digital identity) is limited only to people, excluding legal persons, since profiles of legal entities have been the subject of our previous research e.g. [Wieloch, 2011].

The identity is defined by a set of attributes characterizing a person [Rannenberg et al., 2009b]. Obviously, a digital identity cannot capture all characteristics of a person. Therefore, it refers to a partial, reductive representation [Rannenberg et al., 2009a, Rundle et al., 2010, Barisch, 2009, Nabeth, 2009]. Traditionally, an identity is considered as a permanent, timeless and unconditional entity, persisted in a kind of a datastore in order to be accessible many times for a long period of time. On the other hand, it can also be understood as something temporary, created and used on-the-fly within a single session or as something conditional that exists only in a given context [Identity Commons, 2012, Rannenberg et al., 2009a]. Moreover, the level of control over the information stored within the identity can be different – a person can have full or partial control over his/her identity (user-centric approach), or identity information can be externally controlled: by governments or institutions, companies, commercial entities etc. [Nabeth, 2009].

Abstracting from the three tiers of identity that were identified in [Weik & Wahle, 2008], a digital identity of a real-world person can be simplified to a set of attributes, which describe this individual and his/her roles. These attributes can be a combination of [Hodges et al., 2005, Nabeth, 2009, Rannenberg et al., 2009b, Kubicek, 2010]:

- real-world attributes, such as name, address, biological characteristics, interests, likes/dislikes, competences, skills, functions, social affiliations, performed actions and locations,
- digital attributes, such as connections in social networks, digital traces left by a user (browser history),
- attributes used for identification and authorization purposes, such as logins, passwords, access rights, biometric values, DNA profiles.

Attributes can be permanent (like date of birth, gender), acquired during the life (like diplomas, competences) or temporally assigned to a person (like position, access rights). These attributes can be further categorized according to the domain perspective and activities, in which they are used; for example, different characteristics are needed in e-banking portals and in movie recommender systems. We can therefore say that a person has one identity, but such identity has multiple facets that are used depending on the context [Rannenberg et al., 2009a].

The context depends on some external conditions and state in which the identity is used. It may further determine a set of rules that are applied, regarding the availability and validity of identity attributes. It is connected with the fact that the identity can be used in many aspects of people lives, such as work, leisure, health care, government, travel etc. [Nabeth, 2009, Halperin & Backhouse, 2008]. While some attributes are universal and cut across roles (e.g. preferred language), some are unique to a certain role (e.g. job position). For example, some of information can be connected with the workplace and work environment of the user - when the user is at work, only attributes that are connected with his work life are used. Identification of all possible contexts and providing their clear classification, to the best our our knowledge has not been done yet in relation to the digital identity. Nonetheless, few examples can be listed:

- **Temporal context** is connected with variability in time or seasonality of information about a user. This variability concerns utilization of different identity data, and can result from changes in user's interests or acquiring new attributes [Nabeth & Gasson, 2005a]. In this context, attributes can be differentiated according to specific time-cycles when they are used, e.g. a particular time of a day, a day of a week or a season of a year. For example, if a user likes skiing, but he is interested in this aspect of his life only in the winter, a part of his identity about skiing is used only in the winter season. The temporal aspect of the identity can also be connected with a role of the user (e.g. an employee, a user of a service) or with his/her current state.
- **Location context**, in turn, concerns the situation, when the identity data is used in different places, such as home, work, or city that a user is just visiting. In each of these places, different attributes can be used; for example, at work, attributes concerning user's competences and skills, whereas while searching for an accommodation in a foreign city, attributes connected with user's preferences about a hotel's standard are applied. The

analysis on extension of the context using geo-location information can be found e.g. in [Royer et al., 2009].
- **Device context** emerging from usage of diverse devices, such as laptops, tablets, and mobile phones. Each device can be used in various situations or for different purposes. Thus, attributes in the identity can be associated with different technological contexts [Bhargav-Spantzel et al., 2007, Hovav & Berger, 2009, Halperin & Backhouse, 2008]. For example, if a user uses a business device only for work and a tablet for leisure and fun, different parts of his identity can be used depending on the device the user is currently working on.
- **Social context** of using the identity is connected with a user's affiliation to various groups (e.g. social groups, social networks) and is developed based on his/her online behaviour. This social information can then be exploited to conclude about user's characteristics – based on identities of user's friends or people that are affiliated to the same group [Nabeth, 2009]. Such mechanisms are already widely-used among others in collaborative filtering recommendation systems or reputation systems, and are not of interest to this article.

The identity can be used either in a separated environment (e.g. in a single system or a company) or in many different environments (e.g. by many services, portals or Web sites). Interesting examples of widespread digital identity systems are countries with e-ID systems implemented [Arora, 2008] or e-health identity management systems [Tormo et al., 2013], whose main aim is to simplify government-citizen communication. Those approaches also consider sharing data with external organizations, given that the rigorous security requirements are met. However, such data exchange requires integration of Web identity protocols into those systems, which still remains a challenge.

Another approach assumes that the identity is just one set of claims about a person and typically for each user many digital identities exist [Identity Commons, 2013, Jaquet-Chiffelle et al., 2009]. However, these multiple identities may be collected together under one umbrella based on an association between a set of attributes between providers [Yeluri & Castro-Leon, 2014]. Such an approach allows them to interoperate with each other as a cohesive whole, forming the meta-identity of a person. This enables linking attributes between heterogeneous systems while their ownership stays on a local level [Benantar, 2006]. In this scenario, state of knowledge about a user (set of information stored) can be different for each of the applications [Laurent & Bouzefrane, 2015]. Such consolidated representation of a person's digital identity is called identities' federation [Satchell et al., 2006]. The implementation of the concept of the federated identity allows organizations to securely share confidential user attributes with other trusted providers, and thus increase the usability from the user's perspective e.g. by providing single sign-on (SSO) functionality [Yeluri & Castro-Leon, 2014]. This association of attributes of an entity can also be deduced based on the similarity of attributes between

Table 1 Features of the identity used in further analysis within the paper.

Criterion	Value
Subject	Natural person
Identity's characteristics	Permanent, timeless and unconditional Federated, centralized (one identity with multiple facets) User-centric (controlled by a user)
Attributes stored in the identity	Real-world, digital, identification-related Permanent and temporal Inherited, acquired, assigned
Appliance / Utilization	Used in many environments (Web pages, portals, systems) Used in different contexts (services, places, devices) Used in different roles (Internet user, consumer, employee) Used for different purposes

different providers (e.g. a unique e-mail address), resulting in e.g. joining of user profiles between social media [Gautam et al., 2016].

To sum up, the concept of the digital identity can be perceived and defined in various ways, depending on the identity's subject, characteristics of the identity, attributes that are stored in the identity and the context in which the identity is used. In this paper we focus on the single, permanent, user-centric concept of the identity, which consists of both real-world and digital attributes, and can be used in many different environments and contexts, such as technological, temporal, social or location-based. Due to this fact and the limited length of the paper, cloud based federated identity services, and the underlying theoretical concepts and models were not addressed in this paper. However, they are explained in depth in some of the recent literature [Yeluri & Castro-Leon, 2014, Habiba et al., 2014, Zwattendorfer et al., 2013]. This includes also the concept of Identity as a Service (IDaaS) [Zwattendorfer et al., 2013, Mpofu & Van Staden, 2014], connected mostly with the cloud based IMS.

The solutions and research described in this work mention the federated approaches only if they provide the level of privacy of user attributes/user identity at a similar level as the user-centric approach (see e.g an example of a proxy re-encryption proposed for OpenID by [Nunez et al., 2012] and a more general approach proposed by [Slamanig et al., 2014]. Table 1 includes the summary of features of the identity that are used for further analysis of the identity management systems.

3 Data collection

To provide an overview of the state-of-the-art of identity management systems, we have identified a number of papers and projects, that deal with the concept of the digital identity from the user's perspective.

Table 2 Search criteria.

Criterion	Characteristics
Database	ACM, AISeL, IEEE, SpringerLink
Search fields	Title, abstract, keywords
Search expression	"Digital Identity" OR "Identity Management" OR "Identity Management System"
Search period	2002-2017

In the first step, we have identified the recent research papers on identity management systems. To this end, digital libraries of ACM, AISeL, IEEE and SpringerLink were searched. The irrelevant papers have been sorted out in a multi-stage procedure. The results of the database search were assessed regarding their potential relevance by means of a particular search expression. We analysed only publications from 2002, as older ones were considered outdated. Table 2 summarizes the criteria underlying the database search.

Since the most of the identity management systems are developed outside the academia, as open-source or commercial projects, in the next step these works were also taken into account. They were identified mainly based on a list of projects available in the databases of [CORDIS, 2012, FIDIS, 2013] or listed in [Personal Data Ecosystem Consortium, 2011, Identity Commons, 2012]. The reason for applying such an approach was to ensure the topicality of our research, at the same time not focusing only on the prototypes of the IMS developed within academia, but also on the working solutions implemented in business practice.

From all of the identified identity-related efforts (both research and commercial), only those were selected, which met the criteria for the concept of the identity defined in Section 2 (Table 1). In the next step, the projects and solutions identified as relevant were compared, taking into consideration their functionality. To this end, for research projects we studied the published papers, whereas for outside-academia projects we analysed a broad range of documents, such as technical specifications, contents of projects' web pages or public wikis. This analysis allowed us to group together similar functions of different solutions and define eight high-level use cases of IMS.

Admittedly, one could claim that the conducted search cannot find all potentially relevant publications or projects and that final selection depended on the authors' subjective appraisal. Nevertheless, the search results are complete with respect to the underlying criteria and they were focused not only on identifying the finished work, but also on the work-in-progress, both regarding research and commercial projects.

4 Identity Management Systems

4.1 Definition

The identity of a user can be utilised by various processes, aiming at authentication of a person, granting authorization, providing services or supporting actions of this person [Rundle et al., 2010]. Moreover, as it was indicated in Section 2, a person may have many digital identities, used by different systems or services [Jaquet-Chiffelle et al., 2009]. This situation raises a number of issues, concerning management of personal data stored as many partial identities. The solution to this problem is to establish an Identity Management System (IMS).

In the strict sense, the identity management refers to the process of managing a virtual identity and its attributes [Hovav & Berger, 2009]. The management is facilitated by a set of technologies, procedures and standards that enable identification of a user by various services and exchange of the identity data [Barisch, 2012, OECD, 2011]. Approaches to the identity management differ in terms of management procedures (what parties use the system, who is doing what and what operations are possible on the attributes) and the types of data being stored and managed (e.g. comprehensive profiles or partial identities) [Halperin & Backhouse, 2008].

In most IMS, three main actors can be identified: a user, an Identity Provider (IdP) and a Service Provider (SP)(Relying Party). For each of them, a specific goal for using IMS can be distinguished. IdP is a party that creates, maintains and manages identities. It plays a key role in improving an interaction between users and SP, by providing services such as collecting and storing data in the identity, authentication and authorization of the user and providing identity data to SP [Beres et al., 2007]. Users use services offered by SP and may want these services to be personalised according to their specific needs. However, the development of personalized services requires exchange of information about the user, what can have significant privacy implications and put users privacy at risk [Kobsa, 2007]. Therefore, users may be unwilling to share all their personal information stored in the identity [Chen, 2007]. In some cases, they would also like to have a possibility to remain anonymous without revealing SP their actual identity [Barisch, 2012]. The main goal of SP is to attract and maintain as many users as possible. For this end, SP needs to be supplied with the identity information to perform adaptation and provide better quality services. In some cases, they can also update user's identity with information collected about the user during the usage of services [Bhargav-Spantzel et al., 2007].

Based on the most often brought up division of identity management models, the user-centric model is our focus (shown in figure 1). This model, that showcases the user control over the identity attributes and encourages users to expand their identities, has been mentioned as one of the underlying principles of the Digital Identity 3.0 defined by PwC in [Mertens & Rosemann, 2015]. The user-centric model allows the user to have control over his/her data. The user from a portfolio of identities (sometimes by a proxy of selector) allows to issue all or some attributes to SPs by explicitly giving consent to share identity information [Slamanig et al., 2014]. Although SPs act individually in this scenario, offering collaborative services is possible (but difficult) [Laurent & Bouzefrane, 2015].

Also IDaaS may be considered as a user-centric model, if the provider is not able to gain information about a user without his consent (even when SP authorizes the user and stores his/her data). Cryptographic additions onto the cloud models are proposed by [Nunez et al., 2012, Slamanig et al., 2014, Vossaert et al., 2013, Slamanig et al., 2014]. These additions can adapt IDaaS approaches to be user-centric. Taking into account the recent extensive literature on IDaaS and identity federation models, it is an interesting future research area.

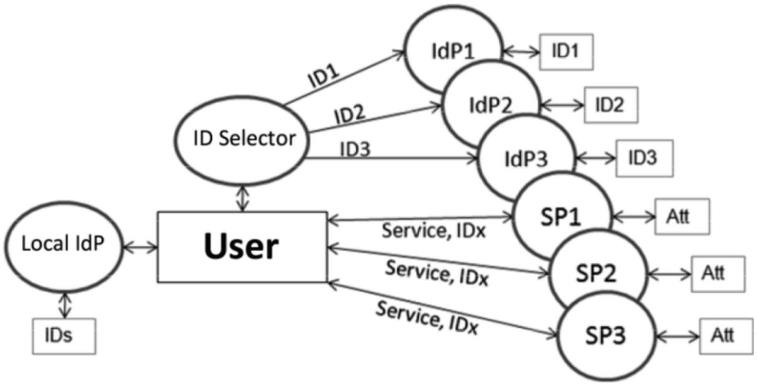


Fig. 1 User-centric model of the Identity Management.
Source: [Laurent & Bouzefrane, 2015]

Taking into account the management procedures and type of data being managed, IMS can be divided into four main types [Meints & Zwingelberg, 2009, Nabeth & Gasson, 2005b, Koch, 2002]: 1) IMS for an account management, 2) IMS for profiling of users, 3) user-controlled IMS, 4) hybrid IMS. Type 1 and 2 are normally used by large organizations, which adapt a centralised approach to the identity management and where administration processes are performed by selected IdP and by users themselves. The main focus of these systems is to assure reliable identification and authorization of a person, not privacy. The implementations of these types of IMS are mainly commercial. Type 3, in turn, is a decentralized and user-oriented IMS, where personal data is typically managed by the user. The privacy protection is important aspect in these systems. Most of them are client-side applications, developed outside a commercial sector (open source, research projects). Type 4 is a hybrid approach that joins characteristics of the types 1-3.

It needs to be underlined that not all existing systems are pure IMS, i.e. systems that implement only identity management functionality [Nabeth & Gasson, 2005b]. There are also partial solutions – systems or applications with another core functionality, but offering identity management functionality, often as an add-on (e.g. Internet browsers).

In this paper, only user-controlled (type 3 or hybrid IMS with user-oriented functions) and pure IMS are addressed. The selected solutions were analysed according to their functionality and use cases, described in the following section.

4.2 Use cases

In order to conduct the comparative analysis of the identity management solutions, eight main use cases of IMS were distinguished. These general use cases were specified based on requirements for IMS presented in research papers as well as functionality of the existing IMS. In the following sections each use case is shortly discussed.

[UC1] Maintaining a single identifier for a user

In contrast to a popular approach when IMS keeps and manages many user's identifiers for different sites [Reed et al., 2008], the maintenance of user's identifier concerns assigning a single identifier to a user's identity, valid in all services that he might use [Cameron, 2005, Bhargav-Spantzel et al., 2007]. Such an identifier can be for example an URI that points to a person's identity.

[UC2] Federating identities

A vision of a single IMS for the whole Web is difficult, if not impossible to achieve. In reality, there may be many different IMS used in different domains and each can store many local identities of a user. Therefore, there may be a need to implement a method that would allow such identities to be managed in a centralised manner and would provide interoperability between them [OASIS, 2009]. The advocates of such bottom-up approach call this vision the Federated Identity Management [Jensen, 2011, Lampropoulos et al., 2010]. The privacy and control of user attributes in such system is done either by a proxy of secure elements or trusted modules [Vossaert et al., 2013],

[UC3] Authentication of a user

The process of user authentication within IMS may be fulfilled in two different ways. On one hand, IdP may provide appropriate user's credentials (stored in the identity) for a specific SP (credential-based authorization). The second approach consists of introducing a single set of credentials, which can be used to log-in to many SP. This can be further expanded with a Single Sign-On

(SSO) mechanism, which allows to pass on an authentication status across different domains [Sun et al., 2011]. Such an approach is nowadays often offered for Facebook and GMAIL users.

[UC4] User authorization to access resources or services

The authorization is a process of determining, whether a user (in our case, an owner of an identity) is allowed to have an access to a particular resource or a service [Hodges et al., 2005]. There are several ways, how this can be accomplished. The first approach assumes that SP asks IdP for information about user's credentials and then, based on the provided credentials and internal policies, it is decided whether a given user should be granted the access. The second approach assumes conducting an authorization process without disclosing details about user's credentials to SP, and thus enables users to remain anonymous. In the third approach, an authorization is based on user's attributes asserted by the IdP [Chen, 2007].

[UC5] Storage of information about a user

An important goal of IMS is to securely store and manage the data and information about users.This may concern distinguishing many partial identities used in different contexts or by different services. The user-controlled IMS additionally assumes that users have an access to information in the identity, including ability to define types of information stored and adding/editing/deleting identity data [Ahn et al., 2007]. This also includes the right to be forgotten in line with GDPR.

[UC6] Exchange of personal information concerning a user

This use case describes sharing identity attributes with other parties. The information exchange may occur for different purposes, such as an automatic form filling, personalization of services or sharing information with other users [Koch, 2002]. The information exchange in user-controlled IMS can be supported by other functionalities. These functionalities enable users to examine, which attributes are being provided to different SP [Cameron, 2005, Angin et al., 2010]. Secondly, they allow to set access policies to the identity or its particular attributes [Leenes et al., 2008], and thus to keep privacy in the process of information exchange by data minimisation (releasing only data required by the SP for a particular service) [Claubeta et al., 2005, Chen, 2007, Ahn et al., 2007].

[UC7] Collection of information about a user

The collection of attributes' values can be realized in different ways [Zigoris & Zhang, 2006]:

- *explicit user modelling* – a user himself inputs the information,
- *implicit user modelling* – capturing information based on user's activity on the Web, using e.g. browser extensions,
- *importing* existing identities from multiple sources (Web portals, social networks, etc.),
- *engaging third parties to issue information* about a user, stored in their internal systems, after receiving permission from a user.

[UC8] Discovery of user's identity provider

In order to use the information stored in the user's identity, SP must be able to learn (discover), where the identity is stored and how to use it. During the discovery process, the SP should automatically gain information, who is the user's IdP and how to communicate with him. The popular examples of solutions for discovering IdP assume that a user inputs the proper information (e.g. URL to his identity) on every site he visits, or IdP is discovered automatically based on a single discovery service [Widdowson & Cantor, 2008].

4.3 Comparison of existing projects and solutions

The use cases described above establish a foundation of comparative analysis of diverse identity management initiatives. We have categorized these initiatives in three groups: 1) protocols and technical standards, 2) projects, in which the user-controlled IMS are developed, 3) initiatives aiming at collaborative development of user-centric identity solutions. This section presents the comparison analyses for each of the mentioned groups.

Protocols

Table 3 presents an analysis of the selected identity-related protocols.From all identified, we selected protocols that may be deployed beyond a single domain (e.g. could be used for the identity management on the Web, thus e.g. Kerberos protocol was rejected), are available as open standards and have a focus on building, storing and processing identities of regular Internet users.

Apart from the basic information about these protocols (development status, main goal and market adoption), we have analysed three dimensions, which are relevant to the topic of this article (please see Table 1):

- attribute exchange: to what extent user attributes can be exchanged via a given protocol,
- user-defined access policies: existence of mechanisms in the protocol, which enable users to control exchange of their attributes and define access rules to the identity data,
- anonymity: enabling a user to benefit from the protocol without disclosing his permanent identifier or credentials to a SP.

The presented comparison of the identity-related protocols indicates that there are already some well-established solutions available. In the future, SAML and OpenID Connect are likely to be the most popular solutions due to their market adoption and support from many organizations. The OpenID Connect is especially interesting solution, since it merges strengths of already popular OpenID 2.0 and OAuth. Still, new protocols is appearing, which aim to introduce some new functionalities to the user-centric identity management. An interesting fact is that these protocols pursue different, sometimes contradictory solutions. For example, while identity in the WebID protocol represents a wide range of user characteristics, in Persona/BrowserID the only attribute is an e-mail address, being an identifier of a user. In the next section we analyse, how these protocols are utilized by different projects.

Table 3: Comparison of the selected identity protocols

Protocol name	Development status	Main goal	Market adoption	Attribute exchange	User-defined access policies	Anonymity
OpenID [Recordon & Reed, 2006, OpenID Community, 2007, Hardt & Ferg, 2006]	first version released in 2005, version 2.0 released in 2007	providing a mechanism to prove that an end user controls an identifier	high; many portals and IdPs, among others Google, Yahoo and Flickr	attribute exchange extension; a standard set of 64 attributes (such as name, address, biography, image)	a user may choose which attributes to share with a given party	a user may present different identifiers to different Relying Parties
OAuth [Hammer-Lahav, 2011]	developed since 2007, version 1.0 released in 2010, version 2.0 in 2012	enabling a third-party application to obtain a limited access to an HTTP service without exposing the resource owner's credentials	high; used among others by Google, Last.fm etc.	users may reach any stored attribute, which belongs to or describes them; type and interpretation of attributes is beyond the scope of the protocol	a user may set detailed policies for accessing his attributes and withdraw them at any time	a user may remain anonymous - may log in and share his resources without revealing his credentials to Relying Parties
Persona (BrowserID) [Gilbert & Upatising, 2013, Team, 2013]	announced in 2011	allowing to sign in to sites using existing email addresses; IdP does not know to which sites a user is logging in	low; there are modules for some popular CMS systems (e.g. Drupal); any e-mail provider can be IdP	no attribute exchange (only an e-mail address is shared)	no access policies (no user attributes exchanged apart from an e-mail address)	a user is not anonymous; e-mail address of a user is always shared with Relying Parties

WebID (FOAF + SSL) [Story & Corlosquet, 2011, W3C Wiki, 2013, Brickley & Miller, 2010]	no final specification yet (work-in-progress, drafts published in 2014)	enabling authentication on the Web, based on the public key infrastructure; linking users and other entities into the Web of Trust and providing a new method of authorization to third party services	low; there are modules for some popular CMS systems (e.g. Drupal)	exchange of attributes from the FOAF specification (such as user relationships with other entities)	defining advanced access policies based on the WebAccessControl Ontology	a user is not anonymous; URL of user's WebID is always shared with Relying Parties, what makes it prone to tampering - some examples of counteracting were provided in [Wild et al., 2014]
SAML [Ragouzis et al., 2006, Cantor & Scavo, 2005]	developed since early 2000's; version 2.0 released in 2008	representing assertions about the authentication, attributes and authorization, using the markup language, applied mainly for a multi-domain SSO scenario	high; notable implementations include the Shibboleth system	any user attributes may be exchanged in different scenarios according to so-called SAML profiles	defining Security Policies for access to protected resources, but not from the point of view of a user and his attributes	a user may remain anonymous - attributes may be exchanged without revealing user's credentials
OpenId Connect [Sakimura et al., 2014]	version 1.0 released in 2014	provides similar mechanisms to OpenID 2.0 in more API-friendly way; might be used also by native and mobile applications	growing; currently used, among others by Google	a set of standard or additional claims about a user may be exchanged between the IdP and relying parties	a user may choose, which attributes to share with a given party	a user may present different identifiers to different relying parties

Projects

Table 4 presents the comparison of the selected identity-related projects and IMS developed within these projects. Each project has been analysed in terms of providing functionality for eight use-cases, defined in Section 4.2. In the table, we focus only on projects that were under development after year 2010. Therefore some projects, which are discontinued, such as Windows CardSpace [Chappell, 2006] and Light-Weight Identity (LID) were not included in the analysis.

From all analysed systems, only two (Higgins and Project Danube) assume assigning a single identifier to a user, which is then used by all services. In order to prove that the identity under the identifier indeed belongs to the user, a password-protected authentication or the public key infrastructure standards are used. However, this approach has a drawback, since sharing the same identifier with many sites can put users' privacy at risk [Leenes et al., 2008]. Therefore, in other systems the so-called "directed identity" approach is used [Reed et al., 2008], where IMS stores many user's identifiers and a user can choose, which identifier he wants to use for a specific site.

Moreover, within the analysed IMS, the solutions for discovering a user's identity include automatic discovery of IdP. However, in some cases (Persona, OpenPDS, PrimeLife) this functionality is supported only partially, since the discovery service is available only for those SP, who have previously integrated their services with the IMS or information about IdP must be provided by a user (Ego). In case of STORK, the IdP needs to register at IdP to provide the automatic discovery functionality.

When it comes to federating identities, only within the SWIFT project a prototype of a federated system was developed, which provides aggregation and interoperation of partial identities managed by different IdP [Lutz, D., et al., 2009]. In most of the analysed projects (Personal, Higgins, Di.me, Prime, Ego, MIA), the IMS only allows to collect in one place (by single IdP) multiple profiles used in different domains and manage them through a single interface. In case of SkIDdentity, there is an identity broker, which based on received authentication requests from SP, forwards this request to an appropriate authentication service, and then returns the received data to the calling SP. However, these local identities are not related to one another and different IdP cannot exchange the identity information, such as credentials, login status and attributes. This is mainly caused by the fact that usually an agreement between organizations in the federation is necessary, concerning a set of identifiers and attributes, which will be used by all parties to refer to the same user [Ragouzis et al., 2006]. This requirement makes it hard to scale the federation idea to the Web [Sun et al., 2010].

Among the analysed projects, predominates the usage of standardized authentication protocols, such as OAuth, OpenID or SAML (Personal, Higgins, Di.me, OpenPDS, SKIDentity, STORK) or credential-based authentication, based on data stored in the identity or provided by the user (PrimeLife, SWIFT). Within the authentication process, some of the solutions additionally support usage of nicknames (PrimeLife, SWIFT), anonymous usage of services and avoiding unintended linkability of many user's sessions (Di.me, PrimeLife Identity Mixer [IBM Research Zurich, 2011], Ego). PRIMA provides also non-impersonation property – no one except the user who is in a possession of a secret can be authorized [Asghar et al., 2016]. Only SWIFT offers its users a functionality of the single sign-on [Rajasekaran, H., et al., 2010].

Also the authorization process is accomplished in several ways. The first one assumes that the authorization is based on the user's credentials provided by the IdP and the SP decides, whether the authorization should be granted, taking into account the internal policy (PrimeLife, Di.me, Ego). Similarly, the authorization can also be based on user's attributes aggregated into the identity (SWIFT). Another approach is based on popular authorization protocols, such as SAML, Kerberos, X509 (Higgins, Personal, OpenPDS, CREDENTIAL). In case of using OAuth 2.0 protocol, the authorization can be provided without disclosing details about the user's credentials to the SP. Only personally unidentifiable information is sent in the form of an authorisation decision statement – the decision about granting an access is made by the IdP based on the credentials of the identity.

In almost all analysed systems (apart from OpenPDS), data stored within the identity can be categorised or grouped according to the different criteria. Some of the systems (Persona, Higgins, SWIFT) provide data containers, such as gems or cards, which contain either specific types of information [Gilbert & Upatising, 2013], or information about a certain aspect of user's activity (role-specific information) [OECD, 2011, Rajasekaran, H., et al., 2010]. Others provide a function to distinguish multiple partial identities, used for dealing with different SP. Besides, almost all systems enable users to view, manage (add, edit or delete information and set access policies) and control information stored in their identities. The systems that enable such functionalities are often called the Personal Data Service (PDS) [Higgins, 2013]. Some of the systems additionally provide an encryption of the data, in order to make it impossible to be read by an unauthorized third parties [Personal Inc., 2013, Higgins, 2013, Detlef et al., 2015]. An interesting solution for modelling the identity data is adapted within the Di.me and Ego systems. In the former aggregation, integration and synchronisation of user's information is both driven and supported by the Semantic Web technologies [Scerri et al., 2011]. Storage of user's data in the Ego system, in turn, is based on a semantic model, using Wikipedia categories structure [Węckowski & Małyszko, 2013]. MIA, in turn, is a mobile solution that can be used with a smartphone and provides identities for physical and electronic identification [Terbu et al., 2016].

All of the analysed systems enable an exchange of an identity information with a SP or other users, and users can control an identity and examine, which information is being provided to a SP and which parties have an access to a certain part of their identity. The systems allow a user to define and set access policies, where he can grant authorization to a specific part of his identity and thus ensure selective release of attributes (data minimisation). Moreover, some of the systems support a specific functionality, such as:

- ensuring user's privacy through an exchange of encrypted (personal) and anonymous data (SWIFT) or lack of exchange of raw identity data and providing instead only aggregated information and answers to questions (OpenPDS),
- risk detection mechanism, which warns a user of an undesirable disclosure of the sensitive data (Di.me),
- assisting a user in understanding privacy policies and providing policy enforcement via so-called "sticky policies" (cryptographical association of policies to the encrypted identity data)(PrimeLife),
- deduction which access policy should be applied based on other policies applied within a federation (SWIFT),
- circle of trust, where identity federation service and IdPs ("colleagues") trust each other and store the relevant data on colleagues' sites (STORK),
- user can give consent for data sharing in various ways: implicitly (by providing credentials), explicitly for particular data types/attributes before data is collected, explicitly with data values after data collection (SWIFT),
- an inference engine for generating proofs that prove the possession of a particular attribute without disclosing private information (PRIMA).

When it comes to collecting the identity data, the majority of the systems support explicit user modelling (Personal, Higgins, PrimeLife, SWIFT, Ego) and import of existing information and profiles from external sources (Personal, Higgins, Di.me, Project Danube, SWIFT, SkIDentity, MIA). Di.me and SWIFT additionally enable synchronisation of identity data on various user's devices. Only three systems (OpenPDS, Ego and Higgins) assume bidirectional information exchange and capturing information based on user's activity on the Web.

To sum up, the conducted review of the IMS indicates that all analysed systems provide a wide range of functionalities, encompassing all defined use-cases. Furthermore, all of them meet the criteria of the user-controlled IMS, where the control over information stored in the identity is shifted from SPs to users, in particular in terms of the possibility to decide, who has an access to particular identity attributes. A great importance is also put on assuring user's privacy, while using the IMS. On the other hand, there is a lack of interoperability between different IMS and exchange of user's information between different IdPs. The only sign for providing such interoperability in the future is the usage the existing identity-related protocols and standards within most of the analysed IMS.

Table 4: Comparison of the selected IMS

Project name	**Maintaining a single identifier [UC1]**	**Federating identities [UC2]**	**User authentication [UCIII]**	**User authorization [UCIV]**	**Storage of information [UCV]**	**Information exchange [UCVI]**	**Collection of information [UCVII]**	**Discovery of IdP [UCVIII]**
Di.me [Scerri et al., 2011, di.me project, 2010]	not supported	not supported	SALS specification for XMPP protocol or OAuth authentication; support for unlinkability	credential-based authorization	semantic data model; user-controlled multiple partial identities; private data storage	P2P exchange based on XMPP protocol; risk detection mechanism; user-controlled access policies	ontology-based synchronization; identity import	plugins/ gateways for external services
Higgins [Higgins, 2013, Lampropoulos et al., 2010, Suriadi et al., 2008, Recordon & Reed, 2006]	a single identifier, used on all webpages	not supported	single authentication, using OpenID, SAML or other protocol	token-based authorization, using WS-Trust specification and SAML, Kerberos, or X509 protocol	common data model; PDS; user-controlled context-specific data containers (cards)	user-controlled access policies and bi-directional exchange process	implicit and explicit modelling; identity import	automatic discovery (Higgins Browser Extension)
OpenPDS [Human Dynamics, MIT Media Lab, 2012, OpenPDS Project, 2012, de Montjoye et al., 2014]	not supported	not supported	OAuth 2.0 or OpenID authentication	OAuth 2.0 authorization	single-identity model; PDS; data mapped to 3 scores (Social, Activity, Focus)	no exchange of attributes or raw data; user-controlled access policies and multiple sharing levels	explicit modelling using Funf open sensing framework [The Knight Foundation, 2013]	partially supported

Personal [Personal Inc., 2013]	not supported	not supported	OAuth 2.0 authentication	OAuth 2.0 authorization	encrypted data model; PDS; data containers (gems); 2 types of data - sensitive and non-sensitive	secure sharing of encrypted data via Persona API; user-controlled access policies and exchange process	explicit modelling; identity import	partially supported
PrimeLife [Camenisch et al., 2011, Leenes et al., 2008, Project PRIME, 2008]	not supported	not supported	credential-based authentication; PKI; support for pseudonymity and unlinkability	credential-based authorization	user-controlled multiple partial identities	user-controlled exchange process; access policy negotiation; sticky policies; released data tracking	explicit modelling	partially supported
Project Danube [Project Danube, 2012, Markus Sabadello, 2011]	a single URI-identifier (IName)	interoperability between IdP	support for many authentication protocols (i.e. OAuth, OpenId, OStatus)	link contracts for the XDI protocol	PDS built on XRI and XDI technology; user-controlled multiple partial identities	PKI-based exchange; access control based on link contracts	identity import and mapping to XDI	automatic discovery of the XDI endpoint (XRI Resolution, Webfinger or LRDD)
SWIFT [SWIFT Consortium, 2009, Lutz, D., et al., 2010, Rajasekaran, H., et al., 2010, Lutz, D., et al., 2009]	not supported	aggregation and interoperation of many partial identities, managed by different IdP	credential-based authentication; SSO; support for pseudonymity	attribute-based authorization	user-controlled context- and role-specific data containers (cards)	user-controlled exchange process; anonymous attribute transfer; XACML-based deductive policies for the federation	explicit modelling; identity import and aggregation; identity transfer between devices	automatic discovery using XRI and XRDS

Ego [Węckowski & Małyszko, 2013, Filipowska & Małyszko, 2013]	not supported (to ensure unlinkability of user sessions)	not supported	custom authentication protocol with a focus on support for unlinkability; the SP only receives a temporary identifier of the user	not supported	semantic data model, based on Wikipedia categories structure	user-controlled access policies for different parts of the user model	explicit modelling; implicit modelling performed jointly by different SPs; IdP's module, controlling the collection process	partially supported (a user manually provides the SP with the URL to the IdP)
SkIDentity [Detlef et al., 2015]	not supported	partially supported in the form of identity broker	credential-based authentication; support for various protocols (SAML, OpenID, OAuth) and e-ID tokens; support for pseudonymity;	not supported	user-controlled cloud identity; storage of encrypted data; decentralized storage of an identity on the user's system	user-controlled exchange process	identity import and transfer between devices	automatic discovery
STORK [Heppe et al., 2011]	a single identifier	partially supported via STORK federated eID service	SAML-based authentication	support digital signature; certificate validation based on OCSP gateway	data stored by IdP	user-controlled access policies and multiple sharing levels	information issued by third parties	supported based on URL; circle of trust concept
FutureID [Bruegger & Roßnagel, 2016]	not supported	partially supported as intermediate service between credential technologies and SP	credential-based authentication; support for many authentication protocols	not supported	not supported	not supported	user-controlled access policy	decentralized discovery mechanism

My Identity App (MIA) [Terbu et al., 2016]	not supported	aggregation of partial identities managed by different IdP	credential-based authentication; support for various protocols (SAML, OpenID, OAuth)	time-limited, one-time tokens; barcode, QR or short string tokens	mobile ID solution; centralized data storage	data exchange after verification of permissions; access policies; exchange of encrypted data	identity import	automatic discovery
PRIMA [Asghar et al., 2016]	not supported		credential-based authentication using SAML and OpenID; support for unlinkability and non-impersonation		locally stored credentials; no interaction with IdP needed to consume services	user-controlled access policies; inference engine for attribute proofs	explicit modeling and identity import	partially supported (trust relation between SP and IdP needed)
CREDENTIAL [Hörandner et al., 2016]	not supported	not supported	custom authentication protocol based on proxy re-encryption and redactable signatures; hardware-based multi-factor mechanisms	not supported	user-controlled cloud identity; single identity model; storage of encrypted data based on custom cryptographic mechanisms	encrypted data sharing; selective disclosure mechanisms; user-controlled access policies; data exchange using SAML or OpenID Connect	explicit modelling	not supported (user is delegated to IdP)

Initiatives

Apart from protocols and projects, the identity management systems are also a subject of interest of several initiatives, which intend to consolidate and coordinate efforts in the area of the identity management. These initiatives gather people and organisations interested in defining requirements for IMS and development of identity-related technologies. Since most of the work in these initiatives happens within the work groups, we compared areas of interest of these groups and related them to our use-cases. The working groups in the well-known initiatives are presented in Table 5. In our analysis only these working groups were taken into account, which are active and advocate the user-centric approach to the identity management.

The Identity Commons is the only organisation, whose activities cover all use cases defined. The working groups within Identity Commons generally aim at creating and developing an interoperable, universally-adopted user-centric identity layer for the Internet [Identity Commons Working Groups, 2013]. Moreover, Identity Commons assists efforts to create transparency in the operations of the identity systems and their associated services. The works within this initiative cover a wide spectrum of activities, from providing an Internet-scale identity interoperability (OSIS) and establishing persistent, privacy- protected identities (XDI.org), through defining the semantics of an identity and machine-readable description of attributes (Identity Schemas), creation of tools and protocols that help users to supervise an exchange of the identity data (Project VRM, Personal Data Ecosystem Consortium, Higgins Project), up to promoting solutions related to the identity and the law in this regard (ID-Legal).

Another well-known organisation is the Kantara Initiative, which mission is to ensure secure, identity-based interactions, while preventing misuse of the personal information. Therefore, its activities focus on collecting requirements for the development and operation of the Identity Assurance Framework [Wasley & Brennan, 2012] and verification of identity providers. The works within the Kantara Work and Discussion Groups [Kantara Initiative, 2013b] concern defining a global framework and requirements for federation of identity (Federation Interoperability, Identity Assurance, Business Cases for Trusted Federations), handling identity's attributes (Attributes in Motion), increasing user's control over the authorization of the exchanged data, as well as defining access policies for storing user's personal information (Information Sharing, User Managed Access).

The works of [OpenID Foundation, 2013] focus on promoting and enabling OpenID technologies among users and facilitate uptake of the OpenID solution in such fields as logging into a website (Account Chooser), enabling a Single Sign On (Native Applications), securing authorization and authentication process, sharing and collecting the identity information (OpenID Connect, Backplane Protocol) and discovering the Identity Provider (OpenID Connect).

[Kerberos Consortium, 2013a] in turn, aims at improving Kerberos authentication and authorization solutions and providing a secured, universal single sign-on environment for federated realms (MIT Kerberos Team). The Kerberos Consortium is also involved in developing new technologies for personal data ecosystems, in which people can manage their personal data more efficiently and equitably (MIT-KIT WG).

Apart from the initiatives presented in Table 5, it is worth to mention the FIDO Alliance [FIDO Alliance, 2017] that focuses mainly on authentication solutions that might be used in IMS. FIDO is an international alliance whose aim is to provide an interoperable ecosystem of hardware-, mobile- and biometrics-based authenticators that can be used with many apps and websites. They develop technical specifications that define open and interoperable mechanisms and work on specification on new, formal standards for security devices and browser plugins.

The comparison of activities within different identity-related initiatives shows that on one hand, the ongoing works cover all general use-cases and functionality of IMS. On other hand, there are many groups with overlapping research areas, handling similar issues. Admittedly, some of them cooperate (for example, MIT-KIT and IMA or Information Sharing, which work both under Kantara Initiatives and Identity Commons) and use the existing identity standards, frameworks and best practices. However, there is still a lack of coordination of activities, and exchange of knowledge and experience between different initiatives.

Table 5: Comparison of the selected identity-related initiatives with regard to the coverage of the selected IMS use cases and the working groups' interests.

Project name	**Maintaining a single identifier [UC1]**	**Federating identities [UC2]**	**User authentication [UCIII]**	**User authorization [UCIV]**	**Storage of information [UCV]**	**Information exchange [UCVI]**	**Collection of information [UCVII]**	**Discovery of IdP [UCVIII]**
Identity Commons [Identity Commons, 2013]	OSIS	Project VRM, Personal Data Ecosystem Consortium, ID-Legal	OSIS, XDI.org	OSIS, XDI.org	OSIS, Project VRM, Personal Data Ecosystem Consortium, XDI.org, Higgins Project, Identity Schemas, ID-Legal	OSIS, Project VRM, Personal Data Ecosystem Consortium, XDI.org, Higgins Project, ID-Legal	OSIS, Project VRM, Personal Data Ecosystem Consortium, XDI.org, Higgins Project, ID-Legal	OSIS, XDI.org, ID-Legal
Kantara Initiative [Kantara Initiative, 2013a]		Business Cases for Trusted Federations, Federation Interoperability, Identity Assurance		User Managed Access	Attributes in Motion, Information Sharing, User Managed Access	Attributes in Motion, Information Sharing, User Managed Access	User Managed Access	User Managed Access
OpenId Foundation [OpenID Foundation, 2013]	OpenID Connect, Account Chooser		OpenID Connect, Account Chooser	OpenID Connect, Native Applications		OpenID Connect, Backplane Protocol	Backplane Protocol	OpenID Connect
Kerberos Consortium [Kerberos Consortium, 2013b]		MIT Kerberos	MIT Kerberos, MIT-KIT	MIT Kerberos, MIT-KIT		MIT-KIT		MIT-KIT

5 Summary and Outlook

The goal of the paper was to deliver the survey of solutions in the field of the digital identity and to compare these solutions from the perspective of use cases emphasizing the most popular application scenarios of the identity management systems. Starting from a general summary of the digital identity definitions, we examined the selected identity-related protocols, projects and initiatives. Being on different stages of development, the analysed projects and solutions gave us an overview of the current and planned approaches to solve the different issues related to IMS.

From all analysed protocols and IMS, only some have relatively high market adoption. Among the protocols, these are SAML and OpenID Connect, which merge advantages of other popular protocols, namely OpenID 2.0 and OAuth. Among the IMS, the most popular are Personal and OpenPDS. Personal provides tools and APIs for companies to integrate it into their websites and applications, thus enabling data import from popular social portals, like Facebook or LinkedIn. Personal provides also connection to over 300 institutions to fetch bills, statements and digital receipts and is offered as a Web, mobile and cloud-based solution [Personal Inc., 2013]. Moreover, it supports most of the distinguished use cases, therefore has the highest potential to become more widely used on the Web. The latter – OpendPDS – is an open-source solution, available via a mobile application. It is still under development and its coverage may expand significantly, since common libraries for OpenPDS are planned. However, usage of an openPDS-based solution on a device without appropriate resources for fast data processing and generating answers (data aggregation) might be impracticable, since latency in data processing might be unacceptable for users.

The works on other identity management solutions, such as Higgins and Windows CardSpace, despite intensive support from the community in the past and a technology contribution from well-known companies, such as Microsoft, IBM, Novell, Oracle, and Google, recently were significantly reduced or even stopped. Similarly it is with the IMS being results of research projects (Di.me, Swift, PrimeLife, Ego) and which have not been implemented in real life. The exemption is Di.me, which was validated as a platform for attendees of conferences and events. Its features are highly valued in the context of event organization, and thus has a potential to be adopted within the event management sector [Alonso et al., 2013].

The analysis conducted within this paper revealed that most of these issues relate to the development of user-oriented solutions in the digital identity domain and are either already solved (at least partially), or are a topic of interest of the ongoing research. Basic functionalities of IMS, such as best practices regarding authentication or authorization, are already well-developed and adopted. However, an extensive research is still needed in several fields. Firstly, the analysis of information exchange between an identity and its environment should be researched, especially when it comes to providing an interoperability between different systems working within an identities' federation. It is important to analyse types of information that IMS should store and share, parties involved in this process as well as possible ways of conducting the exchange, while, at the same time, providing a user with control over this process. There is therefore a need for a model of an identity data exchange, which from the one hand can be used by many different systems and by different service providers, and on the other hand would ensure users that privacy of their data is not at risk (by providing a secure data exchange and if desired, an ability to remain anonymous).

Secondly, the research on possible solutions for user modelling in terms of building an identity and using it in different contexts is also needed. In this aspect, it needs to be taken into account that users aim at being able to use services adapted to their needs, but at the same time they do not want to worry about their privacy [Krasnova et al., 2009]. However, in order to perform personalization, a large amount of personal information needs to be collected, what can have significant privacy implications [Kobsa, 2007]. The privacy and security issues should be main aspects taken into account, while building a user's identity. Moreover, as users tend to work on more devices and use their personal data in different contexts and roles, traditional profiling methods became insufficient. There is also a need for a secure identity model that firstly, would be built based both on information provided by a user himself and his activities on the Web, and secondly, would be able to distinguish and group user's attributes with regard to different contexts and roles, in which the identity may be used. These challenges have a great potential for implementation in practice and should be taken into account in the course of future projects.

Acknowledgements The work published in this article was partially supported by the project titled: "Ego – Virtual Identity", financed by the Polish National Centre of Research and Development (NCBiR) (contract no. NR11-0037-10).

References

[Ahn et al., 2007] Ahn, Jae-wook, Peter Brusilovsky, Jonathan Grady, Daqing He, & Sue Yeon Syn 2007. Open user profiles for adaptive news systems: help or harm? In Proc 16th Int. Conf. on World Wide Web, pages 11–20, New York. ACM.

[Alonso et al., 2013] Alonso, David, Sophie Wrobel, & Richard Wacker 2013. White paper: Using di.me digital.me on Business Conferences and Smart Events. `http://www.dime-project.eu/DescargarDocumento.aspx?idd=5206`.

[Angin et al., 2010] Angin, P., B. Bhargava, R. Ranchal, N. Singh, M. Linderman, L.B. Othmane, & L. Lilien 2010. An Entity-Centric Approach for Privacy and Identity Management in Cloud Computing. In 29th Symp. on Reliable Distributed Systems, pages 177–183, New York. IEEE.

[Arora, 2008] Arora, Siddhartha 2008. National e-ID card schemes: A European overview. Information Security Technical Report, 13(2):46–53.

[Asghar et al., 2016] Asghar, Muhammad Rizwan, Michael Backes, & Milivoj Simeonovski 2016. PRIMA: Privacy-Preserving Identity and Access Management at Internet-Scale. arXiv preprint arXiv:1612.01787.

[Banihashemi et al., 2016] Banihashemi, Sepideh, Alireza Talebpour, Elaheh Homayounvala, & Abdolreza Abhari 2016. Identifying and Prioritizing Evaluation Criteria for User-

Centric Digital Identity Management Systems. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 7(7):45–54.

[Barisch, 2009] Barisch, Marc 2009. Modelling the impact of virtual identities on communication infrastructures. In Proc 5th ACM workshop on Digital identity management, pages 45–52, New York. ACM.

[Barisch, 2012] Barisch, Marc 2012. Design and Evaluation of a Sytem to Extend Identity Management to Multiple Devices. PhD Thesis, Universitat Stuttgart.

[Bassi & Horn, 2008] Bassi, Alessandro, & Geir Horn 2008. Internet of Things in 2020, Roadmap for the Future, Version 1.1. http://www.iot-visitthefuture.eu/fileadmin/documents/researchforeurope/270808_IoT_in_2020_Workshop_Report_V1-1.pdf (2013-10-23).

[Benantar, 2006] Benantar, Messaoud 2006. Access control systems: security, identity management and trust models. Springer Science & Business Media.

[Beres et al., 2007] Beres, Yolanta, Adrian Baldwin, Marco Casassa Mont, & Simon Shiu 2007. On identity assurance in the presence of federated identity management systems. In Proc 2007 ACM workshop on Digital identity management, pages 27–35, New York. ACM.

[Bernstein et al., 2011] Bernstein, M.S., A. Monroy-Hernandez, D. Harry, P. Andre, K. Panovich, & G. Vargas 2011. An Analysis of Anonymity and Ephemerality in a Large Online Community. In 5th Int. Conf. on Weblogs and Social Media, Menlo Park. The AAAI Press.

[Bhargav-Spantzel et al., 2007] Bhargav-Spantzel, Abhilasha, Jan Camenisch, Thomas Gross, & Dieter Sommer 2007. User centricity: A taxonomy and open issues. Journal of Computer Security, 15(5):493–527.

[Brickley & Miller, 2010] Brickley, Dan, & Libby Miller 2010. FOAF vocabulary specification 0.98. http://xmlns.com/foaf/spec/20100809.html (2013-11-20).

[Bruegger & Roßnagel, 2016] Bruegger, Bud P, & Heiko Roßnagel 2016. Towards a Decentralized Identity Management Ecosystem for Europe and Beyond. In Open Identity Summit, pages 55–66.

[Camenisch et al., 2011] Camenisch, Jan, Simone Fischer-Hübner, & Kai Rannenberg (eds) 2011. Privacy and Identity Management for Life. Springer.

[Cameron, 2005] Cameron, Kim 2005. The laws of identity. http://msdn.microsoft.com/en-us/library/ms996456.aspx (2013-08-14).

[Cantor & Scavo, 2005] Cantor, Scott, & Tom Scavo 2005. Shibboleth architecture. http://shibboleth.internet2.edu/shibboleth-documents.html (2013-11-20).

[Chappell, 2006] Chappell, David 2006. Introducing Windows CardSpace. https://msdn.microsoft.com/en-us/library/aa480189.aspx.

[Chen, 2007] Chen, Zhikui 2007. A Privacy Enabled Service Authorization Based on a User-centric Virtual Identity Management System. In 2nd Int. Conf. on Communications and Networking in China, pages 423–427, New York. IEEE.

[Clarke, 2008] Clarke, Roger 2008. Dissidentity. Identity in the Information Society, 1(1):221–228.

[Claubeta et al., 2005] Claubeta, Sebastian, Dogan Kesdogan, & Tobias Kolsch 2005. Privacy enhancing identity management: protection against re-identification and profiling. In Proc 2005 workshop on Digital identity management, pages 84–93, New York. ACM.

[CORDIS, 2012] CORDIS 2012. Community Research and Development Information Service. http://cordis.europa.eu/home_en.html (2013-08-12).

[de Montjoye et al., 2014] de Montjoye, Yves-Alexandre, Erez Shmueli, Samuel S Wang, & Alex Sandy Pentland 2014. openpds: Protecting the privacy of metadata through safeanswers. PloS one, 9(7):e98790.

[Detlef et al., 2015] Detlef, Hühnlein, Max Tuengerthal, Tobias Wich, Tina Hühnlein, & Benedikt Biallowons 2015. Innovative Building Blocks for Versatile Authentication within the SkIDentity Service. In et al., Detlef Hühnlein (ed), Open Identity Summit 2015, pages 141–152. Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn.

[di.me project, 2010] di.me project 2010. . http://www.dime-project.eu/ (2013-11-14).

[Ferdous & Poet, 2012] Ferdous, Md Sadek, & Ron Poet 2012. A comparative analysis of Identity Management Systems. In High Performance Computing and Simulation, 2012 Int. Conf. on, pages 454–461. IEEE.

[FIDIS, 2013] FIDIS 2013. The FIDIS database on identity management systems. http://imsdb.fidis.net (2013-10-28).

[FIDO Alliance, 2017] FIDO Alliance 2017. https://fidoalliance.org (2017-12-08).

[Filipowska & Małyszko, 2013] Filipowska, Agata, & Jacek Małyszko 2013. Towards using Wikipedia for Building User Identities. In Proceedings of the NLP and DBpedia workshop, Sydney, Australia.

[Gautam et al., 2016] Gautam, Bhaskar, Vivek Jain, Sourabh Jain, & B Annappa 2016. Profile Matching of Online Social Network with Aadhaar Unique Identification Number. In Cloud Computing in Emerging Markets (CCEM), 2016 IEEE International Conference on, pages 168–169. IEEE.

[Gilbert & Upatising, 2013] Gilbert, Connor, & Laza Upatising 2013. Formal analysis of BrowserID/Mozilla Persona. http://www.stanford.edu/~lazau/BrowserIDPersona.pdf (2013-11-20).

[Habiba et al., 2014] Habiba, Umme, Rahat Masood, Muhammad Awais Shibli, & Muaz A Niazi 2014. Cloud identity management security issues & solutions: a taxonomy. Complex Adaptive Systems Modeling, 2(1):5.

[Halperin & Backhouse, 2008] Halperin, Ruth, & James Backhouse 2008. A roadmap for research on identity in the information society. Identity in the Information Society, 1(1):71–87.

[Hammer-Lahav, 2011] Hammer-Lahav, E. 2011. The OAuth 1.0 Protocol. http://tools.ietf.org/html/rfc5849 (2013-11-21).

[Hardt & Ferg, 2006] Hardt, D, & B Ferg 2006. Attribute Properties for OpenID Attribute Exchange. http://openid.net/specs/openid-attribute-properties-list-1_0-01.html (2013-11-20).

[Heppe et al., 2011] Heppe, J, T Schnedier, H Leitold, & R Portela 2011. STORK Overview for new MS. STORK document, November.

[Higgins, 2013] Higgins 2013. PDS Vision. http://wiki.eclipse.org/PDS_Vision (2013-11-14).

[Higgins, 2013] Higgins 2013. Personal Data Service (PDS). http://eclipse.org/higgins/ (2013-08-12).

[Hildebrandt, 2009] Hildebrandt, Mireille 2009. Profiling and AmI. In The future of identity in the information society, pages 273–310. Springer.

[Hildebrandt & Backhouse, 2005] Hildebrandt, Mireille, & James Backhouse 2005. D7.2: Descriptive analysis and inventory of profiling practices. FIDIS Project Deliverable. http://www.cosic.esat.kuleuven.be/publications/article-827.pdf (2013-11-05).

[Hodges et al., 2005] Hodges, Jeff, Rob Philpott, & Eve Maler 2005. Glossary for the OASIS Security Assertion Markup Language (SAML) V2.0. http://docs.oasis-open.org/security/saml/v2.0/saml-glossary-2.0-os.pdf (2013-11-05).

[Hörandner et al., 2016] Hörandner, Felix, Stephan Krenn, Andrea Migliavacca, Florian Thiemer, & Bernd Zwattendorfer 2016. CREDENTIAL: a framework for privacy-preserving cloud-based data sharing. In Availability, Reliability and Security (ARES), 2016 11th International Conference on, pages 742–749. IEEE.

[Hovav & Berger, 2009] Hovav, Anat, & Ron Berger 2009. Tutorial: Identity Management Systems and Secured Access Control. Communications of the Association for Information Systems, 25(1):42.

[Hühnlein et al., 2013] Hühnlein, Detlef, Jörg Schwenk, Tobias Wich, Vladislav Mladenov, Florian Feldmann, Andreas Mayer, Johannes Schmölz, Bud Bruegger, & Moritz Horsch 2013. Options for integrating eid and saml. In Proceedings of the 2013 ACM workshop on Digital identity management, pages 85–96. ACM.

[Human Dynamics, MIT Media Lab, 2012] Human Dynamics, MIT Media Lab 2012. OpenPDS Software. http://idcubed.org/wp-content/uploads/2012/11/OpenPDS-software-from-Human-Dynamics.pdf (2013-11-14).

[IBM Research Zurich, 2011] IBM Research Zurich 2011. Identity mixer. http://primelife.ercim.eu/results/opensource/55-identity-mixer (2013-11-18).

[Identity Commons, 2012] Identity Commons 2012. Identity landscape. http://wiki.idcommons.net/Identity_Landscape (2013-08-12).

[Identity Commons, 2013] Identity Commons 2013. http://www.idcommons.org/ (2013-08-12).

[Identity Commons Working Groups, 2013] Identity Commons Working Groups 2013. Project Wiki. http://wiki.idcommons.net/Working_Group_Descriptions (2013-11-05).

[Jaquet-Chiffelle et al., 2009] Jaquet-Chiffelle, David-Olivier, Emmanuel Benoist, Rolf Haenni, Florent Wenger, & Harald Zwingelberg 2009. Virtual Persons and Identities. In Rannenberg, Kai, Denis Royer, & André Deuker (eds), The Future of Identity in the Information Society, pages 75–122. Springer Berlin Heidelberg.

[Jensen, 2011] Jensen, Jostein 2011. Benefits of Federated Identity Management - A Survey from an Integrated Operations Viewpoint. In Tjoa, A, Gerald Quirchmayr, Ilsun You, & Lida Xu (eds), Availability, Reliability and Security for Business, Enterprise and Health Information Systems, volume 6908 of *LNCS*, pages 1–12. Springer, Berlin.

[Kantara Initiative, 2013a] Kantara Initiative 2013a. http://kantarainitiative.org/ (2013-08-12).

[Kantara Initiative, 2013b] Kantara Initiative 2013b. Work and Discussion Groups. http://kantarainitiative.org/confluence/pages/viewpage.action?pageId=38371527 (2013-11-05).

[Kerberos Consortium, 2013a] Kerberos Consortium 2013a. MIT Kerberos & Internet Trust (MIT-KIT). http://kit.mit.edu/ (2013-11-06).

[Kerberos Consortium, 2013b] Kerberos Consortium 2013b. MIT Kerberos Project. http://www.kerberos.org/ (2013-11-06).

[Kobsa, 2007] Kobsa, Alfred 2007. Privacy-Enhanced Web Personalization. In Brusilovsky, Peter, Alfred Kobsa, & Wolfgang Nejdl (eds), The Adaptive Web, volume 4321 of *LNCS*, pages 628–670. Springer, Berlin.

[Koch, 2002] Koch, M. 2002. Global identity management to boost personalization. In Proc 9th Research Symp. on emerging Electronic Markets, pages 137–147, Basel. University of Applied Sciences Basel.

[Krasnova et al., 2009] Krasnova, Hanna, Oliver Günther, Sarah Spiekermann, & Ksenia Koroleva 2009. Privacy concerns and identity in online social networks. Identity in the Information Society, 2(1):39–63.

[Kubicek, 2010] Kubicek, Herbert 2010. Introduction: conceptual framework and research design for a comparative analysis of national eID Management Systems in selected European countries. Identity in the Information Society, 3(1):5–26.

[Lampropoulos et al., 2010] Lampropoulos, Konstantinos, Daniel Diaz-Sanchez, Florina Almenares, Peter Weik, & Spyros Denazis 2010. Introducing a cross federation identity solution for converged network environments. In Principles, Systems and Applications of IP Telecommunications, pages 1–11, New York. ACM.

[Laurent & Bouzefrane, 2015] Laurent, Maryline, & Samia Bouzefrane 2015. Digital identity management. Elsevier.

[Leenes et al., 2008] Leenes, R., J. Schallaböck, & M. Hansen 2008. PRIME white paper. http://security.future-internet.eu/images/2/27/Prime_White.pdf (2013-08-14).

[Lutz, D., et al., 2009] Lutz, D., et al. 2009. SWIFT Deliverable D503. Prototype Specification . http://www-wordpress.sit.fraunhofer.de/ist-swift/wp-content/uploads/sites/10/2013/10/D503.pdf (2013-11-14).

[Lutz, D., et al., 2010] Lutz, D., et al. 2010. SWIFT Deliverable D504. Simulation, Modelling and Prototypes . http://www-wordpress.sit.fraunhofer.de/ist-swift/wp-content/uploads/sites/10/2013/10/D504.pdf (2013-11-14).

[Manso et al., 2014] Manso, Oscar, Morten Christiansen, & Gert Mikkelsen 2014. Comparative analysis - Web-based Identity Management Systems. http://alexandra.dk/sites/default/files/downloads/It-sikkerhed/Authorization_Systems_Comparative_AI.pdf.

[Markus Sabadello, 2011] Markus Sabadello 2011. Project Danube. http://d-cent.org/fsw2011/wp-content/uploads/fsw2011-Project-Danube.pdf (2013-11-14).

[Meints & Zwingelberg, 2009] Meints, Martin, & Harald Zwingelberg 2009. D3.17 Identity Management Systems – recent developments. FIDIS Project Deliverable . http://www.fidis.net/resources/fidis-deliverables/hightechid/#c1787 (2013-11-27).

[Mertens & Rosemann, 2015] Mertens, Willem, & Michael Rosemann 2015. Digital Identity 3.0: The Platform for People.

[Miorandi et al., 2012] Miorandi, Daniele, Sabrina Sicari, Francesco De Pellegrini, & Imrich Chlamtac 2012. Internet of things: Vision, applications and research challenges. Ad Hoc Networks, 10(7):1497 – 1516.

[Mpofu & Van Staden, 2014] Mpofu, Nkosinathi, & Wynand Jc Van Staden 2014. A survey of trust issues constraining the growth of Identity Management-as-a-Service (IdMaaS). In Information Security for South Africa (ISSA), 2014, pages 1–6. IEEE.

[Nabeth, 2009] Nabeth, Thierry 2009. Identity of Identity. In Rannenberg, Kai, Denis Royer, & André Deuker (eds), The Future of Identity in the Information Society, pages 19–69. Springer Berlin Heidelberg.

[Nabeth & Gasson, 2005a] Nabeth, Thierry, & Mark Gasson 2005a. D2.3: Models. FIDIS Project Deliverable.

[Nabeth & Gasson, 2005b] Nabeth, Thierry, & Mark Gasson 2005b. D3.1: Structured overview on prototypes and concepts of identity management systems. FIDIS Project Deliverable. http://www.fidis.net/resources/fidis-deliverables/hightechid/int-d3100/ (2013-11-05).

[Nunez et al., 2012] Nunez, David, Isaac Agudo, & Javier Lopez 2012. Integrating OpenID with proxy re-encryption to enhance privacy in cloud-based identity services. In Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, pages 241–248. IEEE.

[OASIS, 2009] OASIS 2009. Identity Metasystem Interoperability Version 1.0. http://docs.oasis-open.org/imi/identity/v1.0/identity.html (2013-11-20).

[OECD, 2011] OECD 2011. Digital Identity Management. Enabling Innovation and Trust in the Internet Economy. http://www.oecd.org/sti/ieconomy/49338380.pdf (2013-08-10).

[OpenID Community, 2007] OpenID Community 2007. OpenID Authentication 2.0 - Final. http://openid.net/specs/openid-authentication-2_0.html (2013-11-20).

[OpenID Foundation, 2013] OpenID Foundation 2013. Work Groups of OpenID Foundation. http://openid.net/wg/ (2013-08-12).

[OpenPDS Project, 2012] OpenPDS Project 2012. . http://openpds.media.mit.edu/ (2013-11-14).

[Personal Data Ecosystem Consortium, 2011] Personal Data Ecosystem Consortium 2011. The startup circle. http://personaldataecosystem.org/2011/06/startup/ (2013-08-12).

[Personal Inc., 2013] Personal Inc. 2013. Personal system. https://www.personal.com/ (2013-08-12).

[Project Danube, 2012] Project Danube 2012. Identity and communication for political and social innovation. http://projectdanube.org/ (2013-08-12).

[Project PRIME, 2008] Project PRIME 2008. Privacy and Identity Management for Europe. https://www.prime-project.eu/ (2013-08-12).

[Ragouzis et al., 2006] Ragouzis, Nick, John Hughes, Rob Philpott, & Eve Maler 2006. Security Assertion Markup Language (SAML) V2.0 Technical Overview. http://www.oasis-open.org/committees/documents.php?wg_abbrev=security (2013-11-20).

[Rajasekaran, H., et al., 2010] Rajasekaran, H., et al. 2010. SWIFT White Paper. SWIFT Identity Architecture. http://www-wordpress.sit.fraunhofer.de/ist-swift/wp-content/uploads/sites/10/2013/10/Whitepaper-SWIFT_Identity_Architecture.pdf (2013-11-14).

[Rannenberg et al., 2009a] Rannenberg, Kai, Denis Royer, & Andr Deuker 2009a. The Future of Identity in the Information Society: Challenges and Opportunities. Springer Berlin Heidelberg.

[Rannenberg et al., 2009b] Rannenberg, Kai, Denis Royer, & André Deuker 2009b. Introduction. In Rannenberg, Kai, Denis Royer, & André Deuker (eds), The Future of Identity in the Information Society, pages 1–11. Springer Berlin Heidelberg.

[Recordon & Reed, 2006] Recordon, David, & Drummond Reed 2006. OpenID 2.0: a platform for user-centric identity management. In Proc 2nd ACM workshop on Digital identity management, pages 11–16. ACM.

[Reed et al., 2008] Reed, D., L. Chasen, & W. Tan 2008. OpenID identity discovery with XRI and XRDS. In Proc 7th symposium on Identity and trust on the Internet, pages 19–25, New York.

[Royer et al., 2009] Royer, Denis, André Deuker, & Kai Rannenberg 2009. Mobility and Identity. In The Future of Identity in the Information Society, pages 195–242. Springer.

[Rundle et al., 2010] Rundle, Mary, Eve Maler, Anthony Nadalin, Drummond Reed, & Don Thibeau 2010. The Open Identity Trust Framework Model. http://openidentityexchange.org/sites/default/files/the-open-identity-trust-framework-model-2010-03.pdf (2013-08-14).

[Sakimura et al., 2014] Sakimura, Natsuhiko, J Bradley, M Jones, B de Medeiros, & C Mortimore 2014. Openid connect core 1.0. http://openid.net/specs/openid-connect-core-1_0.html.

[Satchell et al., 2006] Satchell, Christine, Graeme Shanks, Steve Howard, & John Murphy 2006. Beyond security: implications for the future of federated digital identity management systems. In Proc 18th Australia Conf. on Computer-Human Interaction: Design: Activities, Artefacts and Environments, pages 313–316, New York. ACM.

[Scerri et al., 2011] Scerri, Simon, Rafael Gimenez, Fabian Herman, Mohamed Bourimi, & Simon Thiel 2011. Digital.Me towards an integrated Personal Information Sphere.

[Slamanig et al., 2014] Slamanig, Daniel, Klaus Stranacher, & Bernd Zwattendorfer 2014. User-centric identity as a service-architecture for eIDs with selective attribute disclosure. In Proceedings of the 19th ACM symposium on Access control models and technologies, pages 153–164. ACM.

[Story & Corlosquet, 2011] Story, Henry, & Stephane Corlosquet 2011. Web 1.0. Web Identification and Discovery. http://www.w3.org/2005/Incubator/webid/spec/ (2013-11-20).

[Sun et al., 2010] Sun, San-Tsai, Yazan Boshmaf, Kirstie Hawkey, & Konstantin Beznosov 2010. A billion keys, but few locks: the crisis of web single sign-on. In Proc 2010 workshop on new security paradigms, pages 61–72, New York. ACM.

[Sun et al., 2011] Sun, San-Tsai, Eric Pospisil, Ildar Muslukhov, Nuray Dindar, Kirstie Hawkey, & Konstantin Beznosov 2011. What makes users refuse web single sign-on?: an empirical investigation of OpenID. In Proc 7th Symposium on Usable Privacy and Security, pages 4:1–4:20, New York. ACM.

[Suriadi et al., 2008] Suriadi, Suriadi, Ernest Foo, & Rong Du 2008. Layered identity infrastructure model for identity meta systems. In Proc 6th Australasian Conf. on Information security, pages 83–92, Darlinghurst. Australian Computer Society, Inc.

[SWIFT Consortium, 2009] SWIFT Consortium 2009. SWIFT (Secure Widespread Identities for Federated Telecommunications) Project . http://www.ist-swift.org/content/blogcategory/35/46/ (2013-11-14).

[Team, 2013] Team, Mozilla Identity 2013. BrowserID Protocol. https://github.com/mozilla/id-specs/blob/prod/browserid/index.md (2013-11-20).

[Terbu et al., 2016] Terbu, Oliver, Stefan Vogl, & Sebastian Zehetbauer 2016. One mobile ID to secure physical and digital identity. In Open Identity Summit, pages 43–54.

[The Knight Foundation, 2013] The Knight Foundation 2013. The Funf Open Sensing Framework. http://www.funf.org/about.html (2013-11-14).

[Tormo et al., 2013] Tormo, Ginés Dólera, Félix Gómez Mármol, Joao Girao, & Gregorio Martinez Perez 2013. Identity Management–In Privacy We Trust: Bridging the Trust Gap in eHealth Environments. IEEE Security & Privacy, 11(6):34–41.

[Torres et al., 2013] Torres, Jenny, Michele Nogueira, & Guy Pujolle 2013. A survey on identity management for the future network. Communications Surveys & Tutorials, IEEE, 15(2):787–802.

[Vossaert et al., 2013] Vossaert, Jan, Jorn Lapon, Bart De Decker, & Vincent Naessens 2013. User-centric identity management using trusted modules. Mathematical and Computer Modelling, 57(7):1592–1605.

[W3C Wiki, 2013] W3C Wiki 2013. WebAccessControl. http://www.w3.org/wiki/WebAccessControl (2013-11-20).

[Wasley & Brennan, 2012] Wasley, David, & Joni Brennan 2012. Identity Assurance Framework. https://kantarainitiative.org/confluence/download/attachments/38371386/Kantara%20Initiative_IAWG_US%20FPC%20Report_v2.0.pdf (2013-11-18).

[Węckowski & Małyszko, 2013] Węckowski, Dawid Grzegorz, & Jacek Małyszko 2013. On Information Exchange for Virtual Identities: Survey and Proposal. In ICDS 2013, The Seventh International Conference on Digital Society, pages 59–64.

[Weik & Wahle, 2008] Weik, Peter, & Sebastian Wahle 2008. Towards a generic identity enabler for telco networks. In Proc. 12th Int. Conf. on Intelligence in Networks, Bordeaux, pages 20–23.

[Widdowson & Cantor, 2008] Widdowson, Rod, & Scott Cantor 2008. Identity Provider Discovery Service Protocol and Profile. http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-idp-discovery.pdf (2013-11-20).

[Wieloch, 2011] Wieloch, K. 2011. Budowanie profili przedsiębiorstw z wykorzystaniem utożsamiania odwołań semantycznych do przedsiębiorstw w polskich tekstach ekonomicznych. PhD Thesis, Poznan University of Economics.

[Wild et al., 2014] Wild, Stefan, Falko Braune, Dominik Pretzsch, Michel Rienäcker, & Martin Gaedke 2014. Tamper-Evident User Profiles for WebID-Based Social Networks. In International Conference on Web Engineering, pages 470–479. Springer.

[Yeluri & Castro-Leon, 2014] Yeluri, Raghu, & Enrique Castro-Leon 2014. Identity Management and Control for Clouds, pages 141–159. Apress, Berkeley, CA.

[Zigoris & Zhang, 2006] Zigoris, Philip, & Yi Zhang 2006. Bayesian adaptive user profiling with explicit & implicit feedback. In Proc 15th ACM Int. Conf. on Information and knowledge management, pages 397–404, New York. ACM.

[Zwattendorfer et al., 2013] Zwattendorfer, Bernd, Klaus Stranacher, & Arne Tauber 2013. Towards a federated identity as a service model. In International Conference on Electronic Government and the Information Systems Perspective, pages 43–57. Springer.

2.4 Open Data for Profiling: DBpedia in Building User Identities

2.4.1 Open Data: Introduction

Nowadays ”we live in the age of analytics” [67]. Decisions taken are increasingly built on data supported by the intuition. This trend concerns both business and everyday life, however in business ”decisions are [...] based on data of such complexity that a human mind struggles to comprehend” [67].

There are numerous attempts to estimate the amount of data available on the Web. The user generated content is growing tremendously e.g. every minute in June 2018 over 18 million forecast requests were received by The Weather Channel, 3,7 million Goggle searches were conducted, 3,1 mln GB of traffic was generated by Americans and 473 thousands tweets were sent [150]. These statistics refer only to what can be noticed on the Web, not mentioning the data generated by internal company systems, machines and other sources that are currently available on the Web (including the Deep Web). According to the IBM Marketing Cloud study, 90% of the data on the Internet has been created since 2016, and when we apply this exponential trend to upcoming months, we can confirm that we are currently in the Web of Data era [78] (Figure 2.1).

However, the increasing volume is not the only challenge. The research from IDC indicates that about 90% of the data produced is unstructured, meaning that it not only does not follow a data model, but also is textual data [IDC2014]. The authors of this data are mainly people, what impacts the data quality and adds additional challenges for processing of this data.

Why do we need data? We require it to support decisions lowering the risk of incorrect acting, however the diversity of different scenarios is overwhelming. Data may be used to analyse past e.g. to report the activities, to identify trends, to forecast the future taking into account many aspects, to develop models explaining the diversity or identifying anomalies, to develop data-intensive products (e.g. recommending the best solution) and many others. Some data may impact governmental decisions and some other have a direct influence on people’s quality of life. When analysing data, we should not forget the diversity of data and many classifications that try to address this issue. The data may be divided into:

- Quantitative (what you can measure with discrete or continuous values) and qualitative (meaning that data is the outcome of a subjective observation). Quantitative data in statistics is further divided into categorical (nominal, ordinal) and numerical (interval, ratio).

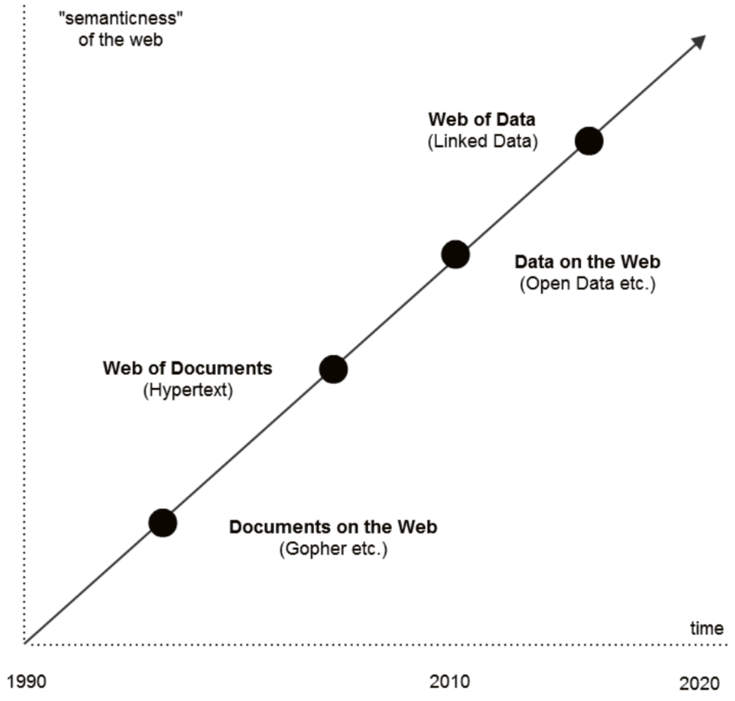


Figure 2.1: Evolution of the Web. Source: [14]

- Structured, unstructured, semi-structured data taking into account the level of structuring and the availability of the underlying data model. The structured data residing e.g. in databases is the easiest to process, while unstructured e.g. customer reviews is the most difficult.

- Emerging from a specific source e.g. machine (logs, statuses) or generated by users (social media, data provided while using a tool, etc.).

- Programming types: primitive (e.g. integer or char) or non-primitive (e.g. array or class).

- Topic of the data e.g. meteorological, historical, cultural, etc.

This classification is by no means complete as the data should always be analysed from a perspective of a domain and a problem that it addresses. Sometimes data is produced to achieve a well-defined goal, while in other cases it is addressed to as a ”useful waste” that emerges from processes carried out by people or machines. The utility of this waste should not be underestimated. However, the data may not be of value for their owner, but there might be third parties that see its potential usage. Therefore, to the list above one more bullet point should be added, namely "closed and open data", distinguishing the data that is made available and may be used for various purposes and the data not easily accessible by third entities.

Definition of Open Data

By definition "open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose" [83]. Another definition extending the previous one says that "open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike" [118]. More business-oriented explanation of the notion may be found in [21], that defines the open data as "data that is freely available to anyone in terms of its use (the chance to apply analytics to it) and rights to republish without restrictions from copyright, patents or other mechanisms of control". All these definitions refer to the issue of openness in terms of freedom, understood similarly as by the Free Software Foundation. The freedom in case of data means free availability, free usage and free redistribution, however not necessarily being for free in monetary terms. What data in particular is subject to this definition? According to [117] open data refers to electronically stored information or recordings e.g. documents, databases, transcripts of hearings, and audio/visual recordings of events. At the same time in case of non-electronic information resources, it is recommended to make these resources available electronically to the extent feasible.

Summarising, openness in case of data should be defined based on data characteristics, namely [118]:

- availability of data and data access: the data must be available as a whole and at a reasonable reproduction cost, in a convenient and modifiable form,
- re-use and redistribution of data needs to be possible, also in case of integrating the data with some other datasets,
- universal participation: there should be no discrimination against fields of endeavour or against persons or groups, meaning that everyone can use the data for any purpose.

Although [83] says that anyone can release his/her data under an open licence for free use, we usually think mostly about government and public sector bodies releasing public information. Therefore, the open data is very often understood as government open data, because government collects enormous amounts of data what is funded by public money and therefore should be shared (and is) with a general audience. The governmental aspect adds some more features to the definition of the open data.

The most comprehensive list of open data features was collected during the meeting of 30 Open Government specialists in Sebastopol, California, USA in 2007. During this meeting, the following set of OGD principles was assembled. These principles were further confirmed in a number of different legal acts, guidelines, best practices, etc. They define open data as data that is [117]:

1. Complete: all public data is made available and this data is not subject to any valid privacy, security or privilege limitations.

2. Primary: data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms (the data is available in its full resolution to enable for building custom solutions and to preserve the data for future users).

3. Timely: data is made available as quick as it is necessary to preserve the value of the data. There should be no delay hindering the data quality.

4. Accessible: the data is available to the widest range of users for the widest range of purposes. This means that the data should be available on the Internet and methods of data preparation and publication should not impact users of a variety of software and hardware platforms. Moreover, the data must be published applying current industry standard protocols and formats, and in case of multiple standards, also using alternative protocols and formats in order to enable for a wide reuse of the data. This also means that automation of the data access must be possible.

5. Machine processable: data should be structured to allow automated processing. This means that data should be properly encoded rather than presented as free-text, images of text, etc. The documentation on the data format and encoding should be also available to potential users.

6. Non-discriminatory: data is available to anyone, with no requirement of registration. The data access should be also possible through anonymous proxies.

7. Non-proprietary: data is available in a format over which no entity has exclusive control. In case of proprietary formats being ubiquitous, multiple formats, including also non-proprietary should be utilised.

8. License-free: data should not be subject to any copyright, patent, trademark or trade secret regulation, however reasonable privacy, security and privilege restrictions may be allowed.

More principles were added to this list by the Sunlight Foundation [152]:

- permanence that refers to the capability of finding data over time without the risk of data being deleted from the public space,
- usage costs: the data should be available at no cost or very marginal costs, as even when the cost is very low it hinders the potential usage and limits the number of users accessing the information.

The features explained above show the character of the open data from the perspective of the open government, however they may also be applied for different data owners as it should be underlined that any organisation can provide open data, so the notion of open data relates to data made public by corporations, universities, NGOs, startups, charities, community groups and individuals, etc. [83].

Why do we need to open data? There are numerous reasons, but before discussing them, it should be noted that from the perspective of economics, data is a resource that may be used by various entities and it still exists (it is inexhaustible). As such the same data may suit various users for diverse purposes and contribute to inter alia [119]:

- innovation how a certain business process is carried out,
- new or improved products or services,
- new knowledge from combined data sources and patterns in large data volumes,
- improved effectiveness and efficiency of services,
- transparency in case of government.

The value of open data is also confirmed by governments and European Commission, that made building a European data economy a part of the Digital Single Market strategy. The initiative aims at enabling the best possible use of the digital data to benefit the economy and society, to unlock the re-use potential and free flow of data across borders to achieve a European digital single market [38].

Classification of Open Data

Data made available may be of different kinds and have a number of potential application scenarios. These scenarios may relate to the following fields [83]:

- Culture: data about cultural works and artefacts collected and held by galleries, libraries, archives and museums; that may contribute e.g. to knowledge on history and culture, development of applications for disabled people enabling them to distantly visit a museum or gallery,

- Science: data produced as a part of scientific research from astronomy to zoology, enabling to extend the existing body of knowledge and develop new products and services.

- Finance: data concerning government accounts (expenditure and revenue) and information on financial markets (stocks, shares, bonds etc) that enables for better management of companies and benchmarking business and economic indicators.

- Statistics: data produced by statistical offices such as the census and key socioeconomic indicators, enabling for building forecasting models for business or profiling customers from a certain region.

- Weather: data used to understand and predict the weather and climate, useful also e.g. in case of forecasting production of electric energy.

- Environment: information related to the natural environment such as presence and level of pollutants e.g. in rivers and seas, of value in case of mining of shall fuel.

This list is not complete, but shows the value of the open data.

On the other hand, not all open data is the same and of the same potential for the economy. Tim Berners Lee [17] proposed a five star model to describe the maturity of open data. The model includes the following stages:

- One star: data is available on the Web, in any format, but with an open licence.

- Two star: data is available on the Web as machine-readable structured data (e.g. excel instead of image scan of a table).

- Three star: in addition to two star, data is presented in a non-proprietary format (e.g. CSV instead of excel).

- Four star: refers to all the above plus usage of open standards from W3C, such as RDF and SPARQL to identify things, so that people can point at your stuff and link it with their data.
- Five star: the highest level of open data maturity refers to the situation when the data is linked to other people's data to provide context.

This model indicates that the emphasis should be put not only on making the data available, but on standardisation of formats and protocols that enable scaling of the approach. The final phase refers to Linked Data and is discussed in the following sections.

Linked Data and Linked Open Data

The simplest definition of the Linked Data says that it is about "using the Web to create typed links between data from different sources" [62], namely referring to the fifth star from the model proposed by Tim Berners Lee, described above. When looking into details, the term Linked Data refers to "data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets" [18].

Both definitions refer to linking of datasets over the Web. It is underlined that the current Web is changing from hypertext links (link documents) to hyperdata links (linking data). In this scenario, data are small components of resources and links enable to drill to the details of those resources. The Linked Data browsers enable to navigate between different data sources by following RDF links and drilling down, benefiting from the granularity of the information. This translates into better search possibilities on the Web, potential of structuring the data as in Wikipedia info boxes, but in automatic manner, meshing up different data through RDF links and because of standardisation easier development of application using this data.

Linked Data is often being referred to as Linked Open Data (LOD), but these are not synonyms. Linked Open Data is Linked Data which is released under an open licence, which does not impede its reuse for free. An example of such licence is Creative Commons CC-BY. Linked Data itself does not have to be open and there is a lot of important use cases when Linked Data is applied internally, and for personal and group-wide data. According to [17], a data set may receive 5-stars in the Tim Berners Lee model, without being open.

Soren Auer discussed the potential of using the Linked Data standard to Open Government

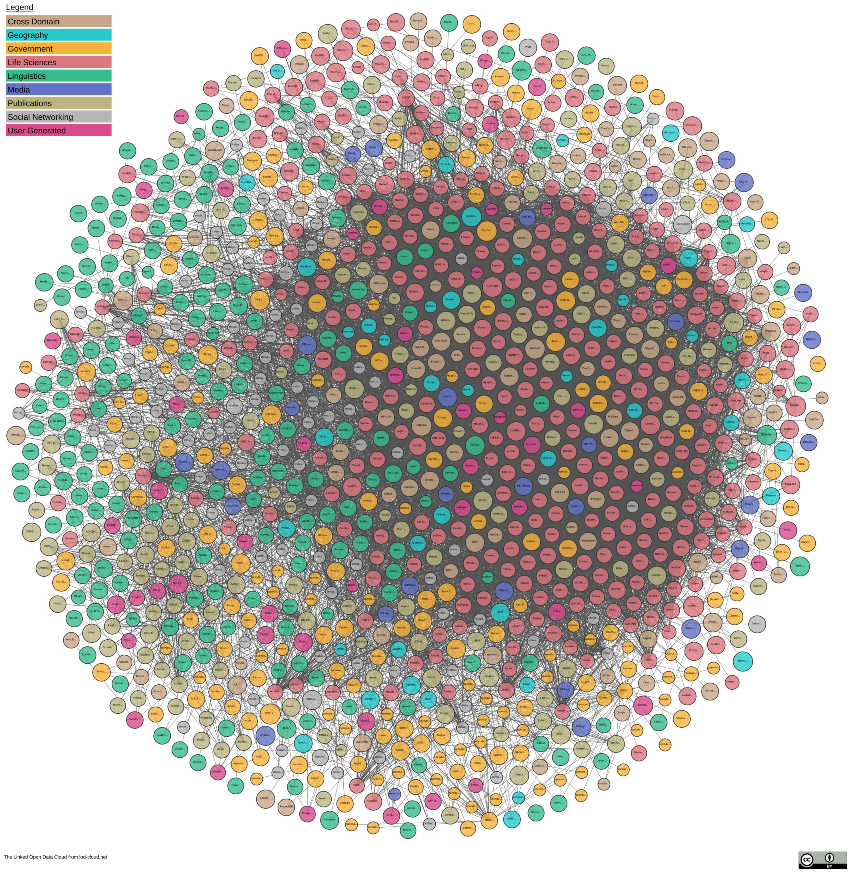


Figure 2.2: The Linked Open Data Cloud. Source: [112]

Data [10]. Currently, the datasets published by local governments are characterised by syntactic heterogeneity (different trees of tags e.g. in XML), semantic heterogeneity (different tags and attributes) and diversity of formats (e.g. XML, CSV, Excel, JSON). Such an approach leads to increased data literacy, but does not contribute to achieving scalability and support wide usage of open data. To achieve the vision of the Web of Data, we need to standardise formats that preserve semantics e.g. RDF, reuse vocabularies and provide tools enabling visualisation of this data.

The current overview of the world of Linked Open Data, named the Linked Open Data Cloud is presented in Figure 2.2. The number of different circles shows the number of different Linked Open Data sources and the relations show how this data is interrelated. The more links, the better potential for the reasoning and the more complex business scenarios. Linked Data when used for description of entities add additional challenges to profiling methods, but also grow a scope of potential usages.

Application examples

A successful example of Linked Data usage concerns British Broadcasting Corporation (BBC) that uses Linked Data as a data integration technology. BBC runs a number of television and radio channels, and every year produces a huge amount of content. BBC stations use separate content management systems that store not only content in terms of tunes or films, but also the textual description of this content. BBC started using Linked Data technologies as a controlled vocabulary to link the content about the same topic from different repositories and also to augment content with additional data from the Linked Open Data cloud [93]. Such an approach is a typical scenario, that shows potential of linking content from distributed data sources and further enable for improved search or recommendation.

Open data may however provide also a different type of value. In Nepal, the Aid Management Platform was developed by the Ministry of Finance to enable for monitoring aid received and budget spending. All organisations benefiting from the budget are required to report details about their funding and programmes carried out, building society-driven open data initiatives including Open Nepal's open data portal, Aid Snapshot, and the Open Aid Partnership [156]. Such open data enables producing analyses for the policy reforms; planning, monitoring and evaluation of various aid programmes or formulating the government's budget, helping to trace gaps between spending and output.

A similar, but more advanced, example may be found in Argentine, where open data was used to improve health services and enable cooperation between society and government to build applications and tools improving access of citizens to health services [138]. This proves that open data can also improve the quality services, which are available to the majority of the citizens.

Linked Data and Linked Open Data for Profiling

In case of profiling, Linked Open Data or Linked Data in general is treated as a technology that may enrich functionalities of tools and expand the number of application scenarios. Profiling with the usage of Linked Data should not be confused with Linked Data Profiling that aims at description of a topic represented by a data source or managing the data itself (clustering, matching, disambiguating, studying patterns within the data, etc.).

Linked Data may however influence how the data model of a profile is expressed or allow expanding features included in the profile by the means of the automatic reasoning. An example of such an approach is presented in the following paper.

Challenges of Using Open Data

The potential emerging from Linked Data for profiling should not be underestimated. However, we should remember about the data related challenges that are still valid. These include the following aspects:

- irrelevance: not all data that may be used should be used, before using a dataset one should remember that even if data set is for free, there is a cost of linking, reasoning, etc. and the value in terms of the final solution should be studied,
- quality: data may be wrong, may contain errors, so before exploiting it, the quality of the data needs to be verified, especially if it is to impact business decisions,
- interpretation: studying documentation of data is important and any assumption taken before learning the data, may be wrong, which points to the importance of the validation of achieved results.

2.4.2 Towards Using DBpedia for Building User Identities

The goal of the following paper is to propose how a profile or a virtual user identity may be described using open data. The focus is in particular on a data structure that enables reasoning over the profile/identity (Linked Open Data). This goal is in line with the following secondary goal of the thesis: ”Analysing how to instantiate a profile of a person or a thing using semantic approaches enabling for diverse application scenarios”.

The paper was published in the proceedings of the 1st NLP&DBpedia Workshop, held during the 12th International Conference on Semantic Web (ISWC2013), 21-25.10.2013, Sydney, Australia. Detailed bibliographic reference is as follows: Filipowska, A., Małyszko, J., 2013, Towards using Wikipedia for Building User Identities, Proceedings of the NLP & DBpedia workshop, pages 1-8.

Towards using DBpedia for building user identities

Agata Filipowska and Jacek Małyszko

Poznan University of Economics
Faculty of Informatics and Electronic Economy
Department of Information Systems
Al. Niepodleglosci 10, 61-875 Poznan, Poland
{firstname.lastname}@kie.ue.poznan.pl,
http://www.kie.ue.poznan.pl

Abstract. Internet offers a number of various services that to maximise the user experience apply different personalisation techniques. An important resource of every personalisation method is a user profile. The more information on the user is available in such profile, the better. Therefore, together with maturing of these mechanisms, the notion of identity emerged. The identity exceeds the user profile with information that is more detailed or enables benefiting from additional functionalities. The information stored within an identity needs to be understandable for different services to be easily reused. This can be achieved using the DBpedia.
The goal of the article is to describe the design of a method that potentially enables providing data to build the user identity, based on his behaviour on the Web. The method is elaborated as well as an example of application is presented.

Keywords: DBpedia, Wikipedia, information extraction, identity

1 Introduction

Most users leave a significant amount of information about themselves on the Web. They abandon their anonymity freely (sometimes unconsciously), in order to stay connected with their friends on the social networking sites, communicate with their government or build their reputation [1], [9]. Also, the service providers want to learn detailed characteristics of their users by using different profiling practices [5], in order to provide a better service and preserve their customers. As a result of these trends, a problem emerged of how the users should establish and manage their presence on the Web, namely their digital identities. This issue is being researched for many years now [5].

One of the major challenges concerning the identity management systems is creation and maintenance of many perspectives on users identity, called virtual identities, most preferably without explicit actions of the user. Virtual identity is understood as a collection of topics concerning specific interest of a user. In this paper we present a method that enables automatic identification of such topics using Wikipedia and information extraction techniques. The method is developed for the Polish language. It utilizes Wikipedia concepts but can easily be extended to DBpedia resources. As there is no Polish DBpedia yet, this will not be covered by this article. However, the work on Polish DBpedia is ongoing and this will be addressed in the future.

The remainder of the paper is structured as follows. Section 2 is devoted to a short summary of virtual identity definitions. In the next section, we indicate current projects and existing approaches that raise the issue of users virtual identities and provide solutions in this area. Section 4 describes the method proposed to identify concepts building the users identity. Finally, in Section 5 we focus on a Use Case demonstrating the application of the method. The article concludes with the final remarks.

2 Definition of Identity

The concept of an identity has been adopted by Information Science as a formal representation of knowledge about a certain person, or any other (digital or real-world) subject. Concerning an identity of a person, it is understood as a set of attributes (permanent or temporary) characterizing a person [13], that is required by providers of services that the person uses [8]. Obviously, a virtual identity cannot capture all characteristics of a person; it is therefore only a partial representation of a subject [13], [14]. Traditionally, an identity is considered as a permanent entity, persisted in a kind of a datastore in order to be accessible many times for a long period of time. However, it can be also understood as something created on-the-fly and used (attached to a person) only during a single session, while a user performs certain tasks or when a particular transaction is performed [7], [13].

More generally, a virtual identity can be defined as a digital representation of a set of claims made by one party about itself or another digital subject [3]. A natural person (a human being) is one example of such entity; other example is a whole organization (i.e. juridical person) [14]. An identity can either be used in a single environment (for example, in a single system or company), or used in many different environments, for example across organizational boundaries. At the same time, different information about every entity is exchanged in different contexts; for example, different user characteristics are needed in e-banking portals and in movie recommender systems. We can therefore either say, that a virtual identity is just one set of claims about a digital subject and for any given digital subject there will typically exist many virtual identities [7], or that each subject has only one identity, but such identity has multiple facets, that are used depending on the context [13].

The identity of a digital subject can be established by combining both the real-world attributes (for example name, address, social security number, physical traits, etc.) and the digital ones (such as passwords, access rights, biometrics, type of encoding, network address and so on) [6]. The information stored in an identity can be used either for the authentication purposes (its goal is to ensure, that a certain person is indeed what he or she claims to be), or as the attribute information (representing the details about the person) [14]. A set of processes relating to the disclosure of the information about the person and usage of this information is called identification [13].

For the requirements of the "Ego - Virtual identity" (Ego) project[1], presented in the paper, the identity is understood as an information structure describing the information needs of a user. This structure is grounded in the Wikipedia concepts' graph to ease its maintenance and assure usefulness while personalizing information content, especially from the information needs evolution point of view. The future work concerns extending the method towards DBpedia resources.

3 Related work

In the following sections, we present the state of the art analysis of the identity management systems on the Internet in terms of the business goals, that they pursue and the functionalities, that they provide. We identify the most important projects and solutions in the area of identity management systems, that may benefit from the approach we suggest. The main projects that we concentrated on are following: FIDIS[2], SWIFT[3], PICOS[4], PRIME[5], STORK[6], ProjectVRM[7]. Moreover, there exist also frequently updated lists of identity-related efforts[8].

In addition to the above-mentioned projects, a number of already implemented solutions were analyzed. These solutions however mainly focus on the authorisation aspect, leaving behind the notion of user representation e.g. the OpenID protocol describes a user with a limited set of attributes only [11]. Similar, authorisation focused, approaches are e.g. [12], [2], [4]. An interesting, and comparable to ours effort is WebID [15] that uses FOAF vocabulary to describe a user.

Some of the solutions are widely used in business, for example the OAuth protocol [9] or various OpenID implementations, while some of them are at earlier stages of development and adoption, e.g. WebID and Higgins.

Finally, its also very important to indicate, that several organizations have emerged and aim at consolidating and coordinating efforts in the area, of which

[1] http://kie.ue.poznan.pl/en/project/ego-virtual-identity
[2] http://www.fidis.net/
[3] http://www.ist-swift.org/
[4] http://www.picos-project.eu/
[5] https://www.prime-project.eu/
[6] https://www.eid-stork.eu/
[7] http://projectvrm.org/
[8] For example: http://personaldataecosystem.org/2011/06/startup/, http://blogs.law.harvard.edu/vrm/development/, accessed on 15/10/2013
[9] It is used for example by Facebook, Google and Last.FM

the most important are probably Kantara Initiative[10], Identity Commons[11] and formerly Liberty Alliance[12].

It can be easily noticed, that the concept of virtual identities is heavily studied. Nevertheless, as it is a wide field to investigate, different areas of virtual identity creation, maintenance and usage can be explored by different projects. To the best of our knowledge, the approach focusing on automatic creation of users virtual identity that links experience from the fields of information extraction and Wikipedia does not exist.

4 Approach and methods used

This section presents details of the approach we apply to create the identity of a user. The phases of creating the users virtual identity are as follows:

Phase 1: Tracing user behavior. The first step towards building a user's identity concerns identification of topics of user's interest. Of course, these topics can be entered manually by a user (a so-called explicit user modeling [17]), but the identity management systems usually provide additional functionalities to make the whole process more effective.

There is a lot of information about a user even before she or he starts using a given identity management system. Such information is often spread across multiple domains such as web portals, social networking sites, etc. Therefore, the identity management systems can try to somehow import and aggregate information about the user from such sources automatically. To make that feasible, the user's data export mechanisms must be made available by owners of such systems. An example of such initiatives are Data Liberation Front[13] and Data Portability Project[14].

We build the identity of a user based on a wide range of his activities on the Web. Our goal is to engage the service providers in this process, as discussed in [16]. At the current stage of the experiment, we focus on building user's identity based on analysis of the Web pages the user visited. To that end, we have implemented a Web browser plug-in, which a user has to install and have it enabled while browsing. The plug-in extracts (structural, XSLT extraction) the main content of the website and commits it on the server.

Phase 2: Analysis of the visited Web sites. The content that is uploaded to the server is analyzed using the lexical extraction module to identify the differentiating phrases and assign a topic. For the list of topics that are the most representative for the whole content of the network, we chose the Wikipedia categories and concepts list.

The extracted content of the website is analyzed using NLP to identify named entities, cross references, etc. and as a result provide a set of words (surfaces

[10] http://kantarainitiative.org/
[11] http://www.identitycommons.net/
[12] http://projectliberty.org/
[13] http://www.dataliberation.org
[14] http://dataportability.org/

existing in the text, further being referred to as phrases), that will be subject to further processing. What is important, that the approach works for the Polish language and is contextual.

Phase 3: Building a representation of a website for the needs of the identity building. The most crucial step, from the point of view of this paper, as well as for the user acceptance of the system being developed, is indication of a topic, the website mentions. This is done in the following steps.

Firstly (in the preparatory phase), all Wikipedia pages are processed in order to identify concepts (Wikilinks) that appear on these pages in order to learn a phrases-concepts mapping, similarly as it was done by [10]. This process is repeated periodically. Thus, we have obtained 5.150.143 phrase – concept mappings. This mapping is ambiguous, as many phrases may point to many different Wikipedia concepts (on average, each phrase points to 1.21 concepts, but there are some phrases that are mapped to up to 4000 concepts). Still, based on that for each phrase we are able to retrieve a list of candidate concepts.

The method of indication of a topic of a website assigns to each phrase from the text (f) concepts from Wikipedia ($c1$ – $c6$ in Figure 1) obtained as indicated in the previous paragraph. Then, for these concepts ($c1$ – $c6$), the upper level categories of concepts are indicated ($c11$ – $c51$). The Wikipedia category structure enables to build a whole tree over the initial concepts that were assigned, e.g. for concept *Peter Higgs*, based on the Polish Wikipedia structure, we retrieve categories such as *Scottish Physicists*, *Born in 1929*, etc. Currently, we use only three levels within the tree (experimentally evaluated). Then, using the bottom-up propagation method the first-level concepts (mapped from phrases extracted from the website content) vote for the upper level concepts. The bottom up propagation measure combines five frequencies:

- The number of times a phrase from the article text refers to a concept from the Wikipedia.
- The number of times a phrase (surface form) appears in the Wikipedia.
- The number of times a given concept is referenced within the Wikipedia.
- The frequency of a word in the language (in our case the Polish language).
- The number of sub-concepts of a concept.

As a result of the bottom-up propagation, we identify a concept (not necessarily the top-level one), that is the most probable topic of the website. Afterwards, the phrase from the website being the most strongly connected with the concept assigned as a topic, is removed from the initial list of phrases and the procedure is repeated for the remaining phrases. While experimenting, we identified that for most of the articles three iterations are enough to provide the most meaningful concepts describing the website's topic.

These concepts are then mapped on the user's virtual identity. Each new package of topics, changes the initial identity. The weights assigned to different topics within the identity, reflect also maturing in time. Also, user may support this process by manually extending the list of automatically assigned categories.

The user identity created by the system, may be then further used for the needs of personalization of websites visited by the user. The Ego system is to pro-

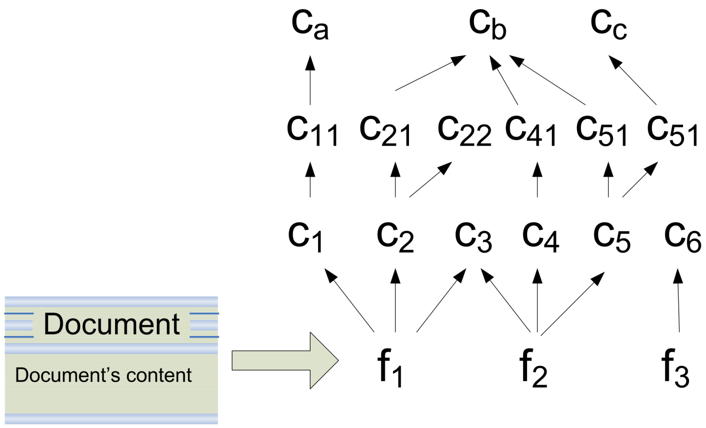


Fig. 1. The multi-layer representation of the article: phrases extracted (f) and Wikipedia concept hierarchy (c).

vide a number of functionalities enabling for sharing and encrypting the identity, authorizing a service provider as well as enabling user to manage the identity and to access it [16].

5 Use Case-based Validation

The presented approach is about to be validated with the real users, who committed to use Ego for a certain period of time and share their experiences. Up till now, the Use Case-based validation has been performed. For the sake of clarity, we present details based on one news article only. The article concerns the Noble Prize Winner Peter Higgs (it is in Polish and is available at Polskie Radio website[15].

The content of the article was extracted and loaded in the database as a logical document (this concerns the topic and the content of the article; menus, comments etc. are not further analysed). Then, the lexical extraction rules extracted 44 different phrases from the article e.g. uroczystości (celebration), professor, etc., out of which 31 were mapped on Wikipedia phrases.

For these Wikipedia phrases, 2052 Wikipedia concepts were retrieved (identified by different URLs) including three upper levels (2052 is a total number of concepts in the tree initially representing the topic of the article). The most frequent concepts in the first level mapping were e.g. fizyka (physics), konferencje miedzynarodowe (international conferences), mechanika kwantowa (quantum physics), II wojna światowa (second world war).

[15] `http://www.polskieradio.pl/23/266/Artykul/951564,`
`Profesor-Higgs-zapadl-sie-pod-ziemie-`

Then, the relations between different concept categories were exploited using the bottom up propagation method. After applying the method, the following concepts were identified as the most descriptive for the article (in the order of importance):

- Urodzeni w XX wieku (born in XX century),
- Popularność (Popularity),
- Higgs,
- Szkolnictwo wyższe (Higher education),
- Nauki przyrodnicze (Natural sciences).

These concepts may be then further mapped on the Wikipedia category structure graph representing the users identity, but this issue is beyond the scope of this paper.

6 Conclusions and future work

The goal of this paper was to present a method that enables for identification of topics that are of user's interest using Wikipedia and information extraction techniques, and based on the behavior of a user on the Web. Starting from a general summary of the Virtual Identity definitions, we presented a method that may be used in order to create user identities using Wikipedia. We also demonstrated an application scenario.

The future work will especially be devoted to tuning of mechanisms developed as well as carrying out an extensive validation of the approach with the real users. The major issue that needs additional research is the bottom-up propagation method that should eliminate concepts being pointed from the multiple websites such as e.g. born in XX century.

Further research will also concern changing the Wikipedia to the DBpedia to allow for an extensive reasoning. This could also offer additional functionalities to an identity management system and service providers that will benefit from it. However, the work on the Polish DBpedia is still the ongoing effort.

Acknowledgments. The work published in this article was supported by the Polish Ministry of Science and Higher Education (decision no. 0987/R/H03/2010/10), upon contract with the Polish National Centre of Research and Development (NCBiR) (contract no. NR11-0037-10/2011) on the project titled: "Ego – Virtual Identity" (http://kie.ue.poznan.pl/en/project/ego-virtual-identity).

References

1. M. Bernstein, A. Monroy-Hernandez, D. Harry, P. Andre, K. Panovich, and G. Vargas. An analysis of anonymity and ephemerality in a large online community. In *5th International Conference on Weblogs and Social Media (ICWSM)*, Menlo Park, 2011. The AAAI Press.
2. C. Burton. The Information Card Ecosystem: The Foundamental Leap from Cookies & Passwords to Cards & Selectors. http://wiki.informationcard.net/files/icf-information-card-ecosystem-white-paper.pdf, 2011.

3. K. Cameron. The laws of identity. `http://msdn.microsoft.com/en-us/library/ms996456.aspx`, 2005.
4. E. Hammer-Lahav. The OAuth 1.0 Protocol. *Internet Engineering Task Force (IETF)*, 2011.
5. M. Hildebrandt and J. Backhouse. D7.2: Descriptive analysis and inventory of profiling practices. FIDIS (Future of Identity in Information Society) Project Deliverable. `http://www.cosic.esat.kuleuven.be/publications/article-827.pdf`, 2005.
6. J. Hodges, R. Philpott, and E. Maler. Glossary for the OASIS Security Assertion Markup Language (SAML) V2.0. `http://docs.oasis-open.org/security/saml/v2.0/saml-glossary-2.0-os.pdf`, 2005.
7. Identity Commons. Identity landscape. `http://wiki.idcommons.net/Identity_Landscape`, 2012.
8. Kantara Initiative Wiki. Consumer Identity Workgroup – scenarios, use cases, & definitions v0.3 . `http://kantarainitiative.org/confluence/pages/viewpage.action?pageId=38371527`, 2012.
9. R. Leenes, J. Schallaböck, and M. Hansen. Prime white paper. `http://security.future-internet.eu/images/2/27/Prime_White.pdf`, 2008.
10. D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.
11. OpenID Community. OpenID Authentication 2.0 - Final. `http://openid.net/specs/openid-authentication-2_0.html`, 2007.
12. N. Ragouzis, J. Hughes, R. Philpott, and E. Maler. Security Assertion Markup Language (SAML) V2.0 Technical Overview. `http://www.oasis-open.org/committees/documents.php?wg_abbrev=security`, 2006.
13. K. Rannenberg, D. Royer, and A. Deuker. *The Future of Identity in the Information Society: Challenges and Opportunities.* Springer, 1st edition, 2009.
14. M. Rundle, E. Maler, A. Nadalin, D. Reed, and D. Thibeau. The Open Identity Trust Framework (OITF) Model (White paper). `http://openidentityexchange.org/sites/default/files/the-open-identity-trust-framework-model-2010-03.pdf`, 2010.
15. H. Story and S. Corlosquet. Web 1.0. Web Identification and Discovery. `http://www.w3.org/2005/Incubator/webid/spec/`, 2011.
16. D. G. Weckowski and J. Malyszko. On information exchange for virtual identities: Survey and proposal. *IARIA, 2013*, 978-1-61208-249-3:59–64, 2013.
17. P. Zigoris and Y. Zhang. Bayesian adaptive user profiling with explicit & implicit feedback. In *Proc 15th ACM international conference on Information and knowledge management*, pages 397–404, New York, 2006. ACM.

2.5 Conclusions

The chapter aimed at providing background and foundation for the following chapters of the thesis. Its goal was "to define a profile of a person or a thing and identify features of a person or a thing that may be represented in a profile and may be usable in diverse application scenarios". This goal was further translated into two secondary goals addressed by specific sections of this chapter.

The first secondary goal (G1.1) was to analyse different approaches for describing a profile or an identity of a user or a thing that are applied in different classes of systems, for example identity management systems. Therefore, the goal of Section 2.2 was to describe a system that enables to bridge the gap between identity management and user modelling systems. The proposed solution supports users to automatically create their identities and manage these identities for the needs of different services. The identity should enable not only authorisation, but also personalisation of content displayed to a user.

The paper presented contributes to achieving this goal by: analysis of the related work in the area of managing user data e.g. identities, profiles, etc.; proposing identity lifecycle (starting from creation, supporting its updates, merging different profiles, update by a user, querying, controlling access by different services, etc.) and proposing an architecture of the system that allows providing a user with an invisible personalisation support while browsing the content (the system is to be working in the background). The system implementing these requirements was delivered within the EGO - Virtual Identity project.

Section 2.3 also addressed the secondary goal previously mentioned (G1.1). It was to study functionality and use cases of identity management systems available on the Web. The focus of the work was to describe trends, ideas an shortcomings of existing solutions and identify potential extensions. The research results achieved and described within the paper concern a detailed review of state of the art in the domain of identity management. The notion of a digital identity is analysed from different perspectives and various definitions are provided. The paper also proposes a set of use cases for identity management systems. The use cases focus on improving user experience while utilising data included in the virtual identity. Then, a comparison of selected identity protocols, projects and initiatives taking into account the proposed use cases follows. This issue was also further addressed in the Bachelor thesis of Adam Maćkowiak, supervised by Agata Filipowska. The thesis concerned utilisation of a virtual identity emerging from previous activities of a user on the Web or his profiles on social media, in the process of advising financial instruments.

The G1.2 secondary goal of the thesis was to address how to instantiate a profile of a person or a thing using semantic approaches enabling for diverse application scenarios. This goal was further translated into the goal of the paper to propose how a profile or a virtual user identity may be described using a data structure that enables reasoning over the profile/identity. Here, also the issue of the open data, and in particular Linked Open Data, is discussed. The paper included in Section 2.4 provides a method for building a profile (identity) of a user based on his activities on the Web. The profile is described using categories from Wikipedia. A method that was presented in the paper is developed for the Polish language. The goal was to derive from keywords included in the paper, topics based on Wikipedia categories that may be included in the profile/identity. The paper presented also a potential extension to DBpedia.

The goals defined for this chapter were achieved. The next chapter of the thesis is to study how a profile can be built. The terminology defined here and also some concepts introduced are utilised.

Chapter 3

Profiling Techniques

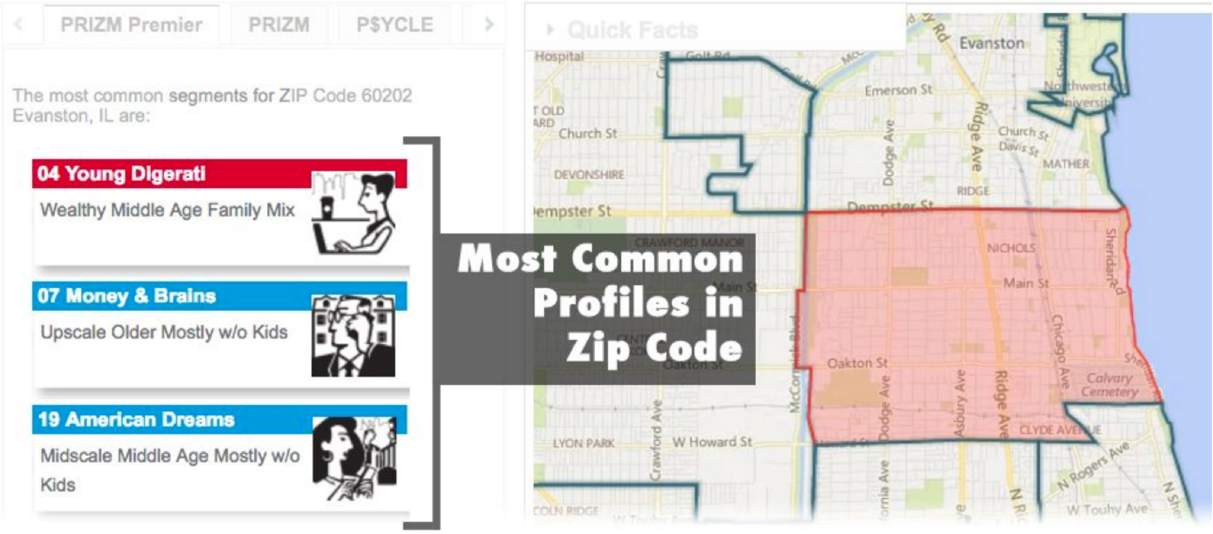
3.1 Introduction

3.1.1 Motivation

Profiling techniques and a scope of a profile greatly depend on the application scenario. These scenarios may be diverse, making researchers study different attributes and behaviours of users. An interesting profiling example that confirms this statement, applied in business practice, is eLoyalty case study described in [143]. eLoyalty proposed a solution that combines a personality (profile) of a client with a personality of a seller on a hotline. The solution is based on the conversation method described within the Process Communication Model developed by Taibi Kahler. The model identifies 6 different personality types, e.g.:

- "Workaholic": person likes facts, focuses on the task to be done and each side conversation is a waste of time for the person.
- "Reactor": a person holding such personality takes care of feelings and building relationships with other people.
- "Persister": focuses on achieving perfection.
- "Dreamer": is a sensitive and introvert person.

Each of the personality profiles is characterised by a number of features describing behaviour of a person with a chosen personality type to enable for identification of a personality. Analysing language forms used by a client and his/her behaviour during the conversation, it is possible to



PRIZM Premier | PRIZM | P$YCLE

The most common segments for ZIP Code 60202 Evanston, IL are:

04 Young Digerati
Wealthy Middle Age Family Mix

07 Money & Brains
Upscale Older Mostly w/o Kids

19 American Dreams
Midscale Middle Age Mostly w/o Kids

Figure 3.1: Extension of customer data. Source: [54]

discover customer's personality type and adjust behaviour of the agent accordingly. Moreover, in case of future contacts, it is possible to select an agent with a personality type best suited for finishing the call with the customer successfully. Similar tests are performed e.g. while selecting a team for the space mission[1].

Profiling does not need to use only data submitted by a user or expressed by his/her behaviour. It is also possible to extend data submitted by a user using e.g. open data described in Section 2.4.1. Such data includes publicly accessible data such as e.g. census, information on a location, public statistics and enables to extend information collected on a person. For example, knowing a zip code of a customer, we may derive what is the character of the neighbourhood he/she lives in (type of houses, level of income, type of a location). An example of such an extension of the customer data is presented in Figure 3.1.

3.1.2 Goals

The goal of this chapter is to analyse profiling methods that enable for describing a user/a thing or relations between users. This goal aligns well with answering research questions on how to model a profile e.g. based on data provided by a user and how to describe relations between entities. To achieve the objective, two secondary goals were defined, namely:

[1] https://pl.scribd.com/document/17867981/The-History-of-the-Process-Communication-Model-in-Astronaut-Selectio

- Describing a user profile w.r.t. user personality and his/her colour preferences. The supplementary goal is to study relations between personality traits and user colour preferences using different methods of analysis.
- Creating a method for describing relations between users focusing on quantitative and qualitative aspects of a relation on the example of a social network.

3.1.3 Structure of the Chapter

The chapter consists of four sections including introduction presenting relation to goals of the thesis and a summary that presents results that were achieved in relation to these goals. Section 3.2 contributes to achieving the first of the secondary goals mentioned, and section 3.3 focuses on the second of these goals.

3.2 Profiling User's Personality Using Colours: Connecting BFI-44 Personality Traits and Plutchik's Wheel of Emotions

The goal of the section is to analyse personalities of users (expressed as personality traits using the Big Five Inventory) and their colour preferences. This goal is in line with the secondary goal of the thesis, namely: "Describing a user profile w.r.t. user personality and his/her colour preferences and studying relations between personality traits and user colour preferences using different methods of analysis"[2].

3.2.1 Introduction

Knowing preferences of a customer or a user is crucial for adoption of products and services. These preferences may relate to e.g. user interests, locations visited or people followed on social

[2] The section is based on two papers:

- Wieloch, M., Kabzińska, K., Filipiak, D., Filipowska, A., Profiling User Colour Preferences with BFI-44 Personality Traits, 10th Workshop on Applications of Knowledge-Based Technologies in Business (AKTB 2018) organised at the 21st International Conference on Business Information Systems, Berlin, Germany, July 18–20, ISBN 978-3-030-04849-5, pp. 63-76.
- Kabzińska, K., Wieloch, M., Filipiak, D., Filipowska, A., 2019, Profiling User's Personality Using Colours: Connecting BFI-44 Personality Traits and Plutchik's Wheel of Emotions, Advances in Intelligent Systems and Computing, 854, pp. 371-380.

media. On the other hand, they may also concern colours of products recommended or displayed to a user. User preferences may emerge from many different factors (environment, previous experiences, etc.), however they are also greatly influenced by a user's personality. In the thesis, while modelling a user, we mostly refer to the notion of profile, however when discussing psychological traits, we will refer to the definition of a personality.

The aim of our research is to analyse users' personalities and their colour preferences with the Big Five Inventory, also known as BFI-44 [87]. Researchers have already recognised the importance of this subject. Zuckerman-Kuhlman Personality Questionnaire [154] or Eysenks [49] personality tests have been previously used to measure personality traits – participants had to grade pictures with specific plain colours and the correlation with their personality was studied. Ferwerda et al. [40] grappled with the issue of predicting personality from colours of posted pictures. In this case, the Big Five model was applied in order to have a full view of personality traits.

This section is structured as follows. A short review of the existing body of knowledge is presented in the next subsection. Section 3.2.3 presents details of the data preparation process. In our approach, images with dominant colours inspired by Plutchik's Wheel of Emotions have been randomly chosen from the Flickr database using the snowball sampling method. They were attached to the questionnaire with the original BFI-44 test. The following section presents findings of the questionnaire. Sections 3.2.5 and 3.2.6 deliver two different models, modelling relation between colour preferences and personality traits. The connections between traits and colours have been examined and also differences between outcomes achieved using different algorithms analysed. Finally, conclusions are presented in the last section.

3.2.2 Related work

Analysis of Pictures

The two most popular approaches used to analyse images on the open social media (OSM) encompass: clustering [75] and determining sentiment of an image [105]. Following Souza et al. [149], one can enlist the following topics connected to OSM pictures' analysis: engagement connected with them [12, 173], self-presentation in social media [48, 146], prediction of age and gender from the visual content [85], recognising the basic personality types from photos basing on the Big Five Inventory questionnaire [40].

Based on their content images may be divided into 8 groups: activities, captured photos, fashion, friends, food, gadgets, pets and selfies. Friends and selfies are the most popular categories on Instagram, forming almost 50% of the all photos. Publishing a selfie online is a way of attracting attention as such photos collect more likes and comments [48].

Users behave differently on each social platform, depending on the type of that platform, user's age and gender. By combining comments, hashtags and content of photos and using appropriate algorithms, it is possible to identify these details and provide classification of users. Thanks to BFI-44, past studies have yielded some important insights into personality, emotions, and reactions of users. Researchers focused on user profiling, photos and texts such as questionnaires [81], comments [102], hashtags [50], and captions [128].

Big Five Personality Traits

The Big Five Inventory (BFI-44) is a personality test (based on the Big Five Model) and consists of 44 questions that measure the Big Five traits [87] such as Extraversion/Introversion, Agreeableness/Antagonism, Conscientiousness/Lack of Direction, Neuroticism/Emotional Stability, and Openness/Closeness to Experience. Following Ahrndt et al. [6], these five personality features can be defined as follows:

- *Extraversion* is related to interactions with other people and gaining the energy from them, contrary to being more independent (e.g. action-oriented, outgoing and energetic behaviour vs. inward, solitary and reserved behaviour).

- *Agreeableness* stems from being trustful, helpful and optimistic, contrary to being antagonistic and sceptical (e.g. cooperative, friendly and compassionate behaviour vs. detached, analytical and antagonistic behaviour).

- *Conscientiousness* is connected to the level of self-discipline and acting dutifully, contrary to spontaneity (e.g. efficient, organised and planned behaviour vs. careless, easy-going and spontaneous behaviour),

- *Neuroticism* reflects the inability in dealing with stress, contrary to emotional stability and confidence, addressed the level of emotional reaction to events (e.g. nervous, sensitive and pessimistic behaviour vs. emotionally stable, secure and confident behaviour).

- *Openness* relies on creativity (e.g. curious, inventive and emotional behaviour vs. consistent, cautious and conservative behaviour).

Some researchers sought to understand other phenomena, such as loneliness or sadness. Pittman [131] established a link between social media use and offline loneliness. Students who are active online usually feel less lonely in real life. By the examination of norms of expressing emotions in OSM, Lup et al. [104] claim that there is a relation between Instagram usage and depression symptoms. Waterloo et al. [170] have carried an extensive study on sadness, anger, disappointment, worry, joy, and pride. These studies proved that positive expressions are perceived better than negative by users across all OSM platforms. Some scholars claim that there are differences in a way of expressing emotions between men and women. In other papers, colours of photos taken in two cities in a certain period of time were compared [68]. Other researchers tried to tackle the problem of sentiment analysis from the visual content [20]. The standard positive/negative/neutral classification can be extended by combining visual content, comments and colours of the photo.

Colours

A universal approach for dividing colours has not been found yet. Some methods are focused on solely blue, green, and red [101]. On the other hand, Ferwerda et al. attached orange, violet and yellow to the aforementioned set [40]. Some scholars investigate so called low-level features, such as chrominance [125]. Plutchik's Wheel of Emotions [132] includes 24 emotions (8 basic emotions with 4 intensity levels) presented on 8 colours (each of them has 4 hues and the least intense one does not carry a meaning).

Despite the representation of emotions as colours, there are no connections between emotions and colours, since it was not the purpose of wheel's author. Among many approaches for choosing basic colours palettes, this work is inspired by those from the common illustrations of Plutchik's Wheel of Emotions which are pink, green, blue, dark blue, red, dark green, orange, and yellow.

When it comes to recognising leading colours of the pictures, it is necessary to define the dominant colour of every pixel [114, 133]. A colour can be found in the colour palette or a lookup table. Determining which colour is dominant is tightly coupled with the number of pixels having the same colour in an image. The colour with the biggest number of pixels is therefore defined as a dominant colour of an image. The dominant colour can be found in the palette of colours. The most popular colour representation scheme is the RGB (Red-Green-Blue) palette [11], but it is complicated to define hues ranges within one colour using this palette. Therefore, HSV scale [24], which is an abbreviation from Hue-Saturation-Value, can be used for this task. After defining

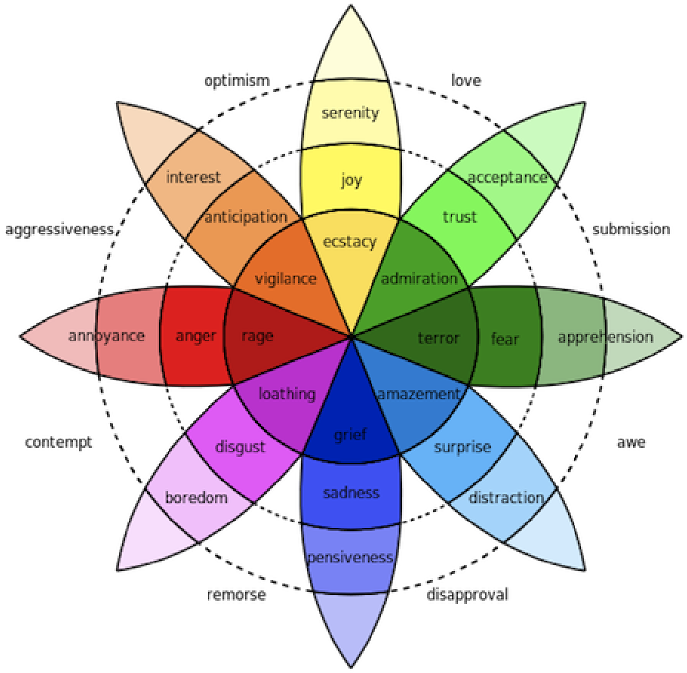


Figure 3.2: Plutchik's Wheel of Emotions. Source: Borth et al. [20]

the range of hue within one colour, the saturation and values have to be stated. On the contrary, defining colour ranges for the RGB colour palette is much more complex. Therefore, using the HSV palette and converting it to the RGB colour palette gained popularity in related research (please see Table 3.1 for a comparison).

Table 3.1: RGB and HSV colour spaces.

Channel	Range	Unit	Description
R	0-255	8 bits	Intensity of red (black to white)
G	0-255	8 bits	Intensity of green (black to white)
B	0-255	8 bits	Intensity of blue (black to white)
H	0-360	degree	Hue
S	0-100	percent	Saturation (bright to dark)
V	0-100	percent	Value (black to white)

Source: Ha et al. [58]

3.2.3 Data collection

Our experiment was based on a questionnaire, which is a combination of the standard BFI-44 test and assessment of Flickr images. The latter are used to examine links between the user's personality and his colour preferences. This section describes the process of preparation of our experiment. It entails Flickr data collection and processing, questionnaire preparation, and the description of the scoring formula used.

Questionnaire

Flickr[3] is one of the most popular OSM platforms. It is image oriented, which makes it particularly useful in the experiment. Flickr has a convenient API, which facilitates preparing a sample. Therefore, we decided to collect pictures from this service by:

1. Following an approach similar to the ones presented by Machajdik & Hanbury [105] and Jamil et al. [84], a random photographer was chosen using the *#photography* hashtag.
2. Using snowball sampling [12, 71, 102], users followed by that photographer were selected (329 users plus the photographer).
3. Following Min & Cheng [114], for each selected Flickr profile, the most recent 100 pictures (or less, if they had fewer photos) have been chosen using Flickr API. This resulted in acquiring 32,056 photos in total.
4. For every photo, a number of pixels of every colour has been encoded using the RGB palette.
5. Values represented in the RGB scale were converted to the HSV model, as an extension of previous ranges in order to enhance their perception by human eyes and therefore simplify the detection of more hues of colours.
6. The downloaded images were divided into 9 categories (8 categories according to chosen basic colours and a category for these colours that are not included in any of the remaining categories).
7. Ranges for specific colours were created using a table of colours[4] and Rapid Tables[5] in order to adjust them to the 8 selected colours.
8. Having assigned the categories of colours, a number of pixels corresponding to one of 9 categories was assigned to each image.

[3] `http://flickr.com`

[4] `https://mehrarodgers.wordpress.com/2013/05/05/final-project/`

[5] `http://www.rapidtables.com/web/color/RGB_Color.htm`

9. Following Min & Cheng [114] and Potluri & Nitta [133], an assumption that the colour that has the largest number of pixels is the dominant colour of the photo and determines its category, was made. Pictures, for which one of the colours from the colour palette was dominant, constitute 20.79% of the initial sample (the rest was assigned to the category 'other colours').

As a dominant colour may be not obvious for a human eye, the acceptance threshold was set taking into account the percentage of pixels of a dominant colour in the picture: At least 70% of pixels in the photo had to belong to a dominant colour [169] to be included in the dataset. The resulting dataset was then manually checked to delete pictures with inscriptions, faces and vibrant symbols that can create bias in further research. From each of the 8 categories of photos, 4 randomly chosen pictures were selected, what resulted in a final sample of 32 photos. The scale for grading pictures in the survey held was the same as for the BFI-44 questions:

- 1 – disagree strongly,
- 2 – disagree a little,
- 3 – neither agree nor disagree,
- 4 – agree a little,
- 5 – agree strongly.

Questions on preferences towards pictures combined with the original 44 questions from BFI-44, were included in a questionnaire that was posted on the Internet.

Scoring Formula

After collecting survey results, the mathematical formula (Equation 3.1) used in the scoring process was developed. The scoring process was carried out as follows:

1. At first, a number of questions for each of the five features was determined.
2. The number of points (P) for every trait was summed up.
3. For every question, the minimal number of points was one. Therefore, the minimal number of points for each trait equals the number of the questions about it (min).
4. The middle of the scale is the number of questions about the trait multiplied by three as it was the middle of the scale (M).
5. Maximum (max) is the minimum multiplied by five which is the largest number of points that could be marked by a participant of the survey.

The result of the calculations is the percentage of a given trait (F). If it is smaller than 50%, then it is treated as a value of the opposite trait by subtracting 100% minus the given value. The same process was performed for scoring photos in order to determine how often a photo was liked.

$$F = \left(50 + \frac{100}{\max - \min}\left(\sum P - M\right)\right) [\%] \tag{3.1}$$

3.2.4 Questionnaire Findings

144 responses to the questionnaire (of which there were 92 coming from women and 52 from men) were collected. The scoring process was performed according to the BFI scoring applying the Equation 3.1. The results of scoring are presented in Figure 3.3. The vast majority of respondents liked blue photos. They graded them 66% on average and the median was close to 69%. Dark green, the second most liked colour, had the mean score equal to 56% and the median about 59%. Drawing conclusions, dark green also has a positive impact on the most of the people. On the contrary, respondents did not like yellow photos, as they scored 42% on average and the median at the level of 44%. The rest of the colours was rated around 50%. The mean and median values are presented in Figure 3.4.

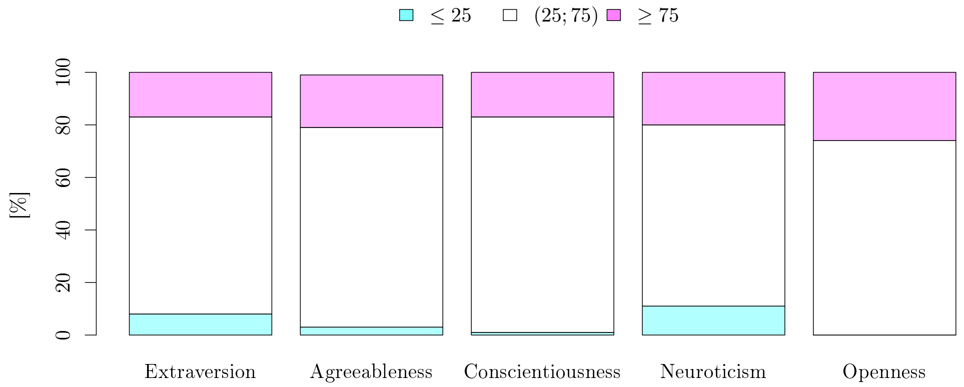


Figure 3.3: Percentage of participants by trait. Source: own elaboration

Regarding personality traits, respondents were divided into 3 groups, for each trait separately (18 in total). For Extraversion, respondents were classified as follows:

- extreme extroverts (Extraversion trait equal to 75% or more),

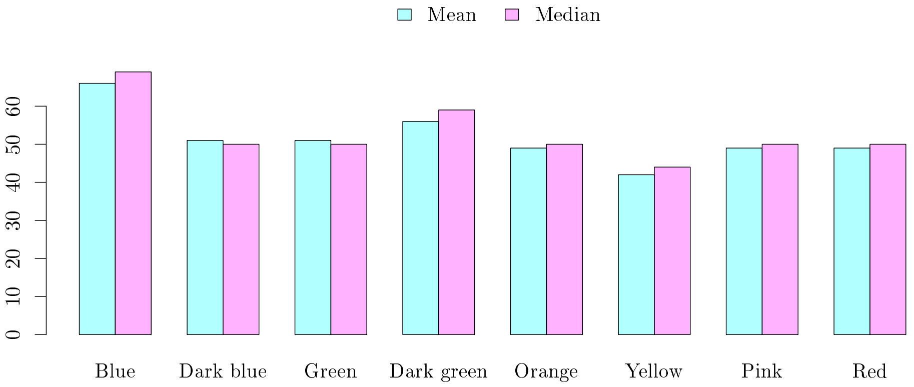


Figure 3.4: Colour preferences. Source: own elaboration

- extreme introverts (Extraversion trait equal to 25% or less),
- people who were between those two ends (more than 25% and less than 75%).

The rest of traits were treated accordingly. For each trait, the moderate (between) group was the largest and there are not many people who present extreme types. There was also no person who was classified into the group of Closeness to experience (Openness is the opposite trait) and there was only one person classified in the Lack of direction group (Conscientiousness is the opposite trait). Figures 3.5 to 3.9 depict mean colour preferences. On each plot, there are two or three series of data - one or two for the first and last quartile and one for moderate values.

Extraversion

Introverts graded most of the colours lower than extroverts and moderate people. For example, blue was marked by extroverts around 70%, moderate people graded it about 66% and introverts at the level of 60% on average. Extraversion or Introversion cannot be determined, as the differences between these two groups are too small to draw conclusions from them (none of the colours is significantly more important comparing extroverts to introverts).

Agreeableness

People who were classified as antagonistic rated dark blue higher than these with extreme high Agreeableness trait. This indicates that dark blue is a factor that differentiates agreeable and

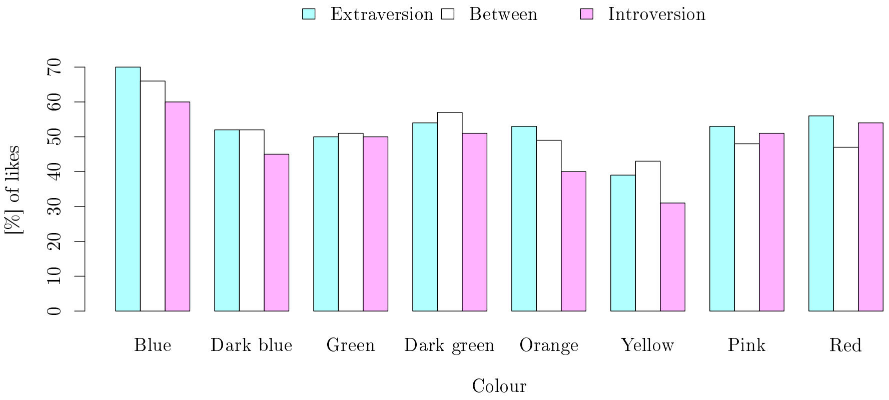


Figure 3.5: Mean colour preferences among extrovert/introvert people. Source: own elaboration

antagonistic people, because antagonistic people rated that colour about 16% higher. With regard to the analysed preferences, these colours are more significant for antagonistic people, so perhaps they are more likely to fancy them. The rest of the colours was graded lower or the same as in the Agreeableness group, except for dark blue and red.

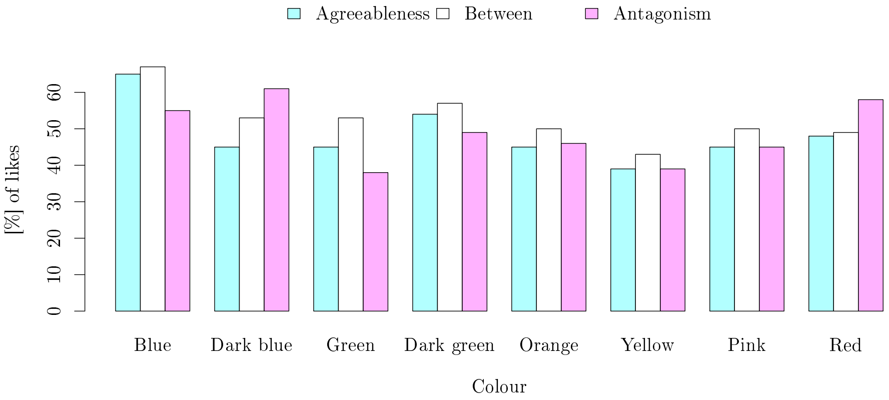


Figure 3.6: Mean colour preferences among agreeable/antagonistic people. Source: own elaboration

Conscientiousness

As only one person in the sample turned out to be a person with Lack of direction, there is no sufficient amount of data to find differences between the people with different personalities.

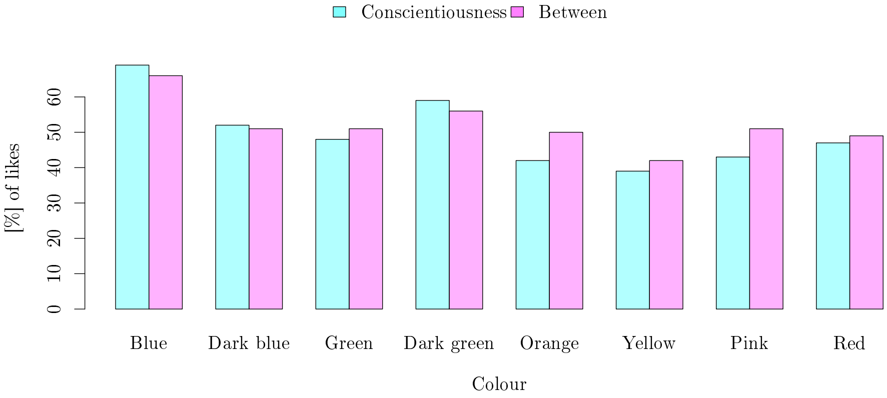


Figure 3.7: Mean colour preferences among conscientious people. Source: own elaboration

Neuroticism

Regarding people who were classified as emotionally stable, their scores for dark green and yellow were higher compared to neurotics. In the case of dark green the average was 25% higher and yellow was rated 21% higher. It leads to the conclusion that it is likely that people who are emotionally stable like more dark green and yellow colours.

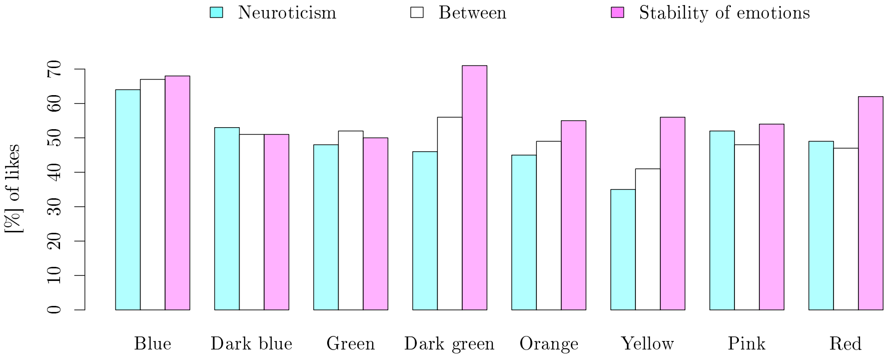


Figure 3.8: Mean colour preferences among neurotic/stable people. Source: own elaboration

Openness

None of the respondents was classified into the Closeness to experience group, therefore a comparison of these groups cannot be conducted.

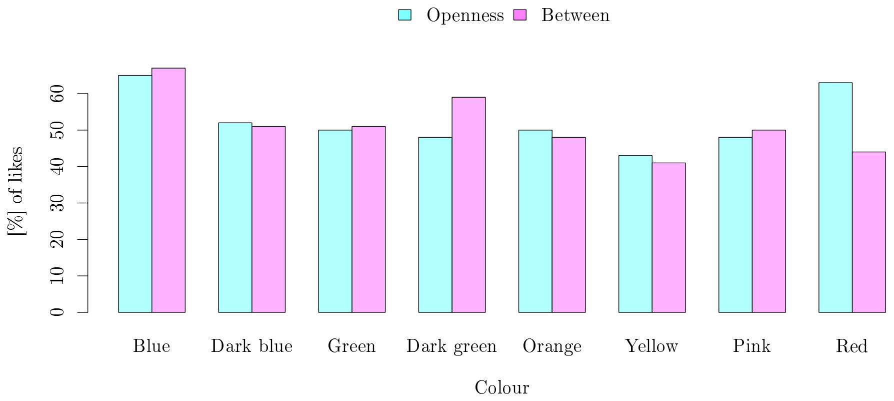


Figure 3.9: Mean colour preferences among open/closed to experience people. Source: own elaboration

3.2.5 Establishing Links Between BFI-44 and Colours: Regression

To study relation of colour preferences to personality, two diverse methods were applied, namely regression and association rule mining. In case of the first of the models created, the least-squares method was applied for obtaining the model. As a result, 5 different models (for each trait separately) were developed. We used Gretl and standard linear regression. Statistical significance was set at the $\alpha = 0.05$ level [96, 142]. To choose models which fit the data best, R^2 values and cross-validation have been used. After rejecting insignificant variables and choosing the right models, the final regression formulas are as follows:

$$\begin{aligned}\text{Extraversion} &= 0.527720 \cdot blue + 0.336480 \cdot orange\\ \text{Agreeableness} &= 0.526617 \cdot blue + 0.172206 \cdot darkgreen + 0.239923 \cdot orange\\ \text{Conscientiousness} &= 0.532727 \cdot blue + 0.215184 \cdot darkgreen + 0.192177 \cdot red\\ \text{Neuroticism} &= 0.424213 \cdot blue + 0.262295 \cdot green + 0.185821 \cdot red\\ \text{Openness} &= 0.510553 \cdot blue + 0.225479 \cdot orange + 0.337031 \cdot red\end{aligned}$$

Table 3.2: Evaluation of models created.

	Average	**Standard deviation**	**R-squared**	**Adjusted R-squared**	**F-statistics**	**P-value (F)**
Extraversion	53.51563	17.53756	0.884671	-0.196736	544.6285	2.50e-67
Agreeableness	59.22068	16.79134	0.885038	-0.554949	361.8292	5.27e-66
Conscientiousness	59.51003	15.38935	0.909029	-0.460811	469.6478	3.64e-73
Neuroticism	53.84115	19.90443	0.822532	-0.485069	217.8366	1.00e-52
Openness	64.70486	13.6531	0.915415	-0.997652	508.6554	2.16e-75

Source: own elaboration

The statistical features (quality assessment) of the obtained models are presented in Table 3.2. All variables in each model are statistically significant. Uncentered R^2 values (notice the lack of intercepts) for every model are close to 0.9 and indicate a good explanation of the dependent variables by independent variables. Observations emerging from these models are summarised in Table 3.3.

Table 3.3: Significant colours for each trait from linear regression models.

Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Blue	Blue	Blue	Blue	Blue
Orange	Dark Green	Dark Green	Green	Orange
	Orange	Red	Red	Red

Source: own elaboration

As expected, blue is significant for all personality traits, which means that this colour has a low discriminatory value. Dark green is significant for Agreeableness and Conscientiousness. In the case of Openness, red and orange were important. Some colours are not strongly linked with any of the personality traits, such as yellow and pink.

3.2.6 Establishing Links Between BFI-44 and Colours: Association Rule Mining

Association rule mining was performed in R using the Apriori algorithm [3]. As an entry requirement for rule forming, the value of support parameter was set to 0.1 and confidence to 0.75 (which means that candidates for rules with lower values of at least one of these parameters were discarded). Rule mining resulted in 28 association rules. Each of them can be perceived as an implication (transactions T) between colours and traits, which formally can be written as a set of items $I = \{i_1, i_2, \ldots, i_n\}$. We define a rule as $X \Rightarrow Y$, where $X, Y \subseteq I$. Since the algorithm does not operate on continuous values, we discretise both: traits and colour preferences using quartiles. For each trait or colour, 1st quartile contains values from 0 to 25%, 2nd quartile has a range of (25,50], 3rd quartile has values in the range of (50,75] and 4th quartile contains values from 75 to 100%. The outcomes of analysis are presented in Table 3.4 and Figure 3.10. It should be noted, that the analysis was performed altogether for all traits, and also for each trait separately.

Summarising results, orange is important for open, extravert, conscious and agreeable people, whereas red was chosen by agreeable and conscientious people. Dark blue is a factor that distinguishes extraverts, agreeable, conscientious and open people. There are also some interesting rules generated to compare preferences in case of two or more personality traits. For example, if someone is an extravert from the 3rd quartile (from medium to high), likes dark green and rather does not like red or yellow. It also means that this person has Openness value also from the 3rd quartile. Agreeable people from the 3rd quartile, who do not especially like orange and dark blue or red, have Conscientiousness value from the 3rd quartile. This rule works also for conscientious people from the 3rd quartile. If they did not like red and orange or dark blue, then they are agreeable people from the 3rd quartile. Also if someone is open (from medium to high) and does not like yellow and pink or dark blue, then this person perhaps likes blue. These results may be further compared with outcomes of our previous research. There are some colours for regression model and Apriori algorithm that turned out to be significant in both approaches. Table 3.5 sums up our findings. For Extraversion, orange colour was identified significant using both methods. Orange is also important for Agreeableness and Openness. In the case of Openness, blue and orange go together in both approaches. Red aligns with Conscientiousness and green goes with Neuroticism.

Table 3.4: Top 10 rules with the highest confidence.

LHS	RHS	support	confid.	lift
{Blue=(50,75], Dark.green=(50,75], Yellow=(25,50]}	{Openness= (50,75]}	0.1250	0.9000	1.3500
{Blue=(50,75], Yellow=(25,50], Pink=(25,50]}	{Openness= (50,75]}	0.1181	0.8947	1.3421
{Blue=(50,75], Dark.blue=(25,50], Dark.green=(50,75]}	{Openness= (50,75]}	0.1181	0.8947	1.3421
{Green=(50,75], Yellow=(50,75]}	{Openness=(50,75]}	0.1042	0.8824	1.3235
{Green=(25,50], Pink=(25,50]}	{Openness= (50,75]}	0.1389	0.8696	1.3043
{Red=[0,25]}	{Openness= (50,75]}	0.1736	0.8621	1.2931
{Blue=(50,75], Orange=(50,75]}	{Openness= (50,75]}	0.1250	0.8571	1.2857
{Orange=(25,50], Red=(25,50]}	{Cons.= (50,75]}	0.1181	0.8500	1.6320
{Orange=(25,50],Red=(25,50]}	{Agreeabl.= (50,75]}	0.1181	0.8500	1.4571
{Blue=(50,75], Dark.blue=(25,50], Orange=(25,50]}	{Openness= (50,75]}	0.1111	0.8421	1.2632

3.2.7 Summary and Conclusions

This section presents how colours can be connected with personality traits. A questionnaire based on Big Five Inventory (BFI-44) and a sample of photos (in colours from colours' palette inspired by Plutchik's Wheel of Emotions) was created. The developed models show the relation between colours and the strength of the trait. Colours and emotions related to specific personality traits

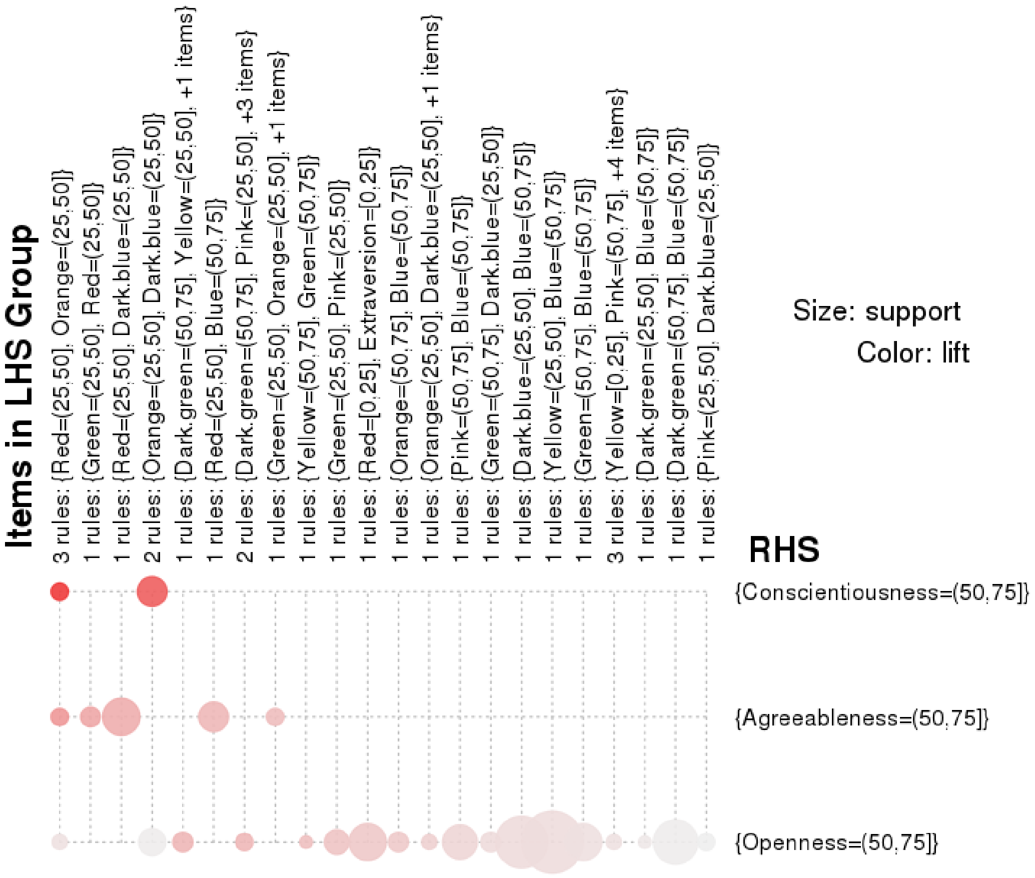


Figure 3.10: Influence of colours on a personality trait (rules with confidence exceeding 0.75). Source: own elaboration in R

may lay foundations for creating the profile of user's preferences in OSM. In our case, the data was collected from Flickr, but it could be taken from any social network that relies on images. A larger data sample might be considered in the future work. One may also test different machine learning approaches in the process of profiling. A sample from a different part of the world might be considered to spot cultural differences. It is worth to notice that results from this study can be helpful for marketing, due to revealed ties between colours and personality traits.

3.3 Modelling the Strength of Relations in Telecommunication Social Networks: A Theoretical Background

The goal of the section is to identify features that should be taken into account when studying relations between users, especially when it comes to qualitative descriptions of relations. In addition, we would like to study what impacts the strength of a relation between users in social networks, including telco social networks. These goals contribute to achieving the second goal of this chapter: "Creating a method for describing relations between users focusing on quantitative and qualitative aspects of a relation on the example of a social network". It should be underlined, that the paper included in the thesis covers the results achieved in the area of modelling strength of relations only partially, as results are described in reports from the Dynamic Social Network project carried out in cooperation with Orange SA (and are partially subject to the non-disclosure agreement).

The paper presented was submitted to the Springer Journal ”Quality & Quantity”[6] and is currently under evaluation. The identifier of the submission is: QUQU-D-18-00851.

[6] `https://link.springer.com/journal/11135`

Noname manuscript No.
(will be inserted by the editor)

Modelling the Strength of Relations in Telecommunication Social Networks: A Theoretical Background.

Bartosz Perkowski · Agata Filipowska

Received: date / Accepted: date

Abstract Theoretical modelling and empirical studies on networks of diverse entities have been a subject of a large body of recent research. Methods from the network analysis are applied in such domains as epidemiology, city planning, marketing and social networks.

Social Network Analysis (SNA) focuses on exploring structures created by individuals through relations with others. These relations may be of diverse nature and characterised by numerous attributes. Some of these attributes concern strength of a relation and may explain quality of relation between entities. Previous studies of social networking services such as Facebook and Twitter have shown, that the quantitative approach is insufficient to accurately describe the relation occurring between two individuals. It is imperative to properly model also the qualitative attributes.

Social networks in Telecommunication emerge from communication using mobile phones. This communication is registered in Call Detail Records (CDR) and based on CDRs relations between individuals can be modelled. Previous attempts to determine the strength of relations used only a number or a duration of phone calls. In this paper, an attempt was made to determine the qualitative attributes of relations, to lay a foundation for the derivation of a

Bartosz Perkowski
Poznan University of Economics and Business
Department of Information Systems
Al. Niepodleglosci 10, 61-875 Poznan, Poland
Tel.: +48 61 639 27 97
E-mail: bartosz.perkowski@ue.poznan.pl

Agata Filipowska
Poznan University of Economics and Business
Department of Information Systems
Al. Niepodleglosci 10, 61-875 Poznan, Poland
Tel.: +48 61 854 36 32
E-mail: agata.filipowska@ue.poznan.pl

new method, to better represent the phone-based relations between individuals.

Keywords Social Network · telecommunication · social relations · relation strength · quality of relations

1 Introduction

A social network is a social structure made up of actors (which can represent individuals but also organizations), dyadic ties, and social interactions between actors. It can also be defined as a social structure consisting of individuals that are connected with each other through various types of relations, and between which information is exchanged on the basis of commonly shared norms and values [1].

Social Network Analysis (SNA) provides a set of methods for analysing the structure of the social network as well as a variety of theories explaining the patterns observed in these structures. In particular, social networks are used to analyse interactions between individuals and the impact of these relations on individuals' behaviour. The sociology defines an individual as a person, a being of rational nature of his own existence [2]. In social networks, an individual may represent, depending on a profile created with the use of social website, a person, a group of people sharing common interests and an organisation. Any two individuals can form a relation. A relation is a connection between two actors of a social network. The types of relations (ties) distinguished in psychology are: friendship, love, marriage, kinship, co-workers and neighbours [3]. The ties between individuals play a crucial role in forming a social structure, thus it is important to model a social nature of every relation.

The social relation is defined as an interaction between individuals or groups, taking place in accordance with accepted practices and schemas [4]. In order to properly describe a social relation with the use of IT tools, among others, it is necessary to assign to a tie an attribute called a strength of a relation. Using the strength of relation, diverse types of relations may be distinguished. Their value can represent a type of a relation as in the following example: spouses are in a stronger relation than friends, who in turn are closer to each other than acquaintances. Thus the stronger the bond between two users, the higher the strength value.

A social network between people emerges also from communication using telecommunication services. The problem of modelling relations based on the Call Detail Records was previously described in [5]. This problem, in telecommunication social networks, is related with the identification of attributes of a connection, that influence the closeness of a relation. In order to solve the problem, attributes of connections have to be identified, which may have an impact on the relation.

In this paper, the analysis of qualitative attributes influencing relations between individuals will be presented, based on the sociological and psychological factors of human communication. **The goal is to describe features**

of a method for modelling the strength of relations. The research performed is based on the Design Science Research methodology as proposed by [23].

The structure of the paper is as follows. Section 2 presents the state of the art in the domain of modelling relations in social networks. Section 3 focuses on the qualitative attributes of relations in various social networks. The examples of psychological and sociological dimensions of a relation are presented based on the literature analysis. Then, the adaptation of these dimensions in telecommunication social networks is performed, with the specification of variables that can have a possible impact on the strength of relations. Finally, a hypotheses for each dimension are formulated. Section 4 focuses on presenting analysis of the results of an on-line survey. Section 5 provides discussion and conclusions.

2 Relations between Individuals in Social Networks

Identification of mechanisms that make social networks evolve, is a research subject that helps understanding, when a new relation between individuals emerges or when it vanishes, and how in a consequence, the structure of a social network changes in time [6]. The analysis of creation of relations should be preceded with the analysis of behaviour of individuals representing people with different characteristics. Some of these characteristics and their values are assigned to specific types of individuals (creating so called stereotypes), e.g. emotionality to women, tolerance to educated people or violence to gang members. Knowing stereotypes, relations between individuals may be better explained.

Since similarities of individuals are the reason why people contact with each other, thus one of the first theories used in relations' modelling in social networks was the theory of homophily [7]. The schemas of communication resulting from similar stereotypes of individuals are reflected in relations [8]. It was proved that the homophily measures were stronger, when more types of relations occurred between two individuals. Thus, it can be assumed that each relation that occurs based on the homophily of individuals, leads to a higher level of homophily of multiple relations [9].

In social networks, many relations are described binary, meaning the relation exists or not. Such form of a connection provides only residual information about the nature of a relation, thus the network created using such relations does not explain their strength [8]. The identification of a relation's strength can provide more information about the connection of two individuals. Social Network Analysis research proves, that social networks' models are of higher quality when the strength of relations is included [6].

The indication of strong relations depends on two factors: complexity (resulting from multiple relations between individuals, various types of interactions and different roles of individuals) and methods of measuring relation's strength. Based on the direction of a relation, a sender and a receiver can be distinguished. While modelling the directed and undirected relations, the analysis can focus on the presence (in case of binary strength of relations) or on the intensity (weighted relations). Thus, the strength of each relation can be characterised using two variables, as it is usually the resultant of existence and intensity of connection that occurs between individuals [10].

The introduction of a relation's strength started research on the influence of relation's strength on the structure of social networks. Interestingly, it was shown that not only strong relations are meaningful, but also weak relations can play a crucial role, especially in case of linking communities [7]. Moreover, the popularity of social networking services like Facebook[1] or Twitter[2] encourages research on how to properly determine the strength of relations. The analysis of individuals registered on Facebook [11] focused on the diversification of strong and weak relations. The authors pointed out, that having only two types of relations is insignificant, as these two may have intermediate features of weak and strong relations. As a result, four various types of relations were proposed based on variables grouped into six categories: interaction, affinity, time based, network, similarity and distance. Similar work was performed for the Flickr network [12].

Similar examples can be multiplied, however all of these have in common **the use of quantitative attributes to describe relations between individuals** in social networks. Already in the 1970s, Granovetter provided the following definition of the relation's strength [7]:

> "The strength of a tie (relation) is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterise the tie."

According to this definition, measuring the relation's strength should include not only the quantitative attributes, but also qualitative ones, in order to reflect the social nature of a relation. One of the first attempts of applying qualitative attributes in measuring the strength of relations in social networks is [13]. The goal of this research was to verify the following hypothesis: social media can be useful in prediction of a relation's strength based on the qualitative attributes of a relation. Authors selected seven qualitative attributes, which may have an impact on a relation: intensity, intimacy, duration, reciprocal services, structure, emotional support and social distance, and for each attribute defined, how it can be measured using various elements of the Facebook network. Finally, using questionnaire among Facebook users, authors performed the analysis of correlation between defined, quantitative attributes and measured relation's strength. The results showed that there are some variables that have an impact on the strength, including: intimacy, intensity, duration and social distance and that it is possible to determine the qualitative attributes of a relation by means of correspondingly set different numerical variables [13].

Phone calls or SMS messages can also be treated as an evidence of a human interaction, just like text messages and comments exchanged by the Facebook

[1] https://www.facebook.com

[2] https://twitter.com

or Twitter users. These interactions in the telecommunication network describe relations between individuals. There are currently two most common methods used for measuring the strength of relations based on telecommunication events, that include: number of phone calls and the total duration of phone calls. The first measure takes into account the number of outgoing phone calls from one user to another. The second measure includes the total duration of outgoing phone calls from one user to another. In addition, there is also a third method as a combination of the two [15], however none of these reflects the nature of social relations of involved users.

As a part of research on the social nature of relations resulting from telecommunication events, [16] focused on the classification of individuals to various affinity networks based on the degree of kinship and the identification of similar behaviour of individuals/groups characterizing these networks. Each individual was assigned to one of the following affinity types: *family* – any two subscribers as a part of common family accounts, *toll* – numbers starting with specified prefix, *utility* – subscribers assigned to the list of business establishments, *others* – other numbers. Based on the affinity type, the affinity networks were created containing similar individuals. Finally, the identification of characteristics within the networks was performed. These characteristics describe the behaviour of individuals based on the mobile usage and include: amount, length, frequency and engagement ratio of phone calls. Performed experiments resulted in the following characteristics of affinity networks:

- The *average number* of phone calls showed that only families stand out in the higher value.
- The *engagement ratio* describes the reciprocity of communication, with the assumption that the more symmetric relation, the higher the value. Families resulted in the highest engagement ratio, while toll and utilities focused mostly on incoming phone calls.
- Weekdays are crucial in case of non-family groups, while weekends are especially valuable for families.
- Families are characterized by the shortest conversations, longer in case of utilities and the longest for tolls.

The experiments showed that some specific characteristics exist for each type of relations in telecommunication networks. Summarizing, it can be assumed that based on the Call Detail Records it is possible to model the strength of relations between individuals including some of the presented characteristics, or even by identifying new ones.

3 Describing Relations in Telecommunication Social Networks

Behavioural profiles of individuals present how people decide to take various actions, if they will to strive to contact with others or not, how often do they travel, how much time do they spend on maintaining relations with others, etc. The need to identify behavioural patterns for individuals in social networks is justified, because individuals tend to create and develop relations with others with similar profiles, also behavioural [8]. A relation between individuals depends on their mutual closeness, because individuals that have a stronger relation (informal, that results from emotions and feelings, and formal, related to the organizational structure) tend to contact each other more often, than in case of a week relation.

Another aspect of a relation having influence on strength is synchronization, which shows how two individuals interact with each other. Two types of synchronization can be distinguished: bilateral and unilateral. While a bilateral relation means individuals contact each other (communication takes place in both directions), the unilateral relation bases on long messages only in one direction. The length of these messages is crucial in order to identify unilateral nature of a relation.

More attributes that originate from sociology and apply to an activity of an individual within a social network to describe a relation are as follows [17]:

- response time – the time between the event and individual's answer,
- response probability – based on how often an individual responds,
- response priority – indicating the order in which an individual responds to events,
- status – informing when an individual is available,
- list of contacts,
- overall amount of events.

Depending on the emotions concerning relations of individuals, three levels of involvement can be distinguished: (1) concerning only safe topics, which do not need any emotional involvement, (2) focusing on communication aimed at achieving appropriate behaviour of a receiver based on the impact it is having, and (3) requiring emotional involvement leading to mutual understanding of both individuals [3]. The stronger the emotional involvement, the stronger the relation is.

In psychology, each relation, independently of its type and level, can be described using following characteristics [18]:

- dynamics – indicating the intensity of messages exchanged between individuals,
- continuity – informing, if breaks in the communication occur,
- interactivity – based on which the relation can be described as partnership (assuming equal division of incoming and outgoing messages),
- domination (where the division of messages is outbalanced).

Moreover, important aspects of relations are: a way of transmitting a message that may be informative or convincing, and the ability to understand a message that depends on the recipient, location, time, and a method of transmission.

Summarising, various approaches to describe interpersonal relations allow to identify seven dimensions that have an impact on the strength of relations (from perspectives of both psychology and sociology of relations): amount of time devoted to a relation, intimacy, intensity, reciprocity [7], membership in informal social groups [20], emotions [21] and social distance [22]. However, the problem of determining the strength of a relation remains. The question is how these dimensions are represented by the data collected from social networks.

Telecommunication events described in CDRs enable to determine some of the above mentioned dimensions. For example, reciprocity relates to incoming and outgoing events e.g. calls. Because in CDRs a sender and a receiver are precisely defined, the direction of communication is clearly visible. With the knowledge about the direction, the reciprocity can be measured using the psychological models of communication. To examine the reciprocity of communication, the following attributes should be included: direction of events, type of a service used in communication and a number of events exchanged between individuals. In our research the assumption was made based on literature that **unilateral relation should be represented with the weakest strength, and along with the increasing number of returned events, the strength of a relation should rise up to the point, where the number of events going in both directions is equal.**

Another dimension of a relation that may be described using CDRs is the level of intimacy. Psychological literature indicates that a verbal communication is a foundation of interpersonal relations, by focusing not only on the information exchange and understanding, but also on emotions or a voice tone [18]. According to this assumption it can be expected, that similar situation may occur in case of communication using mobiles, thus a phone call should be more intimate than a text message. As a result, **two hypotheses regarding the intimacy of relations were formulated, regarding the type of service used in communication between mobile users and the duration of communication.** Longer communication may indicate, that more information is exchanged and potentially the conversation is more intimate. It should be noted, however, that business communication may also engage in lengthy conversations, while the subjects of such conversations are not concerned with private matters. Therefore, a hypothesis was also extended to distinguish between private and business time within 24 hours of a day, assuming that private topics are not addressed during the working hours. As a result, the time of a day may also be another important attribute in determining the strength of a relation.

Closeness indicates how much two individuals are related, with the assumption, that the closer the relation is, the higher its strength is. The closeness of a relation focuses on the pursuit of individuals to contact others. In order to clarify, which attributes are related with the closeness, two hypotheses have to be verified. Firstly, it is assumed that **the closer the relation, the more often individuals tend to contact each other and exchange more information, often more intimate and emotional**. Another dimension of closeness may be explained by the number of answered and not answered phone calls. Hypothetically, **to reflect the actual closeness of two individuals, their behaviour can be analysed by comparing a share of calls answered and to all calls** (this share should be close to 1). The effect of different types of not answered calls (e.g. single signals, dropped and missed calls) may be also analysed.

In addition, it is worth mentioning that the distance (geographical) between individuals that may be identified using CDRs may also explain a quality of relation. The distance is related to the geographic position of individuals relatively to each other. In order to study, whether in the case of telephone communication the distance is influential, a hypothesis has been formulated. **Supposedly, a bigger distance between two individuals staying in a strong relation should increase the intensity of communication.** Verification of the hypothesis is to confirm or exclude the dependency between the intensity of the communication and the geographical distance.

Table 1 summarises the dimensions along with the hypotheses for each dimension regarding the strength of relations and the attributes that may have an impact on this strength. For four selected dimensions: *reciprocity*, *intimacy*, *closeness* and *distance*, five attributes have been identified: *type of a service*, *duration of a phone call*, *number of telecommunication events*, *time of a day* and *location of individuals*. Each of the attributes is connected with at least one dimension.

Table 1 The dependencies between the attributes of communication and sociological/psychological dimensions influencing the relation strength.

Dimension	**Hypothesis**	**Communication attributes**
Reciprocity	The strength of a relation depends on the share of incoming and outgoing calls.	Type of a service Number of events
Intimacy	A phone call is more important than a text message.	Type of a service
	Time of the day affects the subject of a phone call.	Number of events Time of the day
	The strength of a relation depends on the duration of a phone call.	Duration
Closeness	The strength of a relation depends on the frequency of communication.	Number of events
	Not answered phone calls are an important aspect in measuring the strength of a relation.	Number of events
Distance	Frequency of communication depends on the distance between individuals.	Location

Source: own elaboration

4 Verification of Importance of Communication Attributes on the Strength of Relations

In order to perform a statistical verification of hypotheses formulated in Section 3, a questionnaire was prepared to collect information from users of telecommunication services. The questionnaire was firstly analysed by a team of experts to make sure, that all hypotheses can be verified. The final version of the questionnaire was published online using *ankieta+*[3]. The survey was conducted between September 2015 and January 2016. 390 users of mobile phones took part in the survey. The respondents were divided into 6 groups based on age. The majority of them represented groups from 26 to 50 years of age. The most of respondents (306) had a subscription to telecommunication services and also had a full time job with specified hours (nearly 56%), what made possible to determine the business and the private time. The results of the initial analysis of data showed, that there are no correlations between any of attributes of a person e.g. age and the answers to questions related to the qualitative attributes of relations. As a consequence, the breakdown of respondents by any feature was not used in the further analysis.

The following subsection presents the results of a statistical analysis of hypotheses formulated for qualitative attributes, as presented in Table 1. The analysis of every question was preceded by description of motivation for the question, where the purpose of the analysis was defined and possible variants of the answers were discussed. Each question is presented together with a research hypothesis, which is later verified using a statistical test. Based on the obtained results, the conclusions on the influence of chosen qualitative attributes on the strength of the relations between individuals in the telecommunication social network are specified.

4.1 Statistical Verification of Formulated Hypotheses

Question 1. Which of the following forms of communication has a bigger impact on creating relations between individuals?

When communicating using mobile phones, a few types of services can be used to exchange information between individuals. The most popular are: phone calls and text messages. The purpose of this study was to analyse the impact of the usage of each type of a service on interpersonal relations (this terminology was more understandable for respondents than the notion of strength of relations). For the question, there were three possible answers:

- phone call – indicates that this service has a bigger impact on the strength of a relation,
- text message – means that text form is more valuable,
- both – both, phone calls and text messages have the same impact on the strength of relation.

In order to investigate the dependency between the service being used and the emergence of relations between individuals, a χ^2 test for independence for the null hypothesis H_0 was carried out. The H_0 was as follows: *the development of interpersonal relations does not depend on the type of a service used by individuals to contact each other*. Accepting the null hypothesis would mean that the type of a service used, does not affect the relation between individuals.

[3] http://www.ankietaplus.pl/

Alternative hypothesis H_1 states that *the type of a service is important in the context of managing relations.* Table 2 shows the calculation of χ^2 statistics.

Table 2 Statistical χ^2 test for question 1.

Answers	**Observed value** (O_i)	**Expected value** (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Phone call	223	130	93	8649	66.53
Text message	18	130	-112	12544	96.49
Both	149	130	19	361	2.78
Total	390	390			**165.8**

Source: own elaboration

On the significance level of 5% and with 2 degrees of freedom, the test statistic is improbably large, thus the null hypothesis should be rejected in favour of the alternative one. As a result, the **importance of including the type of service in measuring the strength of relations should be noticed.** In addition, the following regularities can be observed:

- Phone calls are a dominant service and therefore have the greatest impact on forming relations.
- More than 38% of respondents indicated that both services have the same impact on development of relations.
- Text messages are only valuable, if they occur together with phone calls.

In conclusion, the key service that influences the strength of relations for mobile communication is a phone call, while the text message can only be considered as an enhancing element of a relation, when it occurs is in conjunction with a call.

Question 2. Do you agree with the sentence that each minute of a phone call longer than a specific duration, e.g. 10 minutes, has a lower impact on the interpersonal relations?

The goal of this question was to analyse, if there is a correlation between the length of a phone call and the strength of a relation between two individuals. A theoretical assumption with a time limit of 10 minutes was made, which is the reference point for evaluating the importance of a phone call by the respondents. Thus, each respondent had one of three possible options to choose:

- Yes, every next minute is less important – indicates a threshold in the duration of a phone call above which every minute has a lower importance.
- No, every minute is equally important – indicates there is no such threshold, thus every minute has the same value for establishing the relation, irrespectively of the duration.
- No, every next minute is more important – again, there is a threshold in the duration of the phone call, however, after this limit, every minute impacts the strength of relation in a greater manner.

The null hypothesis H_0 for a χ^2 test was formed: *the strength of a relation between two individuals does not depend on the duration of a phone call.* An alternative hypothesis H_1 indicates that there is a dependency between the length of a phone call and the strength of a relation. Table 3 shows the calculation of χ^2 statistics.

Table 3 Statistical χ^2 test for question 2.

Answers	**Observed value (O_i)**	**Expected value (E_i)**	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Yes, every next minute is less important	101	130	-29	841	6.47
No, every minute is equally important	246	130	116	13456	103.51
No, every next minute is more important	43	130	-87	7569	58.22
Total	390	390			**168.2**

Source: own elaboration

The results show, that with a significance level of 5% and 2 degrees of freedom, the test statistic is very large, what causes the rejection of the null hypothesis. On this basis, **it can be assumed that the duration of a phone call is influential when forming relations**. Moreover, by analysing the distribution of responses, the following can be noticed:

- Almost $\frac{2}{3}$ of respondents indicated the equal value of every next minute of a phone call on a relation.
- The strength of a relation is proportionally dependent on the length of a phone call, and thus increases linearly from the first minute.
- By assuming the linear growth of the strength, there is no time limit after which the influence of the length of the conversation on the strength of relations changes.

Question 3. Do you think that not answered calls should be taken into account while measuring the strength of relations?

Another qualitative aspect concerns the not answered calls and their impact on the relation's strength. The respondents had to choose one from three options. It should be noted that two responses indicate the impact of not answered calls, while one remains neutral:

- Yes, the not answered calls impact negatively the strength of a relation (reduce the strength) – means that for a receiver the not answered phone call is treated as an undesirable, causing a weakening of the relation.
- Yes, they impact positively (increase the strength of a relation) – the opposite reaction, for the receiver it is an information about the a contact attempt, and therefore allows the receiver to assert that the caller is seeking to contact him, what strengthens the relation.

– No, they do not impact the relation – not answered calls are neutral for the receiver and do not strengthen or weaken the relation.

To verify the results, for a χ^2 test for independence the following null hypothesis H_0 was formulated: *the strength of a relation is independent from the number of not answered phone calls.* This means that regardless of the number of not answered calls, the strength of a relation remains unchanged. The alternative hypothesis H_1 would mean that statistics on the not answered phone calls should be included while describing the relation. The results of the χ^2 test are presented in Table 4.

Table 4 Statistical χ^2 test for question 3.

Answers	**Observed value (O_i)**	**Expected value (E_i)**	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Impact on the strength (positive and negative)	206	195	11	121	0.62
No impact on the strength (neutral)	184	195	-11	121	0.62
Total	390	390			**1.24**

Source: own elaboration

Assuming the significance level of 5% and 1 degree of freedom, the test statistic is low, thus there is no reason to reject the null hypothesis. Taking into account the results, a **hypothesis about the independence of the relation's strength and the number of not answered calls was confirmed**. This means that while modelling a relation, we may ignore the not answered calls as they do not influence the quality of a relation.

Question 4. Does more frequent phone communication between two individuals (both in the form of phone calls and text messages) reflects a closer relation?

From a sociological point of view, individuals staying in a close relation tend to contact one another more often. This question in the survey was formulated to analyse, if a similar dependency occurs in case of communication using mobile phones. The respondents had two possibilities to choose from: yes – which means compliance with the above mentioned sociological assumption, or no – contradicting the claim that a stronger relation means more frequent phone contact. A χ^2 test for independence for the null hypothesis H_0 was carried out: *the strength of relations does not depend on the frequency of communication between individuals.* An alternative hypothesis H_1 assumes the dependence between the frequent communication and closer relation. Table 5 depicts results of χ^2 test.

On the significance level of 5% and 1 degree of freedom, the test statistic is improbably large, thus the null hypothesis should be rejected in favour of the alternative hypothesis. **The majority of respondents indicated that the frequency of communication is meaningful in relation between individuals, thus the number of telecommunication events should be**

Table 5 Statistical χ^2 test for question 4.

Answers	**Observed value** (O_i)	**Expected value** (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Yes	333	195	138	19044	97.66
No	57	195	-138	19044	97.66
Total	390	390			**195.32**

Source: own elaboration

included while modelling strength of a relation, with the favour of pairs having a high rate of contact.

Question 5. Indicate, which type of a communication model has a bigger impact on the creation of interpersonal relations.

Analysis of the related work results in investigating two models of communication: uni- and bilateral, with a specified impact of both on the creation of relations. To prove that one of these models applies better while studying social network built on communication using mobiles, two responses were specified:

- The bilateral model of communication, when both individuals are active – both perform phone calls and send text messages, is crucial to properly build a relation.
- The unilateral model of communication, when only one individual is active (calls and sends messages) and second one is passive (only receives), is sufficient for modelling relations as a relation may emerge from an initiative of just one individual.

The null hypothesis for χ^2 test of independence was: *the interpersonal relations in telecommunication and their strength are not determined by the model of communication.* In case of an alternative hypothesis, the communication model will have to be identified. Table 6 presents the results of χ^2 test.

Table 6 Statistical χ^2 test for question 5.

Answers	**Observed value** (O_i)	**Expected value** (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Both individuals are active (bilateral model)	376	195	181	32761	168.01
One individual is active (unilateral model)	14	195	-181	32761	168.01
Total	390	390			**336.02**

Source: own elaboration

Using the χ^2 statistical test, with the significance level of 5% and 1 degree of freedom, the calculated theoretical χ^2 was extremely large, what made us to reject the null hypothesis and accept the alternative one. Over 95% of the respondents chose the bilateral model of communication as the one adequate to properly model the relations between users. **Thus, it is important to distinguish the direction of telecommunication events and include**

the number of outgoing and incoming events between individuals. As a result, the proposed method of measuring the strength of relations using qualitative attributes should favour the bilateral relations.

Question 6. At what time do you usually make the private telephone conversations?

People, who are professionally active, devote specific parts of a day to business and private activities. As a result, business time may be correlated with more frequent contacts with associates or business partners and concern more formal conversations. On the contrary, in the private time people have meetings with family and friends, during which informal and more intimate topics are discussed. The goal of this question was to identify, how the time of a day may impact the individual's tendency to conduct private phone calls. The respondents had to choose one answer to the above question from the following:

- During business time – private phone calls are performed at work.
- Not in business time – an individual focuses on private phone calls during the private time, thus during the business time he focuses only on creating and maintaining business relations.
- Regardless of the time of a day – the answer describes a situation, when dividing the day into different parts makes no sense, as the private phone calls occur during the whole day.

A χ^2 test for independence was carried out to confirm/reject the following null hypothesis H_0: *the time of a day does not have an impact on the subject of a phone call.* It means that regardless of the time of a day, individuals conduct private calls. The alternative hypothesis H_1 indicates the dependency between phone calls' topics (and contacted individuals) and the time of a day. Table 7 shows the calculation of χ^2 statistics.

Table 7 Statistical χ^2 test for question 6.

Answers	**Observed value** (O_i)	**Expected value** (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
During business time	10	130	-120	14400	110.77
Not in business time	190	130	60	3600	27.69
Regardless the time of a day	190	130	60	3600	27.69
Total	390	390			**116.15**

Source: own elaboration

With the significance level of 5% and 2 degrees of freedom, the theoretical value of the χ^2 test was very large, what leads to the rejection of the null hypothesis. In addition, by analysing the percentage of responses, it can be seen that two options have an equal number of answers: having private phone calls during the whole day and just after the business time. On this basis, the following conclusions can be drawn:

- The subject of a phone call depends on the time of a day.
- Private phone calls are more often conducted after work.
- In case of communication during the private time, there is a probability that individuals contact each other also during the business time.
- Communication in business hours usually concerns only business relations (business hours are not the time of having mostly private conversations, not taking place after work).

To summarize, the results showed that **there is a need to distinguish between business and private time, with the assumption about a greater importance of communication during the private time for modelling personal relations.**

Question 7. Do you agree with the statement that a bigger distance between two individuals (e.g. emerging from going on holidays or a business trip) leads to more frequent phone communication?

The last analysed dimension of relations is the impact of geographical distance on frequency of contact between individuals. Assuming that the distance impacts the communication, the goal was to check the correlation between the geographical distance and frequency of telecommunication events. There were four possible answers to the question:

- Yes, in case of close relations – more intense phone communication between individuals staying in a close relation.
- Yes, in case of further relations – individuals tend to contact more often with others staying in weak relations.
- Yes, in case of all relations – individuals always have more intense phone communication when the distance gets bigger, independently from a relation type.
- No, in none of cases – it denies the correlation between the distance and the communication frequency.

The null hypothesis was: *the frequency of a contact with the use of telecommunication services does not depend on the distance between individuals* and the alternative one: *the geographical distance impacts the frequency of communication.* Table 8 presents the results of χ^2 test.

Table 8 Statistical χ^2 test for question 7.

Answers	**Observed value** (O_i)	**Expected value** (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Yes, in case of close relations	281	97,5	183,5	33672,25	345.36
Yes, in case of further relations	5	97,5	-92,5	8556,25	87.76
Yes, in case of all relations	56	97,5	-41,5	1722,25	17.67
No, in none of cases	48	97,5	-49,5	2450,25	25.13
Total	390	390			**475.92**

Source: own elaboration

On the significance level of 5%, the calculated χ^2 theoretical statistics was extremely large, what resulted in a rejection of the null hypothesis in favour of the alternative one. Analysis of responses enables to draw the following conclusions:

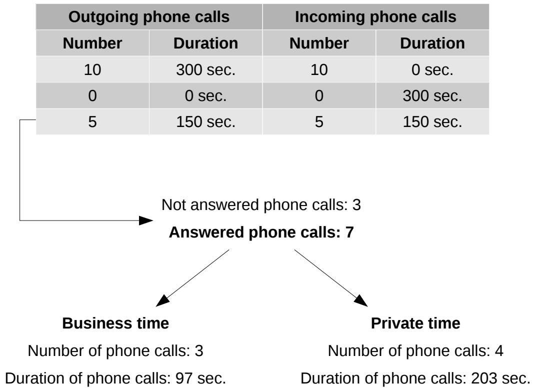
- Almost $\frac{3}{4}$ of respondents indicated, that increased intensity of phone communication is due to a greater distance between individuals in case of closer relations.
- A small share of the responses indicates that other types of relations are not reflected in frequent communication in case of the increase of geographical distance.
- Analysing the strength of a relation should take place over a relatively long period of time, which allows to determine the distance that most of the time is shared by both individuals and, consequently, allows the identification of the intensity of communication that takes place when this distance rises.

5 Discussion and Conclusions

The results indicate that while describing relations that occur in telecommunication social networks, it is important to precisely analyse quantitative attributes of communication based on CDRs as they may be translated into qualitative descriptions of relations within the social networks. Figure 1 depicts three sample relations resulting only from phone calls to validate impact of our research results. The first relation is derived only from outgoing events, the second one focuses only on incoming calls and the third one depicts a bilateral communication. Using the basic and commonly used approach to measure the strength of a relation, each of these exemplary relations has the same value of strength. Even, if for the first two relations this result is valid, in case of the third one concerning the outgoing and incoming calls, the value should be different. This is due to the bilateral model of communication, which, according to the results of the research conducted, has a stronger impact on the creation of interpersonal relations. The presented example, shows the need for a different approach to determine the strength of relations between individuals than a commonly used one.

Based on the obtained results, the following conclusions can be formulated. The *reciprocity* of a relation emerges from the telecommunication events that occur in both directions between any pair of individuals. The measure of the relation's strength should include the direction of events and separately count outgoing and incoming phone calls and text messages.

The *closeness* of relations can be characterised using two variables: frequency of communication and the number of not answered phone calls. According to the results of the survey, not answered phone calls do not impact the relation's strength and they should not be included while estimating closeness. Thus, the attribute that defines closeness should concern only the number of events occurring between individuals. Of course, in case of events both, phone



Outgoing phone calls		Incoming phone calls	
Number	Duration	Number	Duration
10	300 sec.	10	0 sec.
0	0 sec.	0	300 sec.
5	150 sec.	5	150 sec.

Fig. 1 Parameters of a phone communication useful for the determination of qualitative attributes of a relation between individuals. Source: own elaboration

calls and text messages have to be included with the assumption, that phone calls are the basis of each social relation and text messages are used only to increase the strength of an already existing relation. It has to be noticed, that *closeness* depends on the frequency of communication, and the number of phone calls informs how many times both individuals contacted each other. In order to precisely describe the frequency, not only the number of telecommunication events has to be included, but also the period when the activities took place. The activity period was previously used by [Kumar, Rao, Nagpal 2012] to describe the time when two individuals exchanged messages using Facebook. Similarly, activity period in telecommunication social networks can be specified as the number of days, during which two individuals have outgoing and incoming events. Thus, it was assumed that the relation between the two individuals is closer, the more often the communication between them occurs.

The last analysed qualitative attribute of the relation is *intimacy*. The intimacy is represented by the share of communication in business and private time, and also by the duration of phone calls. Both of these features were confirmed by a survey. The results of the survey depicted that every minute of conversation increases the strength of the relation equally. In this case, the duration can be considered in two ways: as the sum of the lengths or as the average length of all phone calls between two individuals. Moreover, the results of the analysis indicated that individuals conduct more intimate conversations in a private time. Thus, while modelling the *intimacy* of relations, the time of a day must be included to calculate the share of telecommunication events occurring during the private and the business time. Due to the significance of the time of day, it has to be assumed that events occurring at different times, even if they are of equal value, should have a different effect on the final strength of the relation.

The future work may focus on the estimation of the qualitative attributes on data and modelling the final formula of the method for estimating the strength of a relation between individuals in a telecommunication social network. Using the conclusions from the survey-based research, for each qualitative dimension of a relation, quantitative parameters were identified that impact one or more of these dimensions. Thus, in order to measure the strength of a relation and reflect the real-world relation with a value, for each of the dimensions a formula has to be developed. Finally, a developed method has to be tested on a sample of telecommunication data.

References

1. Krupski R. (ed.), Zarzadzanie strategiczne. Koncepcje. Metody., Wydawnictwo Akademii Ekonomicznej we Wrocławiu, edition V (2003)
2. Elias N., Społeczeństwo jednostek, PWN (2008)
3. Argyle M., The Psychology of Interpersonal Behaviour, PWN (2002)
4. Znaniecki F., Hałas E., Relacje społeczne i role społeczne: niedokończona socjologia systematyczna, PWN (2011)
5. Perkowski B., Filipowska A., Modelling the Strength of Relations in Telecommunication Social Networks, Studia Ekonomiczne, 234, 140–151 (2015)
6. Liben-Nowell D., Kleinberg J., The Link Prediction Problem for Social Networks, International Conference on Information and Knowledge Management (2003)
7. Granovetter M., The Strength of Weak Ties, The American Journal of Sociology, vol. 78, no. 6, 1360–1380 (1973)
8. McPherson M., Smith-Lovin L., Cook J.M., Birds of a Feather: Homophily in Social Networks, Annual Review of Sociology (2001)
9. Fischer C.S., To Dwell among Friends. Personal Networks in Town and City, Homogeneity in Personal Relations: Stage in the Life Cycle, Chicago Press (1982)
10. Ansari A., Koenigsberg O., Stahl F., Modeling Multiple Relationships in Social Networks, Journal of Marketing Research (2011)
11. Kumar A., Rao T., Nagpal S., Using Strong, Acquaintance and Weak Tie Strengths for Modeling Relationships in Facebook Network, Contemporary Computing, vol. 306, Springer, 188–200 (2012)
12. Zhuang J., Mei T., Hoi S.C.H., Hua X.S., Li S., Modeling Social Strength in Social Media Community via Kernel-based Learning, ACM Multimedia (2011)
13. Gilbert E., Karahalios K., Predicting Tie Strength with Social Media, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 211–220 (2009)
14. Kazienko P., Ruta D., The Impact of Customer Churn on Social Value Dynamics, International Journal of Virtual Communities and Social Networking, vol. 1, no. 3 (2009)
15. Zhang M., Handbook of Social Network Technologies and Applications, Social Network Analysis: History, Concepts, and Research, Springer, 3–23 (2010)
16. Motahari S., Mengshoel O.J., Reuther P., Appala S., Zoia L., Shah J., The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records, Proceedings of the Sixth Workshop on Social Network Mining and Analysis (2012)
17. Karagiannis T., Vojnovic M., Behavioral profiles for advanced email features, Proceedings of the 18th International World Wide Web Conference, Association for Computing Machinery (2009)
18. Morreale S., Spitzberg B., Komunikacja między ludźmi. Motywacja, wiedza i umiejetności, Wydawnictwo Naukowe PWN (2007)
19. Wood J.T., Communication in our lives, Cengage Learning, edition V (2009)
20. Burt R., Structural Holes: The Social Structure of Competition, Harvard University Press (1995)

21. Wellman B., Wortley S., Different Strokes from Different Folks: Community Ties and Social Support, The American Journal of Sociology, vol. 96, no. 3, 558–588 (1990)
22. Lin N., Ensel W.M., Vaughn J.C., Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment, American Sociological Review, vol. 46, no. 4, 393–405 (1981)
23. Hevner A., March S. T., Park J., Ram S., Design science in information systems research, MIS quarterly, vol. 28, no. 1, 75–105 (2004)

Table 3.5: Comparison of two methods' results.

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Linear model	Blue Orange	Blue Dark Green Orange	Blue Dark Green Red	Blue Green Red	Blue Orange Red
Rule Mining	Orange Dark Blue Dark Green	Blue Dark Blue Green Red Orange	Orange Red Dark Blue Dark green Pink	Yellow Green	Blue Dark Blue Green Dark Green Red Orange Pink Yellow

3.4 Conclusions

This chapter belongs the part of the thesis studying definitions and methods of profiling, being also the background for the following chapters. Its main goal was to ”analyse profiling methods that enable for describing a user/a thing or relations between users”. This goal was further translated into two secondary goals addressed by specific sections of this chapter, targeting at:

G2.1 Describing a user profile w.r.t. user personality and his/her colour preferences and studying relations between personality traits and user colour preferences using different methods of analysis.

G2.2 Creating a method for describing relations between users focusing on quantitative and qualitative aspects of a relation on the example of a social network.

In relation to the goal G2.1, the goal of the Section 3.2 was to analyse personalities of users (expressed as personality traits using the Big Five Inventory) and their colour preferences. The results described are based on a survey in which participated 144 respondents. Then, both personality traits and colour preferences of these respondents were modelled. This was to indicate what may be the colour preferences of people with a specific personality type. The data was studied using regression and association rule mining methods. Some colours such as blue or orange were confirmed to be significant for people with a specific personality type, using both methods. This chapter however presents how a profile may be built when user data is available. It is worth to underline that this data may be also collected by studying behaviour of a user.

The goal of the paper addressing the G2.2, was to identify features that should be taken into account when studying relations between users, especially when it comes to qualitative descriptions of relations. The supplementary goal was to study what impacts the strength of a relation between users in social networks, including telecommunication social networks. The paper included in Section 3.3 provides an analysis of the related work in the area of modelling relations between users. Different types of relations are presented and typical approaches used for describing these relations in social networks are discussed. Then, based on a survey carried out between 306 respondents, it is confirmed what features emerging from call logs should be taken into account while working on a description of a relation. These features underline importance of such aspects as e.g. timing of the contact, geographical distance between people, communication channel used or initiator of the contact, for proper modelling of a relation. Taking these features into account, the initial version of a method for describing strentgh of a relation is also proposed.

The following chapters of the thesis focus on vertical and horizontal application scenarios for profiling, showing also an application scenario taking benefit of the method for assessment of a strength of a relation between two entities.

Chapter 4

Profiling of People and Things for Utilities: Smart Grid

4.1 Public Utility: Introduction

Public utility is defined by Encyclopaedia Britannica as an ”enterprise that provides certain classes of services to the public, including common carrier transportation (buses, airlines, railroads, motor freight carriers, pipelines, etc.); telephone and telegraph; power, heat, and light; and community facilities for water, sanitation, and similar services” [36]. Other definition says that ”a public utility is a company that operates as a public-service corporation, and provides essential services to the public such as electricity, telephone service, natural gas, water or postal services” [108]. These definitions and all other that might be spotted in the literature do not differ much and underline the following features of a public utility [2]:

- providing essential services to the public,
- benefiting from the inelastic demand for services, meaning that there is not much influence of external factors on the demand (people need water or electricity for their daily life, and even if price increases there is a dominating part of the demand that is stable),
- often forming part of a natural monopoly as there are not many companies offering substitutional services usually due to a high barrier of entrance,
- owned or regulated by the national, state of local government,
- benefiting from the economy of scale: because of small or no competition, they have multiple customers and can better manage their fixed costs and prices of services offered.

In many countries companies offering public services are either state-owned or operated by the state, however it might be that the public utility is not owned or operated, but only regulated by a national, state or local government. Public utility companies offer diverse services or products: water, natural gas, telecommunication, sewage, etc. However, analysing them more closely shows that although products or services are different, these companies are similar when it comes to operation or challenges they face.

4.1.1 Challenges Faced by Public Utilities

Public utility companies for many years have been perceived as the privileged ones. Supported by governments, offering services everyone needs or desires, could shape prices and benefit from the economy of scale. However, this time finished and nowadays these companies face challenges similar to other market players.

The major, from the perspective of their long-run existence, is the issue of incomes and costs of functioning. The rise in demand, e.g. for electricity in the European Union, has been lower than the general economic growth in recent years (in developing countries this trend is opposite but other phenomena have to be taken into account). This is because of the changes in machines used by the industry, less energy-intensive industries, demographic factors and the increased awareness of people (especially concerning ecology). This leads to decreasing the costs of everyday functioning, especially that because of emerging competition and government regulations, the price cannot be easily increased. In many countries, the government limits how much the public utility can charge consumers, and may insist that even those who cannot afford to pay the market price are still provided with the service [108].

Public utilities face also the challenge of the competitive market. Regulations made the infrastructure being owned by a separate company than the one providing a service and access to this infrastructure should be granted for all companies interested in provision of a service. Such a change made a customer choose, who will offer him/her the service and started the competition regarding the price and quality of service offered [108].

Quality of service is an important factor for today's customer, especially if prices are similar. Customers are more demanding and connected, and expect to be provided not only with a product (water, gas, energy), but also with sophisticated management services [41]. These management services translate into the improvement of the customer service for activities/processes like paying bills, but also getting real time information on services that may influence his/her behaviour [103].

This is also the area where profiling techniques come in place.

One should not forget other challenges such as problems with acquisition of skilled workforce, needed investments in the infrastructure, Internet of Things changes in homes of customers, etc. [135]. There is also one not frequently mentioned issue that concerns the data quality, even of basic customer data possessed and processed by the public utility companies.

4.1.2 Profiling for Public Utilities

Segmentation and profiling are means of enriching the customer experience and bringing the customer service to the next level. Regarding public utility companies, not only customers may be profiled, but also services to further align them to requirements of people. For example, [23] focused on profiling urban activity hubs using transit smart card data to improve city planning and daily operations.

Though profiling is supposed to target individuals, both business and individual customers share similar expectations. For example with regard to the electric energy, they want to reduce consumption, and they know that technology and data analytics can help them do achieve this goal [41]. Therefore, owners of commercial buildings install energy monitors e.g. to predict future requirements. In addition, companies set efficiency targets e.g. Procter & Gamble (100% renewable energy to power the plants) or Walmart (till 2020 reducing building energy intensity by 20 percent from 2010 levels) [41].

Another issue concerns managing relations with customers. Recently, public utility companies were hit by the concept of ”retailization” [134], meaning the development of more direct consumer-to-utility relationships, similar to best practices from such domains as consumer banking or online shopping. Some examples for the electric energy services/market concern e.g. real-time mobile and digital experience, energy efficiency audits, home energy management solutions, and real-time billing and mobile payments [41].

In telecommunication, profiling is important not only to provide better customer service, but also to improve customer retention [80]. In order to prevent churn, it is important to analyse all relevant customer data and develop programmes enabling retention of clients.

It should be noted that even in case of public utility companies, customers are not a homogeneous group, but rather a multi cohort population with different needs, aspirations and expectations towards a product or service offered. Of course, on the market many different classes of software enabling segmentation of customers are offered, however customers on top of

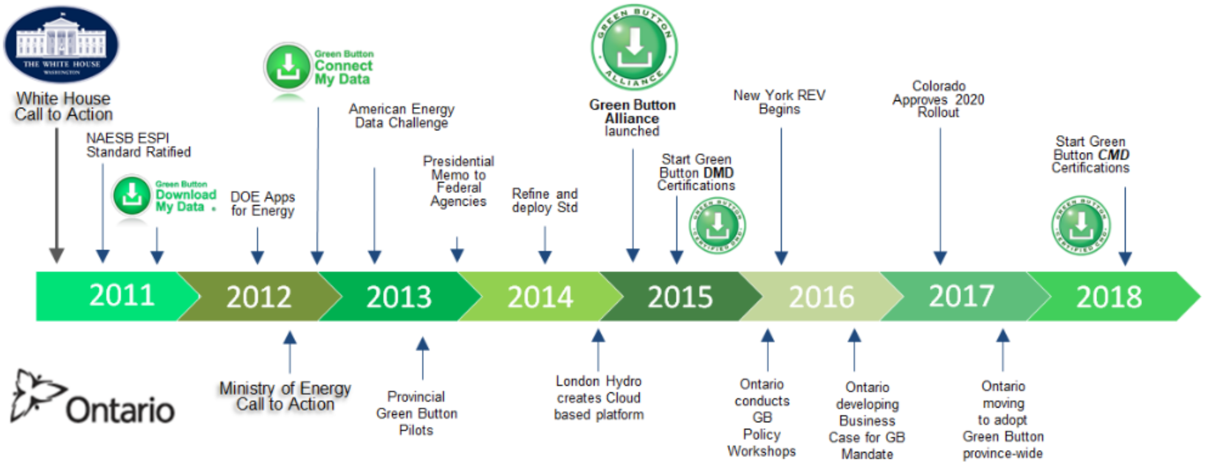

THE WHITE HOUSE
WASHINGTON
White House Call to Action
NAESB ESPI Standard Ratified
Green Button Download My Data
DOE Apps for Energy
Green Button Connect My Data
American Energy Data Challenge
Presidential Memo to Federal Agencies
Refine and deploy Std
Green Button Alliance launched
Start Green Button DMD Certifications
New York REV Begins
Colorado Approves 2020 Rollout
Start Green Button CMD Certifications
2011
2012
2013
2014
2015
2016
2017
2018
Ontario
Ministry of Energy Call to Action
Provincial Green Button Pilots
London Hydro creates Cloud based platform
Ontario conducts GB Policy Workshops
Ontario developing Business Case for GB Mandate
Ontario moving to adopt Green Button province-wide

Figure 4.1: Green Button timeline. Source: [56]

them expect self-service tools and apps that enable them to monitor and change their behaviour [34].

4.1.3 Example: Green Button Initiative

In September 2011, U.S. Chief Technology Officer, Aneesh Chopra challenged utilities to develop the Green Button. The goal was to provide customers with detailed information on their energy usage that is available for download in a simple, ready-to-process format. It was to enable customers at first to understand their usage patterns, and then to take decisions about their energy consumption. Standardisation was needed to assure that new applications may emerge, which may make use of this data [56].

Nowadays, the Green Button initiative provides customers with an access to the data on their electric energy usage. The data is provided in the textual file in a standardised format and may be shared with third parties offering applications for analysing this data and providing recommendations to users [33]. In addition, also business users benefit from this solution. Building owners and property managers retrieve utility-provided Green Button consumption data, and based on it build models, ensuring their buildings perform efficiently. It should be noted that Green Button is not only about electric energy, as water usage and natural-gas usage may be also found via the Green Button.

The Green Button initiative is not the only one, that based on data offered applications utilizing profiling methods to provide additional value to customers. Another example may be the Blue Button that targets the domain of health, and enables users to download health records and then analyse them and shape the healthcare for all family members [Buttons2018]. On the other hand, the Orange Button enables collection, security, management, exchange, and monetizing of solar datasets to target reduction of cost of solar installations [OrangeButton]. Based on those examples, new initiatives targeting new domains certainly will appear to suit the raising demand of customers.

4.1.4 Structure of the Chapter

New profiling initiatives are only a matter of time because of raising awareness of customers and the need for mitigation of risks faced by public utilities that may be observed [55]. Customers are willing to know what are their usage patterns, what is the main cost for them and how to decrease this cost. Therefore, they need access to their usage data. Once they gain it, they start analyses and take actions regarding usage patterns and employment of new solutions such as e.g. solar panels or LED lighting.

Nowadays, we also have a lot of discussions on environmental benefits that come from saving on services offered by public utilities (water, gas, electricity). On the electric energy market, shaving the peaks of usage, could contribute to decreasing usage costs and positively influencing the environment [55]. In our work, we targeted these issues from the perspective of the smart grid, but all challenges still remain valid.

In the context of profiling, being the main subject of this thesis, it should be underlined that profiling and the structure of the profile, greatly depend on the application scenario addressed. The previous chapters presented the potential of the profiling, and diverse methods that may be applied without focusing on a specific domain. This chapter covers details on the efforts related to profiling for the smart grid. Its main goal is **to create a profile of a user or a thing that will be applicable for solutions enabling management of production and consumption of the electric energy in the smart grid**. The secondary goals contributing to achieving the main goal defined are as follows:

- Proposing an architecture of a system for monitoring energy production and consumption in the smart grid, taking into account a profile of an individual prosumer.
- Creating a profile of a user for the needs of electric energy supply: monitoring and describing demand for the electric energy to be used by the system enabling management of the production and consumption of the electric energy in the smart grid.

The chapter consists of four sections including introduction presenting relation to goals of the thesis and summary that presents results that were achieved in relation to these goals. Section 4.2 relates to the first of the presented secondary goals and section 4.3 refers to the second of the goals mentioned.

4.2 Towards Forecasting Demand and Production of Electric Energy in Smart Grids

The goal of the section is to provide requirements analysis and architecture of the system enabling management of the production and consumption of the electric energy in the smart grid. This goal is in line with the following secondary goal of the thesis: ”proposing an architecture of a system for monitoring energy production and consumption in the smart grid, taking into account a profile of an individual prosumer”.

The paper was accepted to the International Conference on Business Informatics Research, 23-25.09.2013, Warsaw, Poland. The detailed bibliographic reference of the paper is as follows: Filipowska, A., Mucha, M., Hofman, R., Hossa, T., Fabisz, K., 2013, Towards Forecasting Demand and Production of Electric Energy in Smart Grids, Lecture Notes in Business Information Processing, 158, pp. 298-314.

Towards forecasting demand and production of electric energy in smart grids

Agata Filipowska[1], Karol Fabisz[1], Tymoteusz Mateusz Hossa[1], Michał Mucha[1], Radosław Hofman[2]

[1] Faculty of Informatics and Electronic Economy, Poznan University of Economics, Poznan 61-875, Poland
[2] Future Energy Sp. z o.o., ul. 28 czerwca 1956 r. 123/20, 61-544 Poznań, Poland
WWW: http://www.kie.ue.poznan.pl, http://www.future-energy.com.pl

Summary. Recently, the electric energy market undergoes serious changes which impact its future structure. They include also an emergence of smart grid encompassing prosumers, being individual market participants that not only consume, but also produce the electric energy. This imposes a need for a new class of tools (and methods) that will support all market players.

The article presents a solution for managing the energy consumption and production in microgrids. We present challenges of managing such networks as well as functionalities of a system, that enables for e.g. preparation of forecasts, tracing the energy consumption or creation of recommendations for the microgrid prosumers, in order to deal with these challenges.

Key words: smartgrid, energy management systems, prosumer, energy production, energy consumption

1 Introduction

Within the last few years, one may notice a rapid changes within many fields of the energy sector. The shift from perceiving electricity as a public good to understanding it as a commodity, which in Europe formally began in Great Britain in 1989 with the revamped Electricity Act [1], continues to advance and spread to other countries [2].

On the other hand, an eco-trend may be observed. This concerns also establishing of international electric energy and greenhouse gas emission certificates' exchange platforms. A political push for environmentally friendly economy resulted in many large renewable energy generation projects, as well as distributed installation of micro-installations in many West and Central European countries, including Germany, Spain, the UK and Czech Republic. For instance in Germany the capacity of energy generation from solar photovoltaic plants has grown from 76 MWp in the year 2000 to over 32 000 MWp in 2012, with an average yearly growth of over 74% [3].

This caused the emergence of a new class of energy market participants, namely prosumers. Examples include a household fitted with PV panels or a farmhouse equipped with a wind turbine. Since they both produce and consume electric energy, prosumers need to trade energy with the grid in case of both surplus and shortage of the energy.

However, the most significant advancement of all is probably the trend towards the installation of smart metering infrastructures (including the EC activities in this area), which allow for a two-way communication with the meters and automated, remote gathering of real-time energy consumption and production data (what is also a technical challenge with regard to the amount of data).

The future of electric energy market comprises the distributed electricity generation, extensive usage of renewable energy sources, facilitating local energy exchange within quasi-independent microgrids, very efficient load balancing using advanced prediction and energy flow management systems, all set in a legal framework enabling real-time market-determined pricing or dynamic tariffs. Also more accurate balance (long-term effect of management) will reduce the demand for energy, because even today 30% of electricity is lost due to lack of energy balance [10].

Realizing the abovementioned vision, requires a system that may:

- communicate with the smart metering infrastructure, in order to gather data and send commands,
- analyse the data to produce meaningful insights regarding market participants and ways in which they use electricity,
- predict supply (from renewable sources) and demand on the electric energy market,
- facilitate transactions and pricing,
- raise awareness about the energy consumption among users.

This publication examines the problem of designing such a system and presents a proprietary solution developed within the Future Energy Management System project[1].

The paper is structured as follows. Section 2 identifies challenges related to management of energy production and consumption in the smart grid. Section 3 presents the system developed to address these challenges. Section 4 provides description of the related work, showing the insights into the field of the smart grid management systems. The paper is summarised in the conclusions section, that presents the main points of the article.

[1] The research results described within the article were achieved within the project: Building a prototype of an innovative system for forecasting the level of consumption and production of electric energy "Future Energy Management System", (Project: UDA-POIG.01.04.00-30-065/10-03, Value: 9.957.022 zł, UE contribution: 5.647.005 zł, period 01.07.2011-31.03.2013), European Funds – for the development of innovative economy (Fundusze Europejskie - dla rozwoju innowacyjnej gospodarki).

2 Management of microgrids: challenges

2.1 Definition of the microgrid

The microgrid is a set of related energy sources, energy consumers and, in some cases, energy storage devices [11]. Each microgrid is connected to the main energy grid, but it should be self-sufficient and work without taking the energy from main electricity grid. The typical microgrid consist of the following elements [12]:

- sources of distributed generation (from renewable sources),
- power inverters,
- control systems,
- connections with the main grid,
- energy storage devices,
- energy consumer devices.

The typical energy sources of distributed generation encompass: biomass energy, geothermal energy, hydropower, ocean energy, wind power and solar photovoltaics [13]. The energy generated from these sources, for example from photovoltaics, must be converted from DC to AC, by using the power inverters. To manage the production and consumption of energy, the microgrid requires an advanced control system to optimize dispatch of energy and to provide load balancing in the microgrid network [14].

The microgrid according to its definition is self-sufficient, but usually it's connected to the main electrical grid, in order to assure a better load balancing in the macrogrid [20]. In microgrids, the energy storage devices are rarely used, because of low efficiency of energy storage technologies. The batteries are used in bungling microgrids, but with an increase in popularity of electric cars, it may end up with using these cars as battery storage devices [15].

The microgrid using the novel smart metering infrastructure is being referred to as smart grid. The rest of the paper, uses these notions as synonims.

Another important concept for the domain of distributed power generation is a prosumer. The prosumers are "professional consumers" or "producer consumers" of electric energy, being in fact also more aware consumers. Referring to the Act on Renewable Energy Sources, a prosumer is a "producer of electricity in micro-installations that is used for his own consumption or for sale (...)" [16].

2.2 Challenges for microgrids

The microgrids bring a lot of new challenges not known or not so much problematic in the large, typical power grids. Balancing supply and demand in microgrids is much more sophisticated, than in the large grids, mainly because of the law of large numbers, that implies that an average demand of a large number of consumers should be close to expected value [24]. In microgrids, the number of consumers is not always sufficient to expect that the law of large numbers works and therefore, the demand or production forecast may be significantly biased [21].

The load balancing is also more difficult in microgrids in comparison with the typical energy grid. The generation of electricity from for e.g. photovoltaics panels and wind turbines is less predictable [17], than e.g. from the fossil fuel. The efficiency of the renewable energy sources often depends on weather conditions, for example, solar and wind power. The systems for management of smartgrids are therefore forced to make predictions on the basis of weather forecasts, what may additionally increase the forecasts' error [18].

Another challenge, concerns explotation of many advanced technologies to manage and control the network. This factor together with the underdeveloped energy infrastructure often requires a large amount of money from investors [23].

The emergence of microgrids also creates a different structure of the energy network - highly decentralized. The concerns arise, if such a market will be able to act not in the laboratory conditions, but in the real world. The consumers are concerned with the price shocks and power blackouts, experts discuss the minimum legal regulatory level [22]. Another often overlooked challenge is the security of the microgrid (also with regard to personal data from the smart meters). The microgrids should be protected against the data leakage and cypher attacks, which can stop production or even destroy the power stations [19].

2.3 Requirements for the system for the microgrid management

The system addressing challenges identified in Section 2.2, that will be able to control and manage the microgrid, should be developed taking into account the following functional requirements:

- generate production and demand forecasts, with the lowest possible error,
- acquire data to improve the forecast models e.g. weather factors data, detailed prosumer data,
- enable management of prosumers, e.g. offerings, recommendations, the overview of energy production/usage,
- support the technical performance management and manage the flow of energy within the microgrid,
- support energy exchange of the microgrid with the macrogrid and manage the energy flow between these grids,
- support communication standards between devices in the microgrid, and between micro and macrogrid, what especially concerns acquisition of data from smart meters,
- support the security standards,
- enable for compliance with the legal requirements (national and European).

The most important feature of the system, from the economical point of view, is generation of low-error forecasts of energy demand, because the bigger the error, the greater the costs associated with the purchase of energy on the market (in the short term). Also, a accurate forecast of energy production in the microgrid is quite important. The knowledge of how much energy we are able to produce and comparing it to the forecasted demand, shows us how much energy we need to buy from the macrogrid, or how much energy we are able to sell [16].

The microgrid management system should also include a module to manage prosumers, in order that collected energy consumption data and detailed data on prosumers, are used also for the advantage of the prosumer. For example a prosumers should be able to list their electrical devices and their usage patterns. Also, the prosumer should be able to describe the specifics of his home, for e.g. type of home insulation (heat losses) or type of home lighting. The main reason for collecting this type of data is generation of forecasts of good quality (even for an individual). Lastly prosumers should provide data about their geographical location, which will allow to match the weather data. Thats why besides of collecting the data on prosumers, it is important to gather data on weather factors which may be useful in generating the energy production predictions.

Another requirement is to provide such a system with specialized mechanisms, which will manage the energy flow, changes in the energy production level like power or voltage, communication with macrogrid in case of excess or deficiency of microgrid production, etc. This kind of system should implement the communication standards between all devices in the microgrid, and between the micro and the macrogrid.

The most discussed issue, is the support of security standards. The microgrid should be protected from outside or internal attacks, which can be devastating to the network stability. In addition to the physical damage of devices, the cyber attacks may aim at the acquisition of users' private data, or data about the operation of the network itself. The security it also essential to meet the legal requirements of setting up a microgrid.

2.4 Scientific and economical challenges

The creation of a microgrid management system brings also a number of scientific and economical challenges.

As it was mentioned in Section 2.3, the forecast error is related to costs in monetary terms. If we overestimate the amount of energy needed by the microgrid, we will need to sell the extra energy to the main grid on low prices. On the other hand, if we underestimate the amount of energy, we will need to buy the difference on the high prices (in short term the energy is more expensive than in the longer horizon). The more accurate forecast, the lower costs of acquiring the electric energy (if needed) or higher income from the production.

Forecasting the consumption for small amount of prosumers is also a challenge, mainly because of the large diversity of the power consumption profiles (and the level of uncertainity). For example, for a group of 4-5 prosumers, their consumption is nearly random. Fortunately, with the rising number of prosumers in the microgrid, the forecast error decreases according to the law of large numbers. For example, for a group of 10 prosumers, the forecast error will be on an acceptable level and for a group of 100 prosumers some methods like a linear regression will give an average forecast error at a level of 2-3%. The chance for improvement of forecasting accuracy for small microgrids, is related to obtaining additional information e.g. on habits or appliances the prosumer possess.

One of the biggest problems in the management of microgrids is that the system needs to react to changes in the environment in the real time. This also concerns the accuracy of forecasts that are essential for planning, so that changes made during the system's functioning are as small as possible. Of course, one has to take into account also many periodic fluctuations like day fluctuations, week fluctuations, month fluctuations and seasons fluctuations. That is why, a microgrid management system, should collect additional data about the prosumer from external sources and use this data to provide more accurate forecast.

3 The concept of a microgrid energy management system

3.1 Market and legal background

The market potential: from the economical point of view, the potential of microgrids is worth billions of dollars [26]. This estimation takes into account a vast number of interconnected electricity users (households, business clients, etc.), their electricity consumption and value of physical infrastructure (transmission upgrades, automation of substations and distribution, smart grid IT and smart meters) [32]. The annual investment in all kind of RES solutions will rise about four times between now and 2030 [36], from 33$ billion annually in 2012 to 73$ billion by the end of 2020, with 494$ billion in cumulative revenue over that period [35]. Therefore, there will appear a need for designing a new kind of systems. It is expected, that in the next five years, energy operators will expand their investments in home and load management systems and storage technologies [33, 25].

In Europe, the microgrid technologies' market is yet underdeveloped, but it is growing rapidly [34]. There are many successful deployments of smart meters and AMI infrastructures in the Western Europe countries (UK, Germany, Spain), but now, the markets of the biggest potential are the ones located in the Central and Eastern Europe (Poland, Slovenia, Slovakia, Czech Republic, Bulgaria) [27]. The value of investment in smart meters in the aforementioned region will reach about 10.3$ billion by 2023 [29]. In other emerging markets from Eurasia, Latin America, Middle East/North Africa, South Africa and South-east Asia, smart meter and AMI infrastructures' market will be worth circa 56$ billion by 2022 [31]. Moreover, the North America Utilities are about to spend over 570$ million by 2016 on home energy management solutions (pilot programs and test deployments) [33]. Even the Brazil, is expected to spend over 27$ billion on total smart grid investments by 2022 [28].

European regulations: from the legal point of view, all European countries must adopt the EU Renewable Energy Directive, known as "The EU climate and energy 20–20–20 package", until 2020. It is an ambitious plan, that represents an integrated approach to climate and energy policy: reducing green–house gas emissions, achieving sustainable development and ensuring the energy security. In the context of smart grids, the legislative package includes directives on the promotion of the use of energy from renewable sources. Moreover, the member states have established national targets for raising the share of RES in their energy consumption (e.g. 10% in Malta, 50% in Sweden) to reach an overall 20% renewables goal in EU [37, 38, 39].

According to the latest multi–annual EU financial framework for the period 2014–2020, the EC decided to propose the creation of a new instrument called "Connecting Europe Facility" (€50 billion total). The main target is to conduct EU investments in transport (€31.7 billion), energy (€9.1 billion) and telecommunications (€9.2 billion) [40]. Moreover, additional €40 billion will be allocated for large–scale deployments of smart grid technologies across the European Union. These investments in key infrastructures can strengthen the Europe's competitiveness, create jobs (forecast of 400 000 places) and promote green energy solutions [41]. Recent studies show that the 20–20–20 policy is mostly affecting and promoting IT/Technical Operations and Engineering/Product Development [42].

Smart Grid market in Poland: as for now, there is no legal framework supporting emergence and functioning of smart grids in Poland. However, the legislation process is in progress and the appriopriate law is expected to be ready in 2013 [16]. Moreover, to encourage RES micro–producers to become active participants of the energy market, financial support schemes for RES, based on green certificates, feed–in tariffs, subsidies, financial instruments and auction mechanism, are being put in place [43, 44]. Moreover, many pilot studies concerning intelligent infrastructure are being carried out. The biggest one is conducted by Energa Operator. The company has already successfully deployed 50 000 smart meters in Kalisz and currently is carrying out research on smart grid system in the Hel Peninsula [45, 48].

To provide the quality collaboration between the information and operational technology, within the context of smart energy solutions, new IT systems (e.g. microgrid energy market management) must be introduced (the market is estimated for 8.6$ billion by 2017) [32]. The demand for smart grid products and systems is estimated for about 2000$ billion over the next 20 years. Thus, there appears a great opportunity to create and use solutions, such as the one described within the paper.

3.2 The microgrid management system concept

In comparison to the traditional energy market, there appear few new entities in the microgrid, most of all prosumers and aggregators. All of the market participants should be equipped with tools that enable to carry out the everyday activities, including decision support, local energy trading and the microgrid management (users, devices, loads, efficiency) [47, 46]. Moreover, these entities should cooperate with one another or even be provided with a single platform. To address the emerging challenges being faced by the microgrid market participants, we propose the concept of Future Energy Management System (FEMS), that is depicted in the Fig. 1.

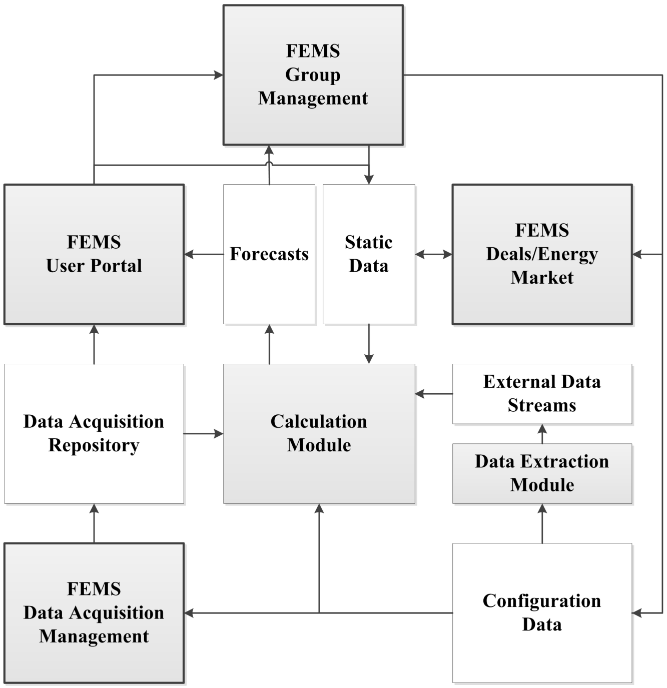


Fig. 1. Future Energy Management System concept

The FEMS system is mainly developed for:

- household administrators of office buildings, apartments, settlements, factories, etc.
- groups of stakeholders e.g. local communities consisting of energy consumers and producers,
- microgeneration administrators and managers.

The architecture of the system is presented in Fig. 1 and encompasses:

- **FEMS User Portal** – being the main user (for all groups of users) tool for managing the energy consumption and production, viewing statistics and tracking personalized recommendations, offering building and electrical energy trading.
- **FEMS Group Management** – the tool prepared for the group administrator that ensures a simple way of managing and analysing energy consumption and production within the whole microgrid subject to management. For instance, the administrator can prepare the energy offers and then send them to the users of the microgrid. Moreover, the tool supports recommendations and communication with users. Finally, it is also responsible for managing permissions and access to the FEMS system and its configuration.

- **FEMS Deals/Energy Market** – component dedicated to energy trading entities, allows the user to answer energy buying/selling inquiries sent by the group administrator.
- **FEMS Data Acquisition Management** – a central tool responsible for the integration with metering infrastructure via two-way communication protocols and data acquisition models. It also handles data measurement, control and reporting.
- **Data Extraction Module** – dedicated to collect data from the Web sources, that might be used for preparation of forecasts and recommendations.
- **Calculation Module** – enabling to prepare recommendations, users' classifications and energy predictions.
- **Static Data** – storage of data describing customers, power delivery points, contracts and devices.
- **Configuration Data** – dictionaries and stereotype definitions.
- **External Data Streams** – encompassing all kinds of data, taken from unstructured resources like weather or events calendar, that are important to energy forecasting methods.
- **Forecasts** – storage of short, medium and long term data describing predicted consumption and production of electric energy.
- **Data Acquisition Repository** – responsible for data retrieval and storage.

We believe, that the described architecture is able to provide all functionalities presented in the Section 3.3.

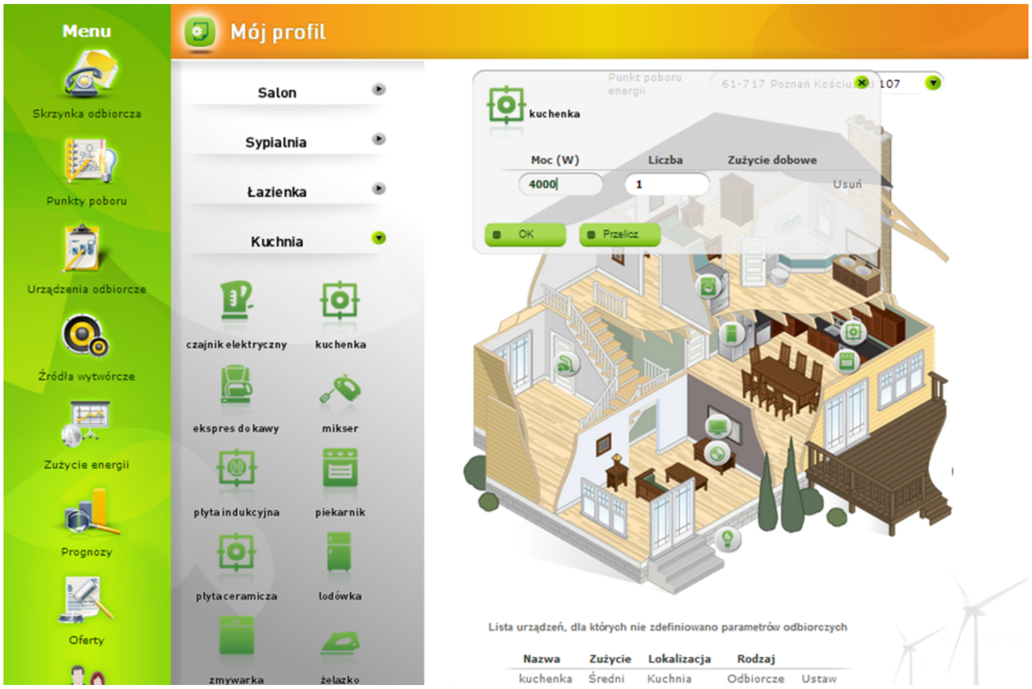
3.3 FEMS Functionalities

The main functionalities of FEMS encompass inter alia acquisition of data from smart meters, reasoning over the grid model, providing recommendations, forecasting and methods for retrieving information from external sources. Moreover, FEMS aims at developing a service that will strengthen the user engagement in the field of smart consumption within microgrid whilst providing listed below functionalities and intuitive, user-friendly interfaces as the one depicted in Fig. 2.

According to analyses that were carried out, these functionalities include:

1. **Monitoring and comparison of defined key performance indicators of energy consumption** – this functionality implies historic, current and future KPIs tracking and comparing them within the microgrid with regard to location, size, number of household members, etc. The user may therefore not only see his KPIs, but also of users of similar energy profile.
2. **Preparing personalized recommendations of actions towards minimising/maximising energy usage/production** – includes presenting hints, based on the prosumer behaviour, possessed devices, calendar, historical consumption and family model (e.g. "Please check the fridge for repair as it consumes 30 percent more energy than last month."). It also presents recommendations dedicated to microproducers, based mostly on weather conditions and device parameters.

3. **Prosumers profiling** – Includes prosumers' profiling via clustering with regard to static (e.g. user type, household devices, stereotype) and dynamic data (historical consumption). These profiles will be further generalised and used by the forecasting and recommendation methods.
4. **Forecasting energy consumption and production** – Includes preparation of short, medium and long-term consumption forecasts based on data from smart meters and production estimation based on system–external data and unstructured data stream retrieved from the Web. Finally, the outputs will be automatically analysed via recommendation mechanism and personalized hints will be presented to users.



Punkt poboru energii: 61-717 Poznań Kościuszki 107

kuchenka

Moc (W)	Liczba	Zużycie dobowe
4000	1	

Lista urządzeń, dla których nie zdefiniowano parametrów odbiorczych

Nazwa	Zużycie	Lokalizacja	Rodzaj	
kuchenka	Średni	Kuchnia	Odbiorcze	Ustaw

Fig. 2. FEMS interface for defining household devices. Source: FEMS system

The user (an individual household) in order to benefit from the solution, is obliged to define:

– personal data (indicating the smart meter possessed),
– consumption and production devices and their parameters,
– energy consumption and production schedule,
– description of the building (its features),
– user preferences,
– family model (stereotype).

The abovementioned parameters can be defined as is depicted in Fig. 2. It presents the interface enabling definition of user stereotype by adding devices to the household with a built-in interactive tool. The users can choose location and then simply drag & drop devices, set their parameters such as the nominal power, their number and energy class, etc. This interface allows also to count the nominal power of all installed devices that might be used for the household energy load forecast or system recommendation methods. By filling in all presented parameters, the user has an opportunity to save money, improve the awareness and actively participate in the local microgrid.

The system implementing the described functionalities following the architecture described above was developed and tested within the Future Energy Management System project financed within the frameword of the Polish Innovative Economy Operational Programme (Zbudowanie prototypu innowacyjnego systemu prognozowania poziomu zużycia i produkcji energii elektrycznej o nazwie 'Future Energy Management System' project: UDA-POIG.01.04.00-30-065/10-00, project value: 9.957.022 PLN, share of the European Union: 5.647.245 PLN, therm of the realization: 07.2011-03.2013).

4 Related work

With the smart metering infrastructure began the dawn of informatization of the energy sector. Thus, IT companies rushed in, trying to deliver a right product for different participants of the energy market. Especially those with experience in development of industrial control systems and analyses of big data.

In Europe, a large part of the development effort is conducted by academics, who are carrying out research projects with the purpose of fulfilling the vision of an energy efficient Europe, as set out by the European Union. Some of these projects, are implemented by start-ups, created by independent entrepreneurs with the goal of satisfying the new market. Among the software-oriented approaches at utilizing the smart metering infrastructure to provide the maximum benefit, we can distinguish those which focus on energy savings on the consumer side (supporting the customer), and those which aim at creation of tools for utilities that allow to run their businesses more efficiently. The latter is addressed mainly by large corporations, which offer solutions for big market players [4, 6, 7, 8]. These corporations take the former approach usually only for the marketing purposes. Only small companies rely on it for profit, seeing the market potential [5].

The large scale efforts of big companies are focused on helping utilities control and maintain the state of the grid, address faults and problems more efficiently, segment their market better in order to offer more value to their customers, and lastly, make it easier to fulfil regulatory obligations. There is also a group of innovative projects for utilities created by smaller companies or academia, as well as projects catering to new participants of the market, based on innovative business models. An example can be creating virtual power plants by establishing clusters of small energy producers [9].

Examples of innovative functionalities offered by ongoing projects are:

- interactive energy use calculators based on manual data input and user-relevant recommendations regarding energy usage – Wattzon[2],
- comparing energy efficiency with the general population and one's neighbors to influence the change of behavior among users by friendly competition – Wattzon, Opower[3],
- communicating with metering infrastructure; support for energy purchasing, contracts, holding accounts; integration of energy management solutions with SCADA systems; integration of customer management with meter data for the purpose of billing – Proximus-IT[4], STC Energy[5], INNSOFT[6], Web2energy[7], ECIX-ORL[8], Opower,
- simulation of energy flow and prices for retail consumers based on statistic methods and stochastic process modelling – resLoadSim[9],
- integration of data visualization and statistics' modules with metering infrastructure – Energy Cap[10], STC Energy.

Other exemplary initiatives in the field, not fully related to functionalities offered by FEMS, are:

- platforms for sharing best practices in executing smart grid initiatives and educating retail consumers to manage their households' energy usage – MeterON[11], EEGI[12], ADDRESS[13], Wattzon,
- efforts to reduce energy needs of buildings by replacing equipment and educating consumers – BEEMUP[14],
- increasing reliance on renewable energy sources and raising awareness about energy usage among consumers – ADDRESS,
- establishing industry standards for smart metering infrastructure, taking into account communication channels and data transfer protocols – OPEN METER[15].

[2] www.wattzon.com
[3] www.opower.com
[4] www.proximus-it.pl/b/zarzadzanie-zakupem-energii/0
[5] www.stcenergy.com
[6] www.innsoft.pl
[7] www.web2energy.com
[8] www.eurocim.pl/ecixorl.html
[9] http://ses.jrc.ec.europa.eu/our-models-portfolio
[10] www.nationalenergyconsulting.com/energycap.html
[11] www.meter-on.eu
[12] www.smartgrids.eu/documents/EEGI/EEGI_Implementation_plan_May\%202010.pdf
[13] www.addressfp7.org
[14] www.beem-up.eu
[15] www.openmeter.com

5 Conclusions

The microgrid market and technologies that enable to manage microgrids are now highly developed. Despite the challenges posed by the management of small electricity grids, there appear new ideas, advanced management techniques, more accurate forecasting models and better tools to predict the market changes. Fortunately, with the development of the microgrid, the accuracy of forecasting the behavior of the microgrid components is improving. The accuracy of forecasts in the microgrids is extremely important because it's strictly related its functioning costs, what is then further reflected in the profitability of microgrid system as a whole.

The market potential of microgrids, especially in the global context is estimated for billions of dollars of investment (but also of revenue). Also on an European market, namely in the CEE[16] region, a lot of investments are performed e.g. by 2033 smart meters investments will reach about 10.3$ billion in the CEE countries.

Currently developed microgrid management systems are mainly prepared to support the management of large smartgrid networks, which are connected by SCADA class systems. On the other hand, there are many small applications that display data on average energy consumption and recommend environmentally responsible behaviour. There are few systems, which support the small microgrids and these type of grids are currently being developed.

FEMS is a system that fits in this gap of the market, as it supports households and individual clients, and it can be successfully applied to small business entities. FEMS supports the analysis of energy consumption and enables to forecast future consumption, as well as production of the microgeneration. The forecasts are generated based on the historical data, but also data acquired from external sources (i.e. Internet weather forcase, TV programme etc). FEMS also supports the microgrid market transactions to allow the energy trading on the local market, using the neighbourhood of renewable energy resources. One of the important modules is also a recommendation system, that operates on data on prosumers in order to effectively educate them in caring for the environment and saving the energy.

FEMS can be easily extended with additional modules and adapted to a variety of markets in the CEE region, as well as around the world. Using this type of management systems for the microgrid may offer many advantages such as: efficient exploitation of renewable energy sources, reduced emissions of carbon dioxide, protection against blackouts, reduction of cost of energy supply, automated acquisiton of data from smart meters and Internet sources, and many more. This type of systems will soon become present in our everyday life, so at this stage we should support their development, seeing that this kind of systems will provide us with clean and green energy in the cheapest way possible.

[16] Central and Eastern Europe countries

References

1. Simmonds, G., "Regulation of the UK electricity industry: 2002", Centre for the Study of Regulated Industries, Bath, UK, 2002.
2. Alsunaidy, A., Green, R., "Electricity Deregulation in OECD (Organization for Economic Cooperation and Development) Countries", Energy 31, no. 6-7 (2006), pp. 769-787.
3. Bundesministerium fur Umwelt, Naturschutz und Reaktorsicherheit, "Development of renewable energy sources in Germany in 2012", Berlin, 2013. http://www.erneuerbare-energien.de/fileadmin/Daten_EE/Dokumente__PDFs_/20130328_hgp_e_ppt_2012_fin_bf.pdf (accessed May 05, 2013).
4. SAP AG, "Energy Data, Smart Meter Analytics Software, HANA In-Memory Computing, SAP.", SAP Business Management Software Solutions, Applications and Services. http://www54.sap.com/pc/tech/in-memory-computing-hana/software/smart-meter-analytics/index.html (accessed March 20, 2013).
5. WattzOn, Inc., www.wattzon.com (accessed March 21, 2013).
6. Siemens AG, Infrastructure & Cities Sector, Smart Grid Division, "Energy meets Intelligence", 2013. http://w3.siemens.com/smartgrid/global/en/products-systems-solutions/Documents/SIEIC_39L_SmartGrid_Update_14s_engl_210x280.pdf (accessed May 05, 2013).
7. Oracle Utilities, "Oracle Utilities Smart Grid Gateway", 2013. http://www.oracle.com/us/industries/utilities/utilities-smart-grid-ds-323531.pdf (accessed May 05, 2013).
8. IBM Corp., "IBM Smart Grid - Solutions - United States", IBM - United States. http://www.ibm.com/smarterplanet/us/en/smart_grid/nextsteps/index.html (accessed April 10, 2013).
9. Fraunhofer Institute, "The virtual power plant - stable supply of electricity from renewable energies", Press Release, 26 March 2013. http://www.fraunhofer.de/en/press/research-news/2013/march/the-virtual-power-plant.html (accessed April 12, 2013).
10. Srinivasan K., Rosenberg, C., "How internet concepts and technologies can help green and smarten the electrical grid", Proceedings of the first ACM SIGCOMM workshop on Green networking (Green Networking '10), ACM, New York, NY, USA, pp. 35-40. http://doi.acm.org/10.1145/1851290.1851298 (accessed May 05, 2013).
11. Kaplan, S. M., "Smart grid: modernizing electric power transmission and distribution; energy independence, storage and security; energy independence and security act of 2007 (EISA); improving electrical grid efficiency, communication, reliability, and resiliency; integra", Alexandria, VA: TheCapitol.Net, 2009.
12. Olszowiec, P., "Autonomiczne systemy elektroenergetyczne malej mocy. Mikrosieci", Energia Gigawat, nr 7-8, 2009.
13. Ren 21, "Renewables 2012 Global status report", Paris, 2012. http://www.map.ren21.net (accessed May 05, 2013).
14. Dimeas, A.L., Hatziargyriou, N.D., "Operation Of A Multiagent System For Microgrid Control", IEEE Transactions on Power Systems, vol. 20, no. 3 (2005), pp. 1447-1455.
15. Markel, T., Kuss, M., Simpson, M., "Value of plug-in vehicle grid support operation", Innovative Technologies for an Efficient and Reliable Electricity Supply (CITRES), pp. 325-332, 27-29 Sept. 2010. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5619785&isnumber=5619765 (accessed May 05, 2013).

16. Sejm RP, "Projekt ustawy o odnawialnych zrodlach energii z dnia 14 grudnia 2012r.". http://orka.sejm.gov.pl/Druki7ka.nsf/dok?OpenAgent&7-020-492-2012 (accessed May 05, 2013).
17. Rezaei, E., Afsharnia, S., "Cooperative voltage balancing in islanded microgrid with single-phase loads", 2011 International Conference on Electrical and Control Engineering (ICECE), pp. 5804-5808, 16-18 Sept. 2011. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6057188&isnumber=6056741 (accessed May 05, 2013).
18. Bando, S., Sasaki, Y., Asano, H., Tagami, S., "Balancing control method of a microgrid with intermittent renewable energy generators and small battery storage", Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE, pp. 1-6, 20-24 July 2008. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4596074&isnumber=4595968 (accessed May 05, 2013).
19. Cheung, H., Hamlyn, A., Mander, T., Cungang, Y., Cheung, R., "Strategy and Role-based Model of Security Access Control for Smart Grids Computer Networks", Electrical Power Conference, IEEE Canada, pp. 423-428, 25-26 Oct. 2007. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4520369&isnumber=4520285
20. Li, P., Li, X., Liu, J., Chen, J., Chen, J., "Analysis of acceptable capacity of microgrid connected to the main power grid", 4th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), pp. 1799-1802, 6-9 July 2011. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5994190&isnumber=5993852 (accessed May 05, 2013).
21. Saad, W., Han, Z., Poor, H.V., Basar, T., "Game-Theoretic Methods for the Smart Grid: An Overview of Microgrid Systems, Demand-Side Management, and Smart Grid Communications", presented at IEEE Signal Process. Mag., 2012, pp. 86-105.
22. Tao, L., Schwaegerl, C., Narayanan, S., Zhang, J.H., "From laboratory Microgrid to real markets - Challenges and opportunities," 8th International Conference on Power Electronics and ECCE Asia (ICPE & ECCE), 2011 IEEE, 30 May 30 - 3 June 2011. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5944600&isnumber=5944368 (accessed May 05, 2013).
23. Colson, C.M., Nehrir, M.H., "A review of challenges to real-time power management of microgrids," Power & Energy Society General Meeting, IEEE, pp. 1-8, 26-30 July 2009. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5275343&isnumber=5260217 (accessed May 05, 2013).
24. Bernoulli, J., "Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis", 1713.
25. GlobalData, "Press Releases — Smart Grids: Microgrid Market Boom on the Way as Europe and Asia Catch-up to US". http://energy.globaldata.com/pressreleasedetails.aspx?prid=705 (accessed April 18, 2013).
26. GlobalData, "Microgrid in Smart Grid - Market Size, Key Issues, Regulations and Outlook to 2020". http://www.researchandmarkets.com/reports/2518946/microgrid_in_smart_grid_market_size_key (accessed April 18, 2013).
27. Hayes, S., Young, R., Sciortino, M., "The ACEEE 2012 International Energy Efficiency Scorecard", American Council for an Energy-Efficient Economy. http://www.aceee.org/research-report/e12a (accessed April 15, 2013).
28. Northeast Group LLC, "Brazil Smart Grid: Market Forecast (2012-2022)", Washington, 2012. http://www.northeast-group.com/reports/Brazil_Smart_Grid_Market_Forecast_2012-2022_Brochure_Northeast_Group_LLC.pdf (accessed May 05, 2013).

29. Northeast Group LLC, ”Central and Eastern Europe Smart Grid: Market Forecast (2013-2023)”, Washington, 2013. http://www.researchandmarkets.com/reports/2556779/central_and_eastern_europe_smart_grid_market (accessed May 05, 2013).
30. Memoori Research, ”The Smart Grid Business 2012 to 2017”, London, 2012. http://www.memoori.com/portfolio/the-smart-grid-business-2012-to-2017/ (accessed May 05, 2013).
31. Northeast Group LLC, ”Emerging Markets Smart Grid: Outlook 2013”, Washington, 2013. http://www.northeast-group.com/reports/Emerging_Markets_Smart_Grid_Outlook_2013_Northeast_Group.pdf (accessed May 05, 2013).
32. Navigant Research, ”Smart Grid IT Systems”, Boulder, Colorado, US, 2012. http://www.navigantresearch.com/research/smart-grid-it-systems (accessed May 05, 2013).
33. IDC Energy Insights, ”Technology Selection: North America Home Energy Management Spending Forecast, 2011-2016”, Framingham, MA, US, 2012. http://www.idc-ei.com/getdoc.jsp?containerId=EI236935 (accessed May 05, 2013).
34. Navigant Research, ”Market Data: Microgrids”, Boulder, Colorado, US, 2013. http://www.navigantresearch.com/wp-assets/uploads/2013/03/MD-MICRO-13-Executive-Summary.pdf (accessed May 05, 2013).
35. Navigant Research, ”Smart Grid Technologies”, Boulder, Colorado, US, 2012. http://www.navigantresearch.com/research/smart-grid-technologies (accessed May 05, 2013).
36. Bloomberg New Energy Finance, ”Strong growth for renewables expected through to 2030, Bloomberg New Energy Finance”, Bloomberg New Energy Finance. http://about.bnef.com/press-releases/strong-growth-for-renewables-expected-through-to-2030/ (accessed May 1, 2013).
37. European Commision, ”20–20–20 package, Climate and energy policy of the EU”, 2012. http://ec.europa.eu/clima/policies/package/index_en.htm (accessed May 05, 2013).
38. Nalco.com, ”Learn More About Our Air Protection Technologies”. http://www.nalcomobotec.com/mb/eu-20-20-20-rule.htm (accessed June 18, 2013).
39. Europedia, ”The EU ”energy-climate” package”. http://europedia.moussis.eu/discus/discus-1230747802-321327-28435.tkl (accessed June 18, 2013).
40. European Union, ”A budget for Europe (2014-2020)”, Europa.eu, Summaries of EU legislation. http://europa.eu/legislation_summaries/budget/bu0001_en.htm (accessed June 18, 2013).
41. European Commision, ”Connecting Europe Facility, 2012”. http://ec.europa.eu/energy/mff/facility/doc/2012/connecting-europe.pdf (accessed May 05, 2013).
42. Emerson Network Power, ”Business-Critical Continuity EU 20–20–20 Directive”, Columbus, OH, US, 2013. http://www.emersonnetworkpower.com/en-EMEA/About/NewsRoom/Documents/EU202020-Report.pdf (accessed May 05, 2013).
43. Green Energy Poland SA, ”5. Zielone Certyfikaty mechanizm dzialania, cena, zasady obrotu. Green Energy Poland SA., Inwestycje z nowa energia”. http://gepsa.pl/322-2/ (accessed June 18, 2013).
44. Kasnowski, J., ”System taryfy gwarantowanej dla fotowoltaiki”, Euroinfrastruktura.pl. http://www.euroinfrastructure.eu/finanse-i-prawo/system-taryfy-gwarantowanej-dla-fotowoltaiki/ (accessed June 18, 2013).

45. Smart-Grids.pl, "Kalisz pierwszym miastem ze Smart Meteringiem. Smart-Grids.pl". http://www.smart-grids.pl/technologie/749-kalisz-pierwszym-miastem-ze-smart-meteringiem.html (accessed June 18, 2013).
46. Dzikowski, J., Filipowska, A., "Wykorzystanie danych z Internetu w prognozowaniu zachowan prosumentow w mikrosieciach energetycznych", Poznan 2012.
47. Dzikowski, J., Filipowska, A., "Przyczynowo-skutkowy model zmiennosci rynku energii elektrycznej", Zeszyty Naukowe Uniwersytetu Ekonomicznego w Poznaniu, Poznan 2012.
48. ENERGA Operator S.A., "ENERGA i Instytut Energetyki: Inteligentna siec coraz blizej". http://www.energa-operator.pl/klienci_indywidualni/informacje.xml?id=3357 (accessed June 18, 2013).

4.3 Profiling of Prosumers for the Needs of Electric Energy Demand Estimation in Microgrids

The goal of the second paper within this chapter is to present and validate a method for estimation of the demand for the electric energy in microgrids based on profiling of a prosumer that enables to determine the energy demand. This paper contributes to achieving the following secondary goal of the thesis: ”Creating a profile of a user for the needs of electric energy supply: monitoring and describing demand for the electric energy to be used by the system enabling management of the production and consumption of the electric energy in the smart grid”. It is worth to underline that the paper targets also the issue of profiling of things.

The paper was published in the International Journal of Energy Optimization and Engineering (IJEOE). Detailed reference: Fabisz, K., Filipowska, A., Hossa, T., 2015, Profiling of Prosumers for the Needs of Electric Energy Demand Estimation in Microgrids, International Journal of Energy Optimization and Engineering (IJEOE), 4, pp. 29-45.

Profiling of prosumers for the needs of electric energy demand estimation in microgrids

Abstract— **Nowadays, a lot of attention regarding smart grids' development is devoted to delivery of methods for estimation of the energy demand taking into account the behavior of network participants (being single prosumers or groups of prosumers). These methods take an advantage from an analysis of the ex-post data on energy consumption, usually with no additional data about profiles of prosumers.**

The goal of this paper is to present and validate a method for an energy demand forecasting based on profiling of prosumers that enables estimation of the energy demand for every user stereotype, every hour, every day of the year and even for every device. The paper presents possible scenarios on how the proposed approach can be used for the benefit of the microgrid.

Keywords-microgrid, prosumer, profiling, prosumer's behavior, energy demand estimation

Introduction

The energy market is changing. New entities, such as prosumers and network aggregators appeared and new management, trading, decision support and forecasting tools are being developed. As a result, challenges and questions arise, such as i.e. "How to encourage the prosumers to data sharing?", "How to make use of the prosumer involvement?", "How to design and make an advantage of the data provided by the prosumers?" and "How to determine a role of a prosumer?" (Shandurkova et al., 2013; Brendal, 2013; Filipowska et al., 2013).

Currently, energy operators prepare forecasts for large groups of energy consumers based on predefined standard energy profiles or another calculation methods like time-series regression, neuro-fuzzy techniques like ARIMA models, etc. (Ghanbarian, Kavehnia, Askari, Mohammadi, & Keivani, 2007; Churueang & Damrongkulkamjorn, 2005). These methods rely on data on historical energy consumption. In many cases, such solutions work well and give low error forecasts results, because of the statistical law of the large numbers (the larger number of energy consumers, the more aligned the forecast).

However, an increasing number of prosumers is equipped with smart meters and the advanced metering infrastructure (AMI) is expanding (Livgard, 2010). In addition, an exploitation of renewable energy sources like wind or sun (photovoltaic panels) becomes more popular, because of the national energy policies and ecofriendly social trends (Parkinson, Wang & Djilali, 2012; Singh, Alapatt & Poole, 2012). Thus, the emergence of the smart grids solutions and intelligent grids, which have their own generation sources and can work almost independently from the main power grid, is being observed (Taft, 2012). Therefore, there is a need for preparing energy consumption forecast of good quality for small groups of prosumers (sometimes even for individuals), independently from the other groups as such prosumer groups can use the energy from their own generation sources. A single consumer or a small group volatility with regard to energy demand is very hard to predict, so energy analysts have a difficult task trying to deliver forecasts with a low error.

Our research (these results are a subject of other submission) shows, that standard forecasting methods like non-linear regression, seasonal exponential smoothing or local methods need aggregated data from at least 60 prosumers to produce well fitted forecasts. When, we use aggregated data from over 60 prosumers, the mean squared prediction error stabilizes. This means that forecasting for groups smaller than 60 prosumers using standard forecasting methods won't provide satisfying results (Hossa et al., 2014).

Therefore, there is a need to supplement standard forecasting methods with other approaches. Besides of gathering the historical energy consumption, the data describing each energy prosumer should be possessed. Such datasets will give an opportunity to explain the volatility of energy consumption even for a single prosumer.

The aim of this paper is to present and validate a method for the energy demand estimation in microgrids based on profiling of prosumers (energy producers and consumers) that enables to determine the energy demand for every user stereotype, every hour of the year and for every device. The proposed approach is tailored to the needs of an innovative microgrid management and decision-support system, namely the Future Energy Management System (FEMS).

The paper is structured as follows. Section 2 introduces the notion of the profile, showing the potential exploitation of user data with regard to energy demand estimation. Section 3 describes the method that enables estimation of energy demand based on the prosumers' data. Section 4 provides a use case-based validation of the proposed method. It includes also possible scenarios on how the suggested approach can be used within a microgrid. Section 5 presents the related work. The paper is summarized in the conclusions section that presents the key points as well as possible directions of our future work.

Prosumer profiling

The notion of a profile in the domain of Information Systems, refers currently to the Web 2.0 tools that enable users for creation of their identities. Such profiles usually include information that characterizes a user including personal, contact as well as additional data such as hobbies, interests, etc. Examples of tools include Facebook (Lampe, 2007), Instagram or corporate portals.

Another type of user profiles is used by recommendation engines, such as Last.fm (http://www.last.fm). This service builds profiles of users regarding the music they listen to. Last.fm is on one hand a social network, where users may add or look for friends listening to similar music, on the other is a complex tool suggesting the user the music he should listen to or displaying him personalized ads. There exist also simple user profiles on the Web that consist of the contact data, transaction history, preferences and other data important and stored from an e-business point of view (Adomavicius, 2005). In this case, the aim of profiling is gathering knowledge that enables preparation of a personalized offer. The research results show that such personalization builds the customer satisfaction and loyalty (Ntawanga, 2008).

This leads us to a definition of a profile that is a source of knowledge on a user on all aspects that may be important from the systems functionalities' point of view (Wahlster, 1989). (Koch, 2000) defined a profile as an abstract representation of a real-world user, in the sense that not all user characteristics are reflected in the model. Only the characteristics that are useful should be taken into account by the system. Similarly the profile is understood in our research. We describe all characteristics of a user that may influence his energy demand. This data may be diverse and include inter alia:

- demographic information on a user e.g. age, gender, nationality,
- family data i.e. the number of people living in the household, including their age and status,
- type of the real estate the users live in e.g. type of heating, cooker, isolation, width of walls,
- style of work, hobbies, etc. – all information characterizing time of the day the user spends at home (using different appliances),
- devices that a user exploits including time of the day, when they are used.

As it may be easily noticed, the data that needs to be gathered by the system is to enable for modeling the user behavior with regard to the energy usage. The demographic and family information, may be used while filling in the gaps in the data e.g. by assigning a user to a correct group (similar from the point of view of his behavior). Moreover, such data enables for delivering of new type of methods for energy demand forecasting taking into account not only the historical demand, but also information on the user.

Towards estimation of microgrids' energy consumption

This section aims at presenting the underlying assumptions as well as the method enabling for energy demand estimation without using the historical data. The proposed method takes into account behavioral and identification variables.

Prosumer stereotypes in microgrids

To better understand the microgrid environment, we distinguish three main categories and some subcategories of the energy prosumers that co-exist within an intelligent network. Based on the energy consumption patterns, we differentiate between the following stereotypes (where a stereotype is a set of data that determines a family type, a building type, a number of persons in a household, appliances used (nominal power, number of devices, an energy class), energy properties of a building, prosumer behavior habits, etc. (Filipowska et al., 2013)):

- households – some of the devices used are constantly plugged in (e.g. a fridge, a TV set), few are used occasionally (e.g. a vacuum cleaner), number of persons varies from one person (single) to many (e.g. 2+2, 2+4 family). This type includes:
    - flats – energy used mainly for lighting, cooking, entertaining,
    - detached houses, energy used as above and in many cases also for heating and cooling,
- business entities:
    - restaurants – depending on the day and the "rush hours", they are characterized by high energy demand in a kitchen (fridges, coolers, electric kettles, etc.) and in a dining room (lighting, air conditioning, etc.),
    - stores – depending on the type of their activity type (e.g. grocery, pharmacy, supermarket) they might use inter alia devices constantly consuming energy such as coolers or fridges as well as those used only during the fixed working hours i.e. 09.00-17.00 or 09.00-20.00,
    - factories – depending on the type of industry (e.g. food, construction, production), they are described by the use of an industrial machinery that has a large energy demand,

    - offices – the highest energy demand occurs during the office working hours, which are e.g. 07.00-15.00 or 09.00-17.00 and the typical devices used are computers, printers, servers, TVs, etc.,
    - wellness, sport & recreation centers – the typically used devices are lights, audio, TVs, special equipment, etc., with working hours e.g. from 09.00 to 22.00,
- institutions:
    - schools & universities – depending on the size and scale, possessing many electric devices, especially computer and IT appliances, but also lightning, cooking and cooling devices,
    - healthcare centers – such as hospitals and health centers, where professional medical equipment and appliances such as medical imaging machines, medical monitors, laboratory equipment, computers, lighting, cooling, etc. are used,
    - churches – that consume a lot of energy for lightning and heating (if it is installed),
    - public utilities & offices – considered as public businesses that provide the public with everyday necessities, entail such features as large amount of electronic devices (computers and other office appliances) and fixed working hours e.g. in Poland from 07.00 to 15.00.

These stereotypes were distinguished taking into account typical classifications of buildings and their usage types (MoE, 1994; CSO, 2013a). Each of them is characterized by a set of different appliances, varying hours of energy consumption, preferences, building type, etc. Additionally, each entity has a diverse overall energy demand that may also vary with the time of the year i.e. within a month a prosumer classified as a flat (i.e. four- person household) may consume circa 200 kWh energy and a prosumer defined as a restaurant may consume ten times more energy (CSO, 2013b) .

Assumptions of the proposed approach

Throughout the past few years, many energy consumption-oriented surveys were conducted. In Poland, the agency responsible for collecting and publishing statistics is the Central Statistical Office (CSO). The main CSO publications addressing the topic of the characteristics of the energy consumers are CSO (2014) and CSO (2012). The aforementioned reports include a detailed information on the energy consumption quantities also with regard to such aspects like the purpose of use, the ownership of the energy, the energy consuming devices and some structural factors that influence the consumption's characteristics. Moreover, they provide information about the energy generation from renewable sources (CSO, 2014; CSO 2012). From the point of view of the proposed approach, the most valuable information, derived from the national statistics, is the one related to the percentage of households using appliances like heating, cooking, mechanical ventilation and air conditioning equipment, lighting, electrical appliances and their parameters. Referring to these statistics (the statistical representative sample size was concerning the responses of 0,0337 % (4576) of Polish households), it was possible to evaluate the probability for predefined stereotype values. Moreover, the data describing the energy load was analyzed and applied so to address and answer one of the crucial problems – what is the possibility of use of a specific device at a specified hour (PSE, 2012; Dzikowski, & Filipowska, 2014).

The approach proposed in this paper, may be perceived as a bridge between the descriptive statistical data, energy provider observations (historical usage) and the currently existing solutions such as intelligent computer systems. It bases on a combination of the prosumer descriptive data and future decisions towards the usage of appliances e.g. energy usage calendar. This information feeds the FEMS tool that analyzes it and recommends further actions (i.e. provides tips for the energy saving) to prosumers and to a microgrid administrator (Filipowska et al., 2013).

Taking all into account, the proposed approach should be considered and analyzed as an ex-ante demand estimation-support method that bases on the data provided or derived by/from the microgrid prosumers as well as the data predicted or anticipated with regard to the prosumer stereotype. This statement implicates the usage of the following data sources (1) the data on a prosumer gathered by the FEMS system that includes inter alia the energy usage calendar data (day, hour, device, appliances parameters), (2) the data regarding the stereotype of a prosumer.

The method parameters and variables

We find the following parameters:

- **α** – a probability that a prosumer of a given stereotype owns a particular device – if appliance is indicated by a user in his profile, the parameter value is equal to 1, if not the probability must be taken from a stereotype,
- **β** – a likelihood that a prosumer of a given stereotype that owns a particular device is about to use it in a given month on a given type of a day – on the beginning the parameter bases on the predefined stereotype that was prepared with regard to the existing statistical reports, lately, the FEMS system calculates the likelihood value; this parameter is further detailed by **η,**
- **γ** – an average number of hours that a given device operates in a given month on a given day – on the beginning the parameter is taken from a specified stereotype and its value is determined by statistical surveys, lately, the FEMS application calculates the parameter value,
- **δ** – a nominal power of a device in Watts – in every case, it is always a user input parameter, however may be taken from a device stereotype,
- **ε** – a number of devices – in every case, it is always a parameter given by a user, if not provided set to 1,
- **ζ** – the probability of the energy consumption by a given stereotype at the specified hour of a day – in the beginning the parameter value is taken from the stereotype and is calculated based on the statistical surveys, lately the values are calculated by the FEMS system,
- **η** – the likelihood that a prosumer of a given stereotype that owns a particular energy unit is about to use it in a given month, on a given day, at the specified hour of a day – the parameter addresses the problem of non-statistical user, who consumes the energy in a specific way.

The listing demonstrated below presents the two auxiliary variables representing the two elements of the main method:

- ***Ex(e)*** – an amount of energy expected to cover a daily energy demand in a particular stereotype on a given type of day. The ***Ex(e)*** parameter takes into account the information on an expected number of hours that a particular device works for a given stereotype in a given month on a given day:

$$Ex(e) = \alpha \times \beta \times \gamma \ \times \delta \times \varepsilon \qquad (1)$$

- ***Ex(h_i)*** – the expected value that a prosumer of a given stereotype that owns a particular energy unit is about to use it in a given month on a given day at a specified hour of a day:

$$Ex(h_i) = \frac{\zeta_i \times \eta_i}{\sum_{i=1}^{24} (\zeta_i \times \eta_i)} \qquad \textit{(2)}$$

Finally, as a consequence of multiplying *(1)* and *(2)*, we are able to calculate the amount of the energy expected (defined as ***Ex(v_i)***) to cover the hourly device energy demand of a user of a particular stereotype on a given type of day:

$$Ex(v_i) = \alpha \times \beta \times \gamma \times \delta \times \varepsilon \times \frac{\zeta_i \times \eta_i}{\sum_{i=1}^{1} (\zeta_i \times \eta_i)} \qquad \textit{(3)}$$

The Section 3 presents the implementation and validation of the proposed approach.

Validation

The following section provides a use case that shows the application and validates the proposed method. Moreover, the usage scenarios on how the suggested approach can be used within a microgrid are identified.

Use case description and validation approach

The following use case is one of the possible scenarios for a microgrid. The input conditions of the method imply that a microgrid aggregator has an access to the user energy usage calendar data via the supporting tool, in this case the FEMS system. Therefore, the most possible appliances' intention of use is determined with regard to the type of a day, hour, device and its parameters. For the purposes of the validation example, we consider a stereotype of the "2+2 persons family, week-day", that is married, has two young children at a school age. One person works from 9 am to 5 pm, leaves at 8.30 am and usually comes back at 5.30 pm. The second person leaves as 6.30 am and comes back at 3.30 pm. Kids are in school from 8 am till 5 pm. Additionally, the family usually eats and spends evenings at home on reading, watching TV and playing with their children. We assume that our prosumer (adult family members) behaves rationally (standard statistical example). The Table 1 presents the method parameters for the household in question. The list of appliances may be derived from a Central Statistical Office (CSO, 2012; CSO, 2014) or a microgrid supporting system e.g. FEMS, where their possession is declared by a user.

There are two issues that should be emphasized with regard to the Table 1. The appliances such as induction hob and instant-flow water heater, based on their nominal power, can significantly affect the overall, daily energy demand. Moreover, for the sake of simplicity, we have decided to exclude a refrigerator from our calculations and assume that it consumes energy constantly (we are aware that its real energy usage may differ on a daily basis, and is affected by the fridge opening frequency, load level, temperature differences, etc.). According to the most present device description, we assume that A++ refrigerator usually absorbs circa 0,63 kWh per day (Turakiewicz, & Pietras, 2013).

Table 1 presents the most important values in the column labeled as *Ex(e)* that corresponds to the amount of energy expected to cover a daily energy demand of a device in case of a particular stereotype on a given type of day. The next step of the method is to distribute the *Ex(e)* values on the day hours for different day types (weekdays, Saturdays and Sundays). Please note that the daily demand of an appliance depends on the day type.

Device	**α**	**β**	**γ**	**δ**	**ε**	***Ex(e)***
washing machine (weekday)	1	0,667	1,4	2200	1	2053,3
laptop (weekday)	1	1	4	60	1	240
oven (weekday)	1	0,5	1	2500	1	1250
electric kettle (weekday)	1	1	0,6	2000	1	1200
vacuum cleaner (weekday)	1	0,5	0,2	500	1	50
iron (weekday)	1	0,5	0,3	2000	1	300
dishwasher (weekday)	1	1	2	1800	1	3600
induction hob (weekday)	1	1	1	7200	1	7200
LED lights (weekday)	1	1	6	5	10	300
instant-flow water heater (weekday)	1	1	1	7000	1	7000
TV (weekday)	**1**	**1**	**4**	**150**	**1**	**600**
TV (Saturday)	**1**	**1**	**7**	**150**	**1**	**1050**
TV (Sunday/holiday)	**1**	**1**	**8**	**150**	**1**	**1200**

Table 1: The appliances and variables assumed for the stereotype "2+2 persons family".

Table 2 presents selected calculations made for the TV in September to show how the energy demand may differentiate through different days of a week.

Hour	**ζ**	**η (w-day)**	**ζ×η (w-day)**	**Ex(h) (W-day)**
00	0,1	0,2	0,02	0,0028
01	0,1	0,1	0,01	0,0014
02	0,1	0,1	0,01	0,0014
03	0,1	0,1	0,01	0,0014
04	0,1	0,1	0,01	0,0014
05	0,1	0,1	0,01	0,0014
06	0,1	0,1	0,01	0,0014
07	0,5	0,4	0,2	0,02797
08	0,5	0,4	0,2	0,02797
09	0,1	0,3	0,03	0,0042
10	0,1	0,1	0,01	0,0014
11	0,1	0,1	0,01	0,0014
12	0,1	0,1	0,01	0,0014
13	0,1	0,1	0,01	0,0014
14	0,1	0,5	0,05	0,00699
15	1	0,5	0,5	0,06993
16	1	0,7	0,7	0,0979
17	1	0,7	0,7	0,0979
18	1	0,9	0,9	0,12587
19	1	0,9	0,9	0,12587
20	1	0,9	0,9	0,12587
21	1	0,9	0,9	0,12587
22	1	0,7	0,7	0,0979
23	0,5	0,7	0,35	0,04895
Sum check			**7,15**	**1**

Table 2: Ex(v) distribution and calculation for TV for a September, week-day.

Table 3 presents results of calculations ***$Ex(v_i)$*** made for week-days, Saturdays and Sundays.

Ex(v)		
Week-day	**Saturday**	**Sunday**
1,678322	22,605	26,087
0,839161	10,172	7,826
0,839161	3,391	7,826
0,839161	2,260	5,217
0,839161	2,260	5,217
0,839161	2,260	5,217
0,839161	2,260	5,217
16,78322	5,651	13,043
16,78322	28,256	58,696
2,517483	50,861	58,696
0,839161	50,861	58,696
0,839161	50,861	58,696
0,839161	50,861	58,696
0,839161	50,861	58,696
4,195804	61,033	70,435
41,95804	61,033	70,435
58,74126	61,033	70,435
58,74126	61,033	70,435
75,52448	61,033	70,435
75,52448	91,550	105,652
75,52448	91,550	105,652
75,52448	91,550	93,913
58,74126	91,550	82,174
29,37063	45,210	32,609
600	**1050**	**1200**

Table 3: Calculated *$Ex(v_i)$* values.

The TV's (150W of nominal power) anticipated energy demand (600W) is depicted in the Figure 1. The highest probability of the electricity demand of this appliance is on Sunday (dark grey line) and Saturday (light grey line) mostly due to the fact, that family spends time at home.

With regard to the specifics of a selected stereotype, a lower probability of energy consumption during the working hours is assumed (due to spending that time working outside home). Moreover, we can anticipate that the possible energy consumption of TV might last from 7 to 9 am on non-working days and from 3 to 12 pm. On Sundays and holidays we might expect rather a constant energy consumption during a whole day. Of course, a probability can differ depending on the hour of a day. Higher energy consumption during evenings might be caused by the fact that people are resting and watching TV shows. We must understand the fact that presented scenario is just one of the many possible scenarios that assume that a family behaves on average. These values may be however, overwritten by the values quantified for a specific user.

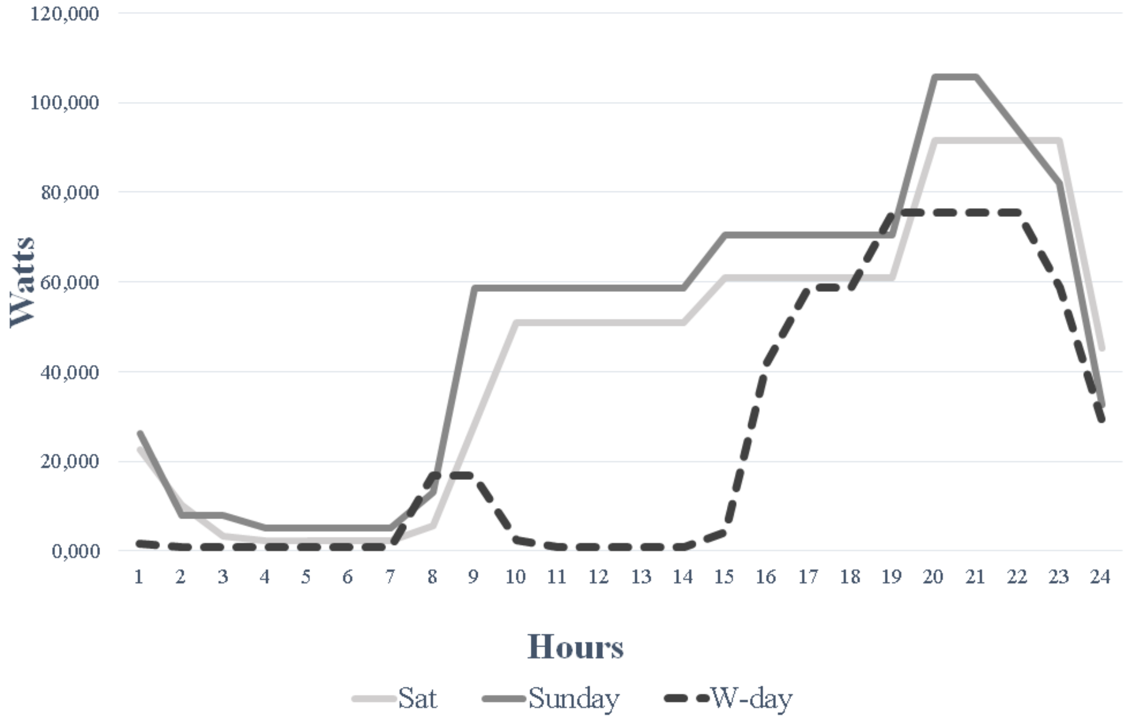


Figure 1: A TV's electricity demand profile obtained with the use of the proposed method for different types of days in September for a "2+2 persons family" stereotype.

Furthermore, after taking into account all appliances' profiles that are listed in the Table 1, the preparation of an overall energy demand profile is possible. Such an energy profile, depicted in the Figure 2 (dotted line), is tailored to a given use case stereotype of the "2+2 persons family, week-day". Moreover, the obtained, possible consumption stereotype profile is compared with the existing standard profiles dedicated to households and used by energy operators (G11). These are typically used by the four main Polish energy operators to prepare demand forecasts such as ENERGA, PGE S.A., Tauron Polska Energia S.A. and ENEA S.A. According to the Polish law, the standard profile is a collection of the data on the average electricity consumption in individual hours of the day estimated for a group of the end users (MoE, 2007).

As it may be observed in the Figure 2, the result of an application of the proposed method differs from the energy profiles prepared for a typical household, but these are only a general representations on how the users behave (generalizing all households into one category). Under the Polish energy market regulations, the rate for the electricity that residential users pay is standardized and results from the electricity tariff – G. Our research shows however, that standard profiles differ significantly from the real usage data (based on data obtained from smart meters or users). This may be due to the fact that they are based mainly on the historical energy usage data of a big group of users and they are not fulfilling the needs of the new generation of the energy consumers/prosumers.

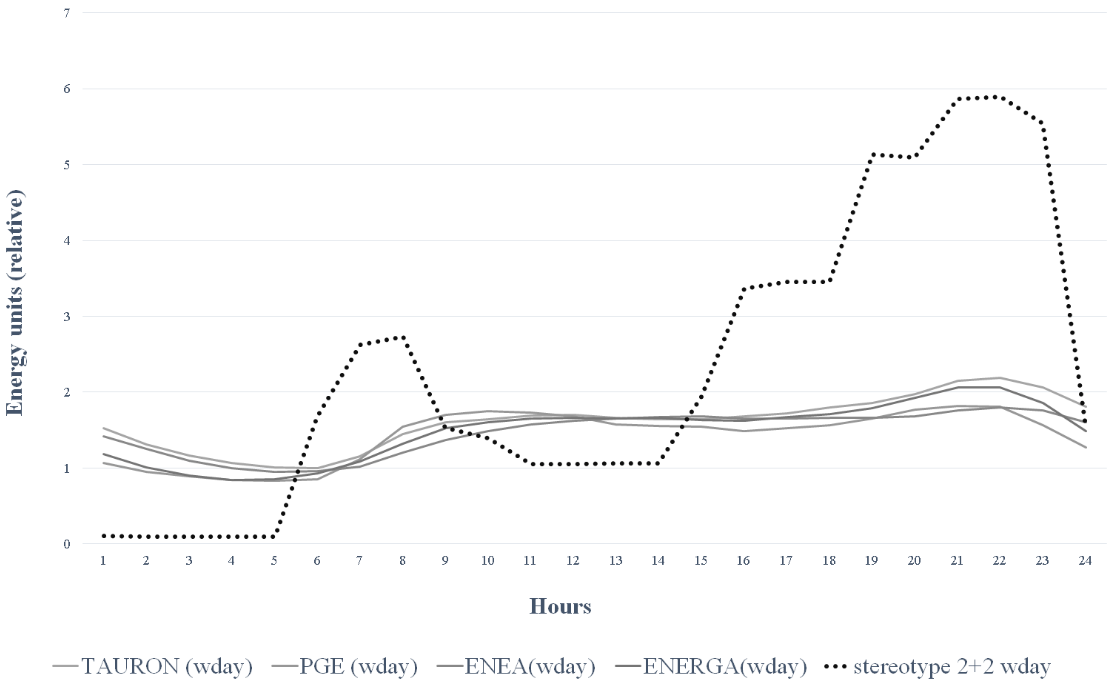
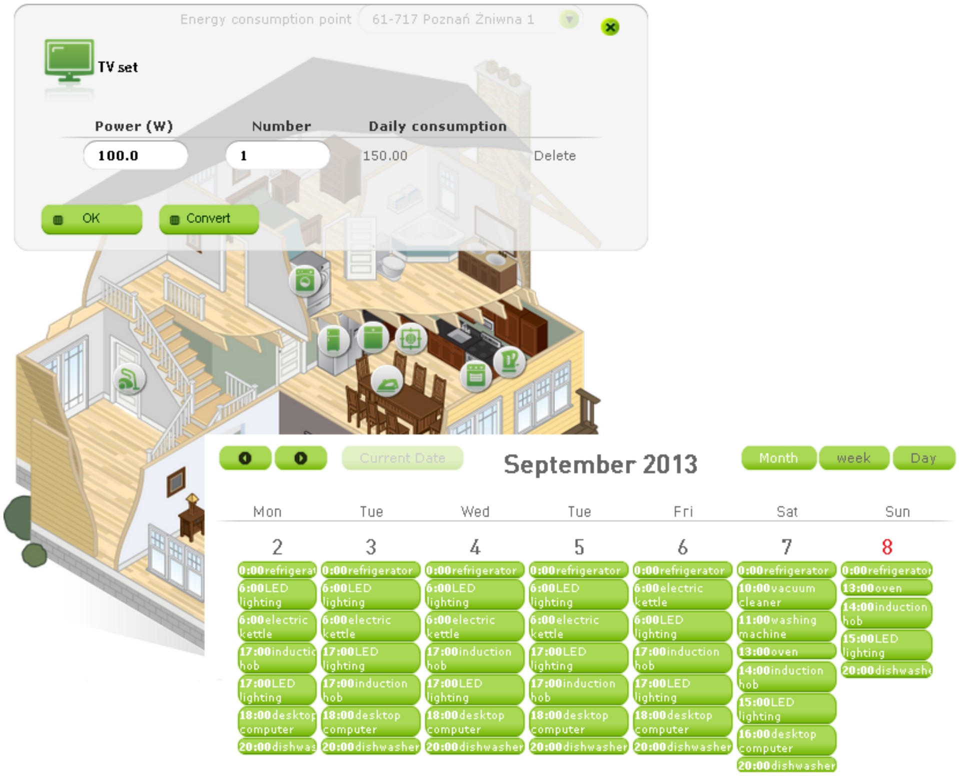


Figure 2: The electricity demand profile obtained with the use of the proposed method in comparison to G11 standard profile typically used by the main polish energy operators (TAURON PE, PGE, ENEA, ENERGA).

The proposed method allows preparing the estimation of the energy consumption that is based on the information delivered by the prosumers (or induced based on their historical usage). The energy consumption level is therefore personalized and estimated in accordance to the "real", not strongly standardized, energy usage patterns. Now, considering the abovementioned profile for a one particular household, an aggregator is ready to prepare similar calculations for the whole group of prosumers in the microgrid. We believe that this way, the aggregator is able to lower the overall costs of the electric energy, strengthen users' awareness (by presenting results to users) and improve the energy efficiency (Filipowska et al., 2013).

FEMS as a tool implementing the proposed method

The proposed method is implemented by the Future Energy Management System (FEMS) – the management supporting system that appeared on the Polish market in 2013. The system was developed for such entities as: prosumers (both single as well as groups of stakeholders e.g. local communities consisting of energy consumers and producers), aggregators (i.e. household administrators) and energy operators (Filipowska et al., 2013). The FEMS aims inter alia at acquisition of data from smart meters, reasoning over the grid model, providing recommendations, forecasting and methods for retrieving information from external sources. FEMS is also a service that can strengthen the user engagement in the field of smart consumption within a microgrid whilst providing intuitive, user-friendly interfaces as the one depicted in Figure 3.



TV set

Power (W)	Number	Daily consumption	
100.0	1	150.00	Delete

September 2013

Mon	Tue	Wed	Tue	Fri	Sat	Sun
2	3	4	5	6	7	8
0:00refrigera	0:00refrigerator	0:00refrigerator	0:00refrigerator	0:00refrigerator	0:00refrigerator	0:00refrigerato
6:00LED lighting	6:00LED lighting	6:00LED lighting	6:00LED lighting	6:00electric kettle	10:00vacuum cleaner	13:00oven
6:00electric kettle	6:00electric kettle	6:00electric kettle	6:00electric kettle	6:00LED lighting	11:00washing machine	14:00induction hob
17:00inducti hob	17:00LED lighting	17:00induction hob	17:00LED lighting	17:00induction hob	13:00oven	15:00LED lighting
17:00LED lighting	17:00induction hob	17:00LED lighting	17:00induction hob	17:00LED lighting	14:00induction hob	20:00dishwashe
18:00deskto computer	18:00desktop computer	18:00desktop computer	18:00desktop computer	18:00desktop computer	15:00LED lighting	
20:00dishwas	20:00dishwasher	20:00dishwasher	20:00dishwasher	20:00dishwasher	16:00desktop computer	
					20:00dishwasher	

Figure 3: The interactive FEMS GUI for adding user appliances and the appliances' usage calendar.

The system was developed to meet the needs of diverse prosumer stereotypes. These stereotypes were derived from the literature as well as while experimenting with the real users' data. The information on a family stereotype, a type of the building, as well as devices owned, is provided by prosumers while creating their personal account (or updated later). Moreover, the users can define their energy consumption calendar, energy consumption preferences, building description, etc. The incentives for providing this information relate to the maintenance fee imposed by the aggregator. Besides the data gathered directly from a family, the system retrieves from various Internet sources data on factors that may influence the energy consumption i.e. weather forecasts and profiles of devices that include data such as the nominal power and the energy class. This way, an aggregator is equipped with a tool that enables him better microgrid management as he provides the users with recommendations and suggestions on how to increase the energy efficiency.

To conclude, both FEMS and the proposed method require some essential data provided by the user e.g. on devices possessed, energy consumption schedule, building description and its main features, preferences and stereotype name. Afterwards, a user has a chance to save some money, improve the energy awareness and actively participate in the local microgrid. In turn, a microgrid aggregator (administrator) may prepare accurate estimations of energy usage of a user and a group he belongs to, and send him recommendations such as "In the nearest future you should think about replacing your old "B" class refrigerator. The new "A++" refrigerator consumes less energy and within a year, I'm sure you would save circa 50% on your energy bill!".

Related work

The related work section is divided into two parts. The first one describes energy forecasting methods. The second considers the current approaches to the energy prosumers profiling. Delivering an efficient forecast (characterized by a high accuracy, that in this case should be understood as a low prediction error) of the energy demand in a short-term horizon is one of the biggest challenges for the energy grid. The accuracy of a short-term forecast determines the costs of the energy trading at the next day or intraday energy market. There are many approaches to the energy estimation in the literature e.g. the regression line method or the curve fitting method that use the data on temperature, humidity, day type parameters and also the historical demand. The forecasting accuracies are very good with an error less than 3% for almost all day types and all seasons (Jain, Nigam, & Tiwari, 2013). The short-term forecasting can be also supported by various regression models (Liu, Huang, & Hsien, 2005; Zhang, & Li, 2011). Moreover, different methods like neural networks, multiple classifier systems, evolutionary programming, orthogonal lest squares and genetic algorithms are used to create the more accurate short-term energy forecasts (Ye et al., 2006; Chan et al., 2011). There are also approaches based on autoregressive methods (ARMA, ARIMA, EGARCH) and exponential smoothing (Contreras et al., 2003; Bowden, & Payne, 2008).

The prediction of the energy demand in the medium time horizon is also being researched. The most sophisticated approaches use the knowledge based systems. Among these, methods which use multi-layer perceptrons with a back-propagation feed-forward algorithms are characterized by a low forecast error (MAPE at level of 2-3%), but they need a relatively larger number of training data in comparison to the canonical statistical methods (Falvo et al., 2006; Salama et al., 2009). For a medium-term forecasting, there are also methods based on autoregressive models, like ARIMA (Churueang, & Damrongkulkamjom, 2005). The long-term energy demand forecasts are mainly used for planning investment in the network assets. One of many approaches is the long-term forecasting based on basic frameworks of fuzzy neural network and particle swarm optimization. This combination of methods enables the intelligent microgrid system to control the electrical supply for achieving the best economical and power efficiency (Wai, Huang, & Chen, 2012). Other methods use e.g. adaptive intelligence networks, nonlinear autoregressive models, but also simple linear regression (Al-Gandoor et al., 2007; Q.Wang, X. Wang, & Xia, 2009). Moreover, for energy demand analysis and forecasting there are approaches based on autoregressive methods like ARIMA, applied individually or combined with genetic algorithms (Gaetan, 2000; Erdogdu, 2007).

The subject of the user profiling is widely described and many solutions are proposed (Poo, Chng, & Goh, 2003). The prosumer profiling is based usually only on the historical energy consumption, without considering other factors (Montanari, & Siwe, 2013). However, there are also approaches that, similarly to what is presented within the article, extend the user profile with other factors that may have an impact on user's energy consumption (Lampropoulos, Vanalme, & Kling, 2010; Rathnayaka et al., 2011). However, they mainly focus on prosumer's consumption, energy generation and energy storage behavior apart from factors such as habits in the energy usage of a particular household, the types of energy consumption or generation devices, other relevant energy characteristics of a household etc. In the literature, there are also examples of the energy generation data obtained from prosumers that are used to support the process of energy generation forecasting (Rathnayaka et al., 2012).

The regular approach to forecasting the energy consumption used by the Polish energy operators considers the application of a standard energy profile. As it was mentioned before, the standard energy profile is a collection of data describing an average, hourly electricity consumption of the day that was estimated for a group of the end users (MoE, 2007). Besides this definition, the Polish energy law does not impose a need for introduction of a detailed method for calculating a standard customer energy profile. Therefore, each energy operator prepares his own method of describing the energy profile and publishes these profiles in the document called "Traffic and operation instruction of distributed network". As a result, every energy operator in Poland has a different method for anticipating the energy consumption. For example PGE Dystrybucja S.A. for the household-oriented tariff, publishes a standard profile in the form of a table, where columns represent the consequent hours of a day and the rows represents next months in a year. Additionally each month is divided in three types of days (from Monday to Friday, Saturdays and Sundays) (PGE, 2013). Furthermore, the energy consumption values are in relative units, so only the interpretation of variability of the energy demand is possible. Such approach ignores the possible abnormal energy usage in the days before important holidays such as e.g. Christmas Eve. Few operators, like i.e. Energa Operator S.A., release a standard energy profile for every hour of every day in the year separately. On the other hand, Tauron Dystrybucja S.A. publishes standard profiles, where days are distinguished by the type (workday or holiday) and by the time of a year (summer or winter day) (Tauron, 2008). In this case the Taurons' approach ignores the differences of the electricity consumption in different months of the year.

While preparing a standard energy profile, a large amount of the energy consumption data, that is gathered from all anonymous energy consumers, is being used. Thus, during the estimation process the law of large numbers (LLN) occurs, thus forecasts of the energy load for a large group of energy consumers are well fitted. Unfortunately, the use of LLN for a single customer or even small group of customers does not give satisfactory results. Therefore, to achieve reliable results, it is necessary to investigate methods based on the prosumers profiling.

Conclusions and future work

In this paper, we presented and validated an energy-demand forecasting method based on prosumers' profiling that enables estimation of the energy demand for diverse user stereotypes, for every hour, every day of the year and for various devices. The method is integrated with the microgrid management tool, namely Future Energy Management System and is subject to further evaluation while real usage.

In our approach, we consider the active user engagement that is manifested by providing the system with all data via the FEMS GUI. Moreover, the method based on the prosumers' profiling aims at changing the widely used approach to forecasting the electric energy load using mainly the analysis of the historic (ex-post) data, in particular as regards the estimation of a load for a single prosumer.

For the future work, we intend to explore and optimize the proposed method, especially with regard to the business stereotypes (companies, restaurants, factories, etc.). Secondly, to ensure the general usability of the method, we plan to rigorously test it using the dataset obtained using the Future Energy Management System (or other available).

Acknowledgement

The research results described within the article were achieved within the project: Building a prototype of an innovative system for forecasting the level of consumption and production of electric energy “Future Energy Management System”, (Project: UDA-POIG.01.04.00-30-065/10-03, Value: 9.957.022 PLN, UE contribution: 5.647.005 PLN, period 01.07.2011-31.03.2013), European Funds – for the development of innovative economy (Fundusze Europejskie - dla rozwoju innowacyjnej gospodarki).

References

Adomavicius, G. & Tuzhilin, A. (2001). *Using data mining methods to build customer profiles*, Computer, vol. 34 (pp. 74–82).

Al-Gandoor, A., Nahleh, Y.A., Sandouga, Y. & Al-Salaymeh, M. (2007). *A multivariate linear regression model for the jordanian industrial electric energy consumption*, The 16th IASTED International Conference on Applied Simulation and Modelling (pp. 386-391), Anaheim, USA

Brendal, B.A. (2011). *A Prosumer Oriented Energy Market*, The IMPROSUME project, IMPROSUME Publication Series #2, NCE Smart Energy Markets, Norway

Bowden, N. & Payne, J.E. (2008). *Short term forecasting of electricity prices for MISO hubs: Evidence from ARIMA-EGARCH models*, Energy economics (pp. 3186-3197), Vol. 30.2008.

Cena Prądu. (2012). *Kalkulator zużycia energii dla gospodarstw domowych – grupa taryfowa G11,* Retrieved November 20, 2012, from http://www.cenapradu.republika.pl/kalkulator.html

Central Statistical Office. (2013). *Construction – Activity Results in 2012,* Warsaw, Poland.

Central Statistical Office. (2012). *Dochody i warunki życia ludnosci Polski – raport z badania EU-SILC 2011*, Warsaw

Central Statistical Office. (2012). *Energy consumption in households in 2009*, Statistical Information and Elaborations, Warsaw.

Central Statistical Office. (2014). *Energy consumption in households in 2012*, Statistical Information and Elaborations, Warsaw.

Central Statistical Office. (2013). *Sytuacja gospodarstw domowych w 2012 r. w świetle wyników badania budżetów gospodarstw domowych*, Retrieved September 19, 2013, from http://old.stat.gov.pl/cps/rde/xbcr/gus/WZ_sytuacja_gosp_dom_2012.pdf

Chan, P.P.K., Wei-Chun, C., Ng, W.W.Y. & Yeung, D.S. (2011). *Multiple classifier system for short term load forecast of Microgrid,* Cybernetics International Conference on Machine Learning, Vol.3, (pp.1268-1273).

Churueang, P. & Damrongkulkamjorn, P. (2005). *Monthly energy forecasting using decomposition method with application of seasonal arima*, *The 7th International Power Engineering Conference* (pp. 223-229), Singapore.

Contreras, J., Espinola, R., Nogales, F. & Conejo, A. (2003). *ARIMA models to predict next-day electricity prices*, *IEEE Transactions on Power Systems* (pp.1014-1020), Vol. 18.

Dzikowski, J. & Filipowska, A. (2014). *Czynniki kształtujące popyt i podaż na rynku energii elektrycznej – podejście modelowe,* Matematyka i informatyka na usługach ekonomii (pp.79-92). Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Erdogdu, E. (2007). *Electricity demand analysis using cointegration and ARIMA modelling: A case study of Turkey*, Energy Policy (pp. 1129-1146), Vol. 35.

Filipowska, A., Fabisz, K., Hossa, T., Mucha, M. & Hofman, R. (2013). *Towards forecasting demand and production of electric energy in smart grids*, Perspectives in Business Informatics Research, 12th International Conference, Poland.

Falvo, M.C., Lamedica, R., Pierazzo, S. & Prudenzi, A. (2006) *A Knowledge Based System for Medium Term Load Forecasting*, Transmission and Distribution Conference and Exhibition (pp.1291-1295), Dallas, Texas.

Gaetan, C. (2000). *ARMA model identification using genetic algorithms*, Journal of Time Series Analysis (pp. 559-570), Vol. 21.

Ghanbarian, M., Kavehnia, F., Askari, M.R., Mohammadi, A., & Keivani, H. (2007). *Applying Time-Series Regression to Load Forecasting Using Neuro-Fuzzy Techniques.* Presented at International Conference on Power Engineering, Energy and Electrical Drives, Setubal, Portugal.

T. Hossa, A. Filipowska, & K. Fabisz. The comparison of medium-term energy demand forecasting methods for the needs of microgrid management. In Proceedings of SmartGridComm, IEEE International Conference on Smart Grid Communications, 2014.

Jain, M.B., Nigam, M.K. & Tiwari, P.C. (2012). *Curve fitting and regression line method based seasonal short term load forecasting*, 2012 World Congress on Information and Communication Technologies (pp.332-337).

Koch, N. (2000). *Software Engineering for Adaptive Hypermedia Systems*, PhD thesis, Ludwig-Maximilians-Universität München.

Lampe, C.A.C, Ellison, N. & Steinfield, C. (2010). *A familiar face(book): profile elements as signals in an online social network*, CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (pp. 435–444).

Lampropoulos, I., Vanalme, G.M.A & Kling, W.L. (2010). *A methodology for modeling the behavior of electricity prosumers within the smart grid*, 2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (pp.1-8), Gothenburg.

Liu, S., Huang, S. & Hsien, T. (2005). *Optimal identification of self-reunion multiple regression (SRMR) model based on regression function for short-term load forecasting*, The 7th International Power Engineering Conference (pp.1-46).

Livgard, E.F. (2010). Smart metering - opportunity or threat to the power industry?. Presented at International Conference on Power System Technology, Hangzhou.

Ministerstwo Gospodarki. (2007). *Rozporządzenie Ministra Gospodarki z dnia 4 maja 2007 r. w sprawie szczegółowych warunków funkcjonowania systemu elektroenergetycznego* (Dz. U. z dnia 29 maja 2007 r.)

Montanari, U. & Siwe, A.T. (2013). *Real time market models and prosumer profiling*, 2013 IEEE Conference on Computer Communications Workshops (pp.7-12), Turin.

Ntawanga, F., Calitz, A.P., & Barnard, L. (2008). *Maintaining customer profiles in an e-commerce environment*, SAICSIT '08: Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries, New York, NY, USA, ACM (pp. 169–179).

Parkinson, S., Wang, D. & Djilali, N. (2012*). Toward low carbon energy systems: The convergence of wind power, demand response, and the electricity grid.* Innovative Smart Grid Technologies, Tianjin.

Poo, D., Chng, B. & Goh J. (2003) *A hybrid approach for user profiling*, System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference (pp.6-9), Washington

Rathnayaka, A.J.D., Potdar, V.M., Dillon, T., Hussain, O. & Kuruppu, S. (2012). *Analysis of energy behaviour profiles of prosumers*, 2012 10th IEEE International Conference on Industrial Informatics (pp.236-241), Beijing

Rathnayaka, A.J.D., Potdar, V.M., Hussain, O. & Dillon, T. (2011*). Identifying prosumer's energy sharing behaviours for forming optimal prosumer-communities*, 2011 International Conference on Cloud and Service Computing (pp.199-206), Hong Kong.

Saha, A.K., Chowdhury, S.P. & Chowdhury, S. (2007) *Application of adaptive network-based fuzzy inference system in short term load forecasting*, International Conference on Information and Communication Technology in Electrical Sciences *(*pp.168-174), Chennai, India.

Salama, H.A.E., El-Gawad, A.F.A., Sakr, S.M., Mohamed, E.A. & Mahmoud, H.M. (2009). *Applications on medium-term forecasting for loads and energy scales by using Artificial Neural Network*, 20th International Conference and Exhibition on Electricity Distribution (pp.1-4), Prague, Czech Republic.

Shandurkova, I., Bremdal, B.A., Bacher, R., Ottesen, S. & Nilsen, A. (2012). *A Prosumer Oriented Energy Market – Developments and future outlooks for Smart Grid oriented energy markets*, The IMPROSUME project, IMPROSUME Publication Series #3, NCE Smart Energy Markets, Norway

Singh, R., Alapatt, G.F. & Poole, K.F. (2012). *Photovoltaics: Emerging role as a dominant electricity generation technology in the 21st century*. International Conference on Microelectronics, Nis.

Taft, J.D. (2012). *Emerging smart grid control trends and implications for control architecture*. Transmission and Distribution Conference and Exposition, Orlando, Florida.

Turakiewicz, J., Pietras, K. (2012). Ile prądu zużywasz w kuchni? Testy sprzętu AGD, *Regiodom*, Retrieved September 20, 2013, from http://regiodom.pl/portal/wnetrze/rtv-agd/ile-pradu-zuzywasz-w-kuchni-testy-sprzetu-agd

URE. (2012). *Charakterystyka rynku energii elektrycznej*, Retrieved November 21, 2013, from http://www.ure.gov.pl/pl/rynki-energii/energia-elektryczna/charakterystyka-rynku/5314,2012.html

Wang, Q., Wang, X. & Xia, F. (2009). *Integration of grey model and multiple regression model to predict energy consumption,* International Conference on Energy and Environment Technology (pp. 194-197), Vol. 1, Guilin, Guangxi.

Wahlster, W. & Kobsa, A. (1989). *User Models in Dialog Systems*, Springer.

Wai, R., Huang, Y. & Chen, Y. (2012). *Design of intelligent long-term load forecasting with fuzzy neural network and particle swarm optimization*, 2012 International Conference on Machine Learning and Cybernetics (pp.1644-1650), vol.4, Xian.

Ye, B., Yan, N.N., Guo, C.X. & Cao Y.J. (2006). *Identification of fuzzy model for short-term load forecasting using evolutionary programming and orthogonal least squares,* Power Engineering Society General Meeting, Montreal, Quebec.

Zhang, M. & Li, L. (2011) . *Short-term load combined forecasting method based on BPNN and LS-SVM,* Power Engineering and Automation Conference (pp.319-322), Wuhan.

4.4 Conclusions

The chapter belongs to the part of the thesis aiming at demonstrating importance of profiling for vertical domains. Its goal was to "create a profile of a user or a thing that will be applicable for solutions enabling management of production and consumption of the electric energy in the smart grid". This goal was further translated into two secondary goals addressed by specific sections of this chapter, namely:

- Proposing an architecture of a system for monitoring energy production and consumption in the smart grid, taking into account a profile of an individual prosumer.
- Creating a profile of a user for the needs of electric energy supply: monitoring and describing demand for the electric energy to be used by the system enabling management of the production and consumption of the electric energy in the smart grid.

In relation to the first of the goals mentioned, the paper included in Section 4.2 was to provide a requirements analysis and an architecture of the system enabling management of the production and consumption of the electric energy in the smart grid. Section 4.2 studies definition and challenges for the microgrid and in relation to them proposes requirements for a system supporting management of prosumers in the smart grid. The concept and the architecture of a system, especially enabling not only management, but also forecasting of production and usage of the electric energy (including acquisition of data from the Web to enable for improved reasoning), are proposed. The system was developed within the Future Energy Management System (FEMS) project which was delivered for a business partner, following the consortium research methodology.

The second goal of the chapter was detailed into presenting and validating a method for estimation of the demand for the electric energy in microgrids based on profiling of a prosumer that enables to determine the energy demand. The paper included in Section 4.3 explains definition of a profile and stereotype in a microgrid. Then, a method for demand estimation taking into account features of a prosumer, including appliances used, is proposed. The application of the method in the smart grid management software is also demonstrated. It should be underlined that the method is nowadays a working method, used for prediction of the electric energy usage by a single household in the FEMS system.

The following chapter concerns application of profiling for another public utility i.e. telecommunication.

Chapter 5

Profiling for Utilities: Telecommunication

5.1 Introduction

5.1.1 Motivation

A significant part of work of the author within the last 8 years concerned researching the domain of telecommunication within the framework of the Dynamic Social Network project carried out in cooperation with Orange SA. Besides of development of methods enabling for modelling a profile of a user or strentgh of a relation between two or more users in the telecommunication network, an emphasis was put on studying business potential related to these methods and scenarios.

The goal of this chapter is to demonstrate how ”to develop a profile of a customer/subscriber to telecommunication services, enabling for personalisation and taking into account the issues of privacy and trust” and to study potential of such a profile from a business perspective. Secondary goals of the thesis, that are addressed in this chapter, include:

- Defining a solution that enables to manage personal information in telecommunication.
- Proposing methods enabling for user profiling based on Call Detail Records data.

5.1.2 Structure of the chapter

The chapter consists of four sections including an introduction presenting relation to thesis' goals and a summary that presents results that were achieved in relation to these goals. Section 5.2 focuses on achieving the first of the secondary goals referred to in this chapter. Section 5.3 relates to proposing methods enabling for user profiling based on Call Detail Records data.

5.2 Managing Personal Information: A Telco Perspective

The goal of the section, and the paper accordingly, is to propose a set of methods for managing personal information to enable new application scenarios. The supplementary goal of the solution envisioned is to empower a user to update and manage his personal data. As such this section contributes to achieving the following secondary goal of the thesis: ”Defining a solution that enables to manage personal information in telecommunication”.

The paper included in this section, was accepted to 19th Conference on Innovations in Clouds, Internet and Networks, ICIN 2016, Paris, 1-3.03.2016 and published in the conference proceedings. The detailed bibliographic reference is as follows: Filipowska, A., Szczekocka, E., Gromada, J., Brun, A., Portugal, J., Jankowiak, P., Kałużny, P., Staiano, J., 2016, Managing Personal Information: A Telco Perspective, 2016 Conference on Innovations in Clouds, Internet and Networks, ICIN 2016, Paris, March 1-3, 2016, pp. 112-119.

19th International ICIN Conference - Innovations in Clouds, Internet and Networks - March 1-3, 2016, Paris.

Managing Personal Information: A Telco Perspective

Ewelina Szczekocka*, Justyna Gromada*, Agata Filipowska†, Piotr Jankowiak†,
Piotr Kałużny†, Arnaud Brun‡, Jean Michel Portugal‡ and Jacopo Staiano§

*Orange Polska S.A., Warsaw, Poland
‡Orange Labs Research, Paris, France
†Department of Information Systems, Poznan University of Economics, Poznan, Poland
§LIP6 - UPMC, Sorbonne Universities, Paris, France

***Abstract*—While paving the way for novel and exciting application scenarios, the foreseen large-scale deployment of *connected objects* poses a number of ethical concerns which, if not timely addressed, can potentially dampen the full potential of this drive. In particular, issues related to privacy and control of the data collected by sensors can undermine users' *trust*, hence hindering several application scenarios due to the perceived sensitivity of the data required. In this work, we describe a first prototype which addresses such issues by granting users full ownership and control of the data their sensors produce. The proposed solution - the Personal Information Management platform - aims at *breaking the data silos*, allowing the spontaneous emergence of ad-hoc and decentralized communities and applications, and *gaining user trust* by providing transparency on personal data and its usage with a user-centric approach. Furthermore, we provide a sample use case application based on the platform, enabling social activities of a community supported by data from objects belonging to Internet of Things and personal information inferred from this data. Finally, we discuss current limitations and elaborate on future developments of the described technology.**

I. Introduction

The current wave of commercial offerings regarding the *Internet of Things* and *connected objects* opens exciting possibilities for designing applications tackling a variety of problems in everyday life. For instance, it will be possible to propose, implement, and test novel approaches to issues of high societal value (e.g. energy consumption optimization, food waste reduction). Nonetheless, it can be speculated that the efficacy of such solutions will be highly influenced by human factors: these connected sensors will allow for extensive collection of sensitive and personal information, as the raw data gathered will encode details about people's behavior, habits, preferences, interactions, social activities and so forth. Looking back at the last years, it is clear that the emergence of the "Web 2.0" and the wide adoption of smart-phones have, on the one hand, allowed for significant technological breakthroughs (e.g. Big Data solutions) while, on the other, the dominant business models have relied on offering "free" services in exchange for the commercial exploitation of personal data: for instance, it is very common to implement user profiling in order to serve targeted advertisements.

Although these models are still widely in use, the striking unbalance in terms of effective data control between the data producers (the users) and the data retainers (the service providers) has lately generated several concerns, bringing policy-makers to start dealing with the issues of privacy and data control, researchers to devise possible solutions, and an increasing amount of users to grow awareness on the matter and find ways to protect their privacy. The focus of this work is on personal data of customers and on building a platform which offers full control over usage of personal data in customers' social activities supported by different services. Moreover, this platform enriches social communication services by injecting into them knowledge derived from personal data in a *trust-by-design* fashion. We believe that this approach can create significant value from personal data for both customers and operators. Previous related work focused on modeling customers' relations based on data from their communications – a first step in building a value from their personal data; it soon became clear that this is not enough to provide a complete offer to a customer. Service providers, and particularly Telco operators, nowadays find themselves in the unfavorable position of losing their customers' *trust*. In this work we propose a prototype solution – the Personal Information Management (PIM) platform – able to grant its users (Telco customers) full control on their data, and to allow them to take conscious and informed decisions to e.g. make their data available to other users or services.

Preliminary user studies have shown significant lack of user awareness of privacy risks, e.g. during their internet activities. Hence, the proposed approach is expected to also contribute towards an increase in customers' awareness on the effects of their data-sharing choices, on how their data is used and by whom.

The main contributions of this paper can be summarized as follows:

- we describe the Personal Information Management platform and depict factors like methods, algorithms, tools allowing for building a smart system predicting and prompting to a user different usages based on his social context;
- we report on a first deployed application based on the

platform along with the feedback obtained in the associated user studies;

- we elaborate on current limitations and future developments of the proposed platform.

II. Related Work

In this section, we provide details on research activities related to this work. In order to contextualize the proposed platform, and to highlight the importance of allowing behavioral data collection in a *trust-by-design* paradigm, we first briefly report on works harnessing the potential of big data processing and we highlight how to adopt and use results of this processing from a single user perspective; then, we describe proposed solutions for granting users control on the data they produce.

A. Behavioral Data Research

In order to obtain useful insights on their customer base and provide advanced analytic capabilities [1], Telcos have in the last years heavily relied on user profiling techniques, which proved particularly valuable to design personalized services and modeling customer groups [2]. The profile of a user can be seen as a *machine processable description* [3] of his/her behavior, encoding several facets such as calling/messaging patterns, i.e. the user's network activity and derived information ([4], [5], [6]), and mobility information, i.e. places the user visits, derived by the location of the Radio Base Stations (RBS) his/her mobile phone connects to ([7], [8]). Furthermore, additional information can be derived by the user's position in the graph of Telco customers [9]; The most employed source of data for these studies is represented by Call Detail Records (CDRs)[1], which hold a variety of information[2] allowing also to estimate the user's current location.

Previous works have shown how the mobility of people is characterized by very similar patterns [10], and display distinct motives over depending on the temporal resolution adopted [11]; these findings confirmed the intuition that people tend to visit only a few places on a regular basis, hence further research has focused on defining and identifying such important locations[3] ([12], [13]). Based on these anchor points, several approaches have tried to uncover user daily travels/trips ([8], [14], [15]) by identifying the places where users stopped, those he/she passed by rapidly (probably not having any activity in there), and so on. Moreover, researchers have exploited area labelling and Points of Interest (POI) to estimate the activity of a user in a given place [16]. Accounting for temporal information, it is then possible to derive even more precise information about user's habits and preferences: in [17], the authors investigated the relations between the important locations in users' lives (home, work, other) and their social interactions, hinting at the impact that mobility profiling can have on deriving information on the users social sphere.

Besides focusing on profiling individuals or customer segments, recent efforts have used aggregated mobility information, derived by CDRs or Global Positioning System (GPS) traces, to tackle problems of societal relevance, such as floodings and emergency response ([18], [19]), mapping the propagation of diseases such as malaria [20] and H1N1 [21], and to predict crime hotspots [22]. These works well exemplify the enormous advantages of big data processing for society, and highlight the need of devising data-sharing solutions able to meet the privacy demands of the data producers.

B. Privacy and Data Control

Research conducted in living-lab settings has recently shown the desire of internet and smartphone users of being granted more control on their own data ([23], [24]). In particular, users have been found to associate higher relative monetary value to their location information, hence considering it the most sensitive behavioral trace. Several privacy-preserving platforms have been developed in the last few years: in particular, the OpenPDS [25] provides a "SafeAnswers" mechanism, allowing data access to collectors only at the user discretion (and at a level of aggregation specified by the user); moreover, systems inspired by crypto-currencies have been developed in a decentralized manner, such as Ubiquitous Commons[4] and Enigma [26].

III. The PIM Platform

Under current paradigms, massive amounts of personal data are automatically collected by service providers: users leave their digital footprints in several systems whenever they e.g. perform a web search, install a mobile app, make a call or message someone, post on social media and so on. Surely, all this data can be used to the users' benefit, by e.g. providing an improved, more responsive experience. Nonetheless, a striking unbalance between users and service providers exists: in fact, the former have no power to act on the uploaded data (e.g. to delete it) nor can control the way their data is exploited by the providers and third parties. With the currently increasing public awareness on privacy matters, this issue seems very likely to undermine customers' *trust*: even if a Telco operator would treat its customers data with the highest caution and responsibility, without transparency it would not be perceived differently from other, less cautious, providers, hence not improving its reputation despite its efforts.

In this work, we present a possible solution to this problem, from a Telco perspective. Based on user-centric data services, the *Personal Information Management* (PIM) platform is a technological facility devised to grant users control over their data: the PIM allows them to visualize the data they produced, to allow for selective actions on it (e.g. partial deletions), to control who has access to it and to do what, and to foster the bottom-up emergence of novel applications by promoting data reuse.

[1] Also called "billing logs".

[2] Who, using which service, has communicated with whom, during which time and where the event happened.

[3] Also referred to as "anchor points".

[4] http://www.ubiquitouscommons.org/

Figure 1 provides a graphical representation of the amount and variety of data collected by Telco operators: data coming from service usage (event logs, mobility, contracts, etc.). It should be noted that collection and storage of such data is delegated to several subsystems (*silos*), and can legally exclusively be used for the purposes they were collected. For instance, this means that billing data can only be used for billing purposes in agreement with a contract between a telco operator and a customer, e.g. when a customer agreed for his data to be used for marketing purposes, his profile is harnessed to propose additional offers or service improvements. Usually, these operations rely on statistical approaches and simple customer segmentation.

The scope of PIM, hence, becomes two-fold: on the one hand, it is instrumental for granting customers full control over the data they generate; on the other, by providing them with the choice of opening their data for specific purposes, it can be seen as an enabler for novel applications of societal and/or business value. The goal of the proposed system can thus be summarized as making the Telco operator data space *secure*, *trustable*, *open*, and *social*.

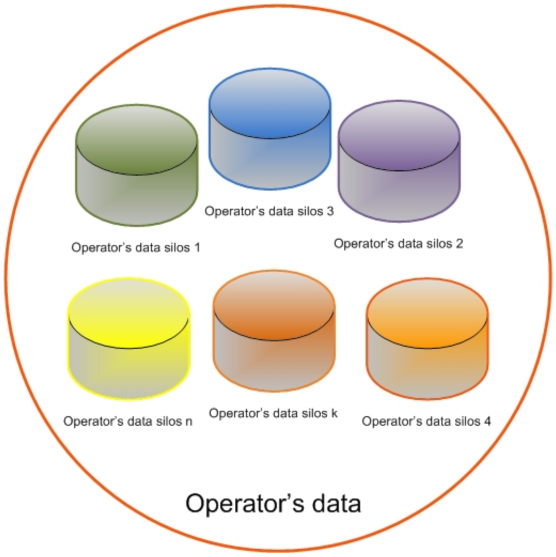


Figure 1. Data collected and protected by Telco operators.

Furthermore, we envision PIM as the platform which will allow Telcos for transition from the model shown in Figure 1 to the one represented in Figure 2. In the latter, all data produced belongs and is managed by the specific customer who produces them: thus, each customer would be enabled to state the following: "I am the owner of data produced by me - my small data belongs to me", "I authorize people, services and applications to view or process all or selected portions of my data", "I let my data being reused when I see my personal or common (social) benefit". A crucial challenge is therefore represented by the need of finding technological solutions able to reduce the burden of data access settings configuration: the system is required to be smart enough to streamline the personal data management experience. For instance, by recommending services relevant to the specific user, or providing adaptive and personalized interfaces.

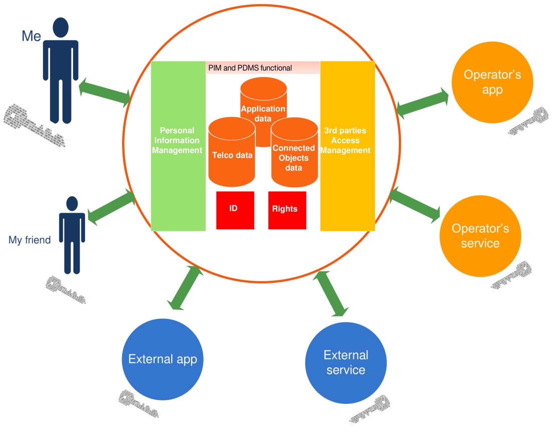


Figure 2. The PIM user-centric data control model.

A. PIM Functionalities

Our goal is to provide each customer with a secure solution that takes care of privacy of his data, extracts knowledge from this data and lets him/her control usage of both his data and the derived knowledge. An important success factor is linked to the customers awareness of the benefits that can derive from their data and appropriate knowledge modeling. Ideally, both the customers and the Telco operators will in the end benefit from this approach.

The most adopted online social networks currently offer centralized infrastructures in the Web [27]. These solutions are said vertical, i.e. they serve users' data only in scope of the same social network. The main disadvantage is the lack of clear rules of personal data and information management: in fact, a user is not aware of how his data is stored and processed, and of how it is used and managed.

Taking into account all of the above, this work is dedicated to elaboration of Personal Information Management, devised as a component of a social communication services platform and as an enabler for social communication services. Previously, we have dealt with the ways of enabling social Telco applications based on customer behavior and relations between customers [28]; in this paper, we focus on personal data aspects.

The PIM lets customers manage their personal data and grants them control over them, and includes a sub-system able to derive knowledge from data. In order to properly collect, store, maintain and transform data, two functional layers were designed, as represented in Figure 3:

- a logical layer, the Personal Information Management, responsible for data processing, knowledge extraction, and management of privacy policies;
- a physical layer called Personal Data Management, which performs data maintenance and operations (e.g. collection, storage, retrieval, physical security) on raw Telco data.

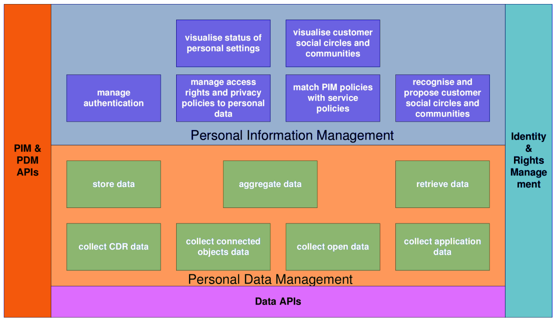

PIM &
PDM
APIs
visualise status of personal settings
visualise customer social circles and communities
manage authentication
manage access rights and privacy policies to personal data
match PIM policies with service policies
recognise and propose customer social circles and communities
Personal Information Management
store data
aggregate data
retrieve data
collect CDR data
collect connected objects data
collect open data
collect application data
Personal Data Management
Data APIs
Identity
&
Rights
Manage
ment

Figure 3. The two-layers architecture of PIM.

An important feature of the sub-system is concerned with ensuring safety of users' data. Personal Information Management functionalities being an integral part of social communication services platform (Social Tool for Telecommunication), are exposed to the external world through a set of APIs. The current implementation provides a set of 22 APIs used to e.g. register a service, retrieve service information, subscription.

In brief, the functional requirements of PIM are:

- to allow user to manage personal data with circles of contacts or communities which govern with own rules of information gathering, sharing etc.;
- to offer smart solutions with a simple interface and let customer manage his/her personal data (e.g. configure privacy settings, check them, confirm system prompts of settings based on user profiling, access visualizations);
- to offer control of different privacy levels based on user data and knowledge derived via customer dynamic profile (i.e. adaptive to the user's evolving interests and needs), social cartography and context (e.g. user relations, usages, location, availability, mood);
- to achieve interoperability among different telco operators' systems.

Usually, a significant part of online social activities consists of sharing of data, information, content which to a great extent has a personal mark. Sharing is in most cases happens through an application allowing the user to expose data to the external world (e.g. friends, classmates, co-workers, students, different groups and communities of people). The subject of sharing can be any kind of customer data (collected or created by him or bound with him by default – e.g. name, address). Specific novel kind of data is related to sensor readings which can include highly personal data (as exemplified by the Quantified Self movement), as well as enviromental and situational measures, which can also be exploited for profiling a particular person's behavior.

Currently, the first implementation of the Personal Information Management system offers relatively simple functionalities concerned with personal data sharing, via different applications to different people belonging to the customer's social circles and communities. Ongoing work will soon allow the introduction of more intelligence to the management of personal information, along with advanced functionalities. Present functionalities of PIM allow:

- to manage authentication of a customer allowing for entering well identified customer into a system;
- to control customer social network visibility;
- to match PIM policies with specific service policies;
- to let a customer set up access rights and privacy policies to his assets (like data, photos, devices) via several services;
- to manage access to customer's data, photos and devices and policies in the context of services and people (e.g. customer contacts' rights within particular services);
- to recognize and propose different social circles and communities for sharing;
- to visualize the status of customer's data access settings;

B. Algorithms, methods and tools for Personal Information Management

For the purpose of introducing intelligence into PIM platform several algorithms, methods and tools should be applied. The customer perspective fits the "small data" perspective elaborated in [29]; nonetheless, Big data methodologies have a significant role for e.g. inferring a customer's background and habits.

In order to provide a customer with personalized and contextual services that can be used on platforms like PIM the raw data needs to be transformed and preprocessed. Basic telecom data is available mostly in form of logs with specific Ids connected to the most important characteristics of a network event (e.g. place, user, service). Data safety and validity is achieved thanks to state of the art algorithms; then, multiple methodologies are adopted to transition from raw call detail records to aggregated data, focusing on different aspects of the logs themselves.

Several algorithms are in the process of being integrated into the platform. These include methods for dynamic behavioral profiling, anomaly detection, social action prediction, and community discovery. By working in the customer's context, such information will be used to provide privacy protection features: for instance the customer will be notified in case of anomalous data-sharing through a given service. One of the algorithms extended, integrated in the platform, and currently under validation is EVABCD [30] for dynamic behavioral profiling. Other methods adopted concern detection of communities based on social strenght of relations, which allow to differentiate contacts between users.

An algorithm that determines home and workplace of a customer based on CDR data can be seen as privacy threatening: nonetheless, a naive solution might consist in labeling such information with unique IDs – disconnected from the users personal data; this way, it is possible to derive higher level knowledge on the customer base while limiting the privacy issue. Following this example, the higher level information derived (where people live and work) can be used in:

Big Data perspective - obtaining information about certain user groups (e.g. identifying students based on POI visits) can give valuable information considering school transportation and providing authorities with suggestion of pedestrian crossings or speed limits on most visited areas.

Small Data perspective - information itself can be used to contextually provide users with offers of meeting people with similar habits and profiles that have agreed for it. For example people living or working in similar areas both use social app focused on certain hobby and receive suggestion of contact. This can be further expanded on with social network analysis providing "friend of a friend" features implemented in Telco network without the need for logging in external apps. Provided with this data it is then aggregated and that there are no ways to identify a given customer, preprocessed data like this aims to represent a real person as much as possible to draw conclusions regarding population based on it.

Many studies have concluded that the standard call logs are a good proxy to infer population characteristics; furthermore, aggregating the data provides us with data sets that can be of a high value for the analysis of:

- Population mobility - traffic, carbon dioxide emission;
- City centered analysis - urban planning [31], use of public transport and its optimization, characteristics of city districts and cross district movement;
- Tourism – identifying tourist movement models and prediction, identifying nationality focused "heatmaps" (e.g. where users from country X mostly visit);

From the academic point of view, several research fields focus on creating models of human behavior, spatial habits or social networks. In these studies, data is used differently from the standard "value oriented" approach (i.e. when user gives his personal data in exchange for personalized advertisements or special offers). This creates opportunities for the user to decide for what purpose his data is analyzed. The social aspect of influencing global research for which data are inaccessible by other means could show users the social benefit of their data preprocessing. On the other hand having easily understandable user information can open new ways of sharing user personal data based on the context. That way user can be more in control of his data, wishing to e.g. share information only when he is at work or only during the time he is travelling between places (for instance, if he wishes to participate only in a transportation study). In this example, the user could be in control not only of the type of the data that is collected, why it is gathered but also when the collection takes place. Such user-centered approach, providing increased transparency on data usage, can encourage users that otherwise would not share their personal logs or other data to do so.

IV. A Use Case: Social Gardening

In this section, we present a first use case implemented using the PIM platform, with the goal of illustrating the platform functionalities: *Social Gardening*, a service to support collective gardening in micro-communities (e.g. between friends, neighbors, or people who share a common interest). Social Gardening is a data-based service: it uses data coming from connected objects (sensors) deployed in a shared garden, information on users location, contacts of people etc., in order to streamline the shared garden management for the users that collectively take care of it. Personal data are thus required: the PIM platform enables the development of these kinds of apps, allowing the reuse of these resources in the service and the sharing of data with selected people for specific purposes.

Thanks to an application, usable from a mobile, a tablet or a computer, gardeners share data provided by connected objects to help take care about plants through social activities. The community can also share content relative to the garden.

The PIM represents the distinguishing feature for the Social Gardening application, and manages information and content sharing in a consistent user-centric fashion.

Figure 4 shows its interface.

Figure 4. The Social Gardening front-end.

A. Use Case Description

The main functionalities offered by the Social Gardening app can be summarized as follows:

- Real time view of the shared garden: see who is available in the garden, connected objects present, etc.

- Information sharing: know the vegetables' needs (data coming from connected objects equipped with sensors) in real time, check weather forecasts, etc.
- Alert/Messaging: receive alerts and notifications from sensors, find volunteers to take care of the garden, etc.
- Socialize: share content (photos, etc.) and news with co-gardeners, organize social events in the garden, etc.

The gardens managed by the application are real world gardens, where people can grow fruits, vegetables, and flowers in a collective fashion. This use case allows us to bridge the world of mobile phones to the emerging scenarios made possible by the increasing commercial availability of sensors, i.e. the Internet of Things. The application is designed to leverage data coming from the sensors the garden is equipped with (cameras, weather stations, humidity and hydration sensors, etc.). Thanks to these sensors, gardeners can keep track of what is happening in the garden and the micro-community can self-organize to obtain the best results: for instance, if the plants need water those in vicinity can take direct action and communicate in real-time with their peers. Moreover, data reuse from other communities of *social gardeners* can allow for the spread of best practices. Functionalities of the Social Gardening app are summarized in Figure 5.

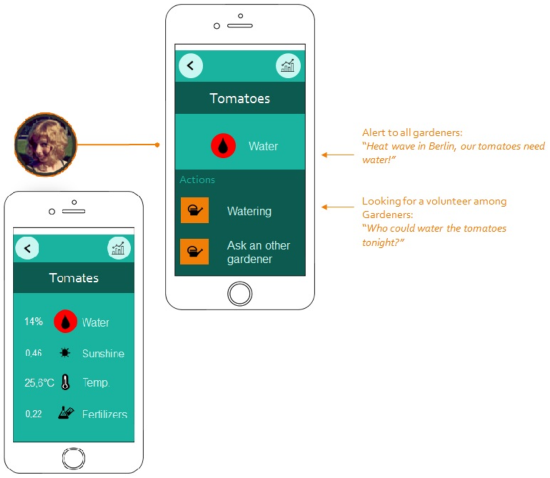

Tomatoes
Water
Actions
Watering
Ask an other gardener
Alert to all gardeners:
"Heat wave in Berlin, our tomatoes need water!"
Looking for a volunteer among Gardeners:
"Who could water the tomatoes tonight?"
Tomates
14%
Water
0,46
Sunshine
25,6°C
Temp.
0,22
Fertilizers

Figure 5. Social Gardening service functionalities.

A use case that can illustrate the need for such service can be the following: Emma, Charles and Léon live in Paris and cultivate together a shared garden in the 20th district. Emma grows tomatoes, Charles some apple trees while Léon takes care of flowers there. When Emma goes for her two-week holidays to London, she can monitor via the Social Gardening application what happens in the garden. For instance, she notices that it is boiling hot in Paris and receives a notification that her tomatoes need to be watered. In this case, she can send a group request for help to his co-gardeners but she can also monitor in real-time who is in the garden (Léon for instance) and ask him directly to take action.

B. User Studies

We ran user studies dedicated to the PIM platform and the Social Gardening service proposal, further evaluations are planned for the next future. Major objectives of the studies were:

- to learn the opinions of potential users on the concepts of new services;
- to determine the factors encouraging people to use the presented services and the possible barriers as to their use;
- to assess the appeal and uniqueness of the concepts;
- to identify potential benefits from the services and the possibilities for their application;
- to gather clues for further work on the services.

The studies were performed in qualitative mode as focus group interviews (FGI), in two sessions. The concept sequence was rotated at each interview; results are summarized below.

A sample of approximately 40 individual customers, divided in two groups, all of Polish nationality, accepted to participate in the user studies; they were selected according to the following criteria: aged 25 to 50; at least with secondary education; regular smartphone/tablet users; interested in technology and open to novel applications; active social media users (often sharing their location, or activities); users of technology for both practical and entertainment reasons; participating in the cultivation of allotment gardens (or house-adjacent gardens). The sample was balanced in terms of gender.

Regarding Social Gardening results can be summarized as:

- the service very well addresses the need for quick and easy access to information about the condition of the plants grown by the user;
- it makes it possible to better organize one's free time (easier decision about the need to go to the garden) or to decide to ask a friend to take care of the plants grown by the user;
- users are worried about the cost of sensors and the risk of them being stolen from the area of an unmonitored garden.

Regarding Personal Information Management can be summarized as:

- for the first group, the service seemed not to evoke major interest, mostly because of limited awareness on privacy issues; they maintained that they did not need a service to manage the sharing of their information; however, after additional questions it turned out that in some situations it was an important aspect for them (e.g. sharing the pictures of their children, HR checking information about them, sharing sensitive pictures via Snapchat;
- the second group of people perceived Personal Information Management quite positively and confirmed interest in this kind of service (offering adjustment of customer

settings with a dedicated interface to Personal Information Management); it must be noted that PIM version presented during this iteration was more mature;
- interviewers noted that while presenting the idea of the service, special attention needs to be paid and emphasis needs to be put on the customers' ability to manage those privacy aspects which are the most sensitive to them.

Interestingly, the results of the user studies about PIM reveal that several respondents seem not too excited about the possibility of managing and having control on their data. Digging deeper into the interviews, though, it becomes evident that the great majority of our respondents were not aware of the reasons why they should care: in fact, as interviewers exposed exemplar situations in which personal data might be misused, they appeared to realize the importance of the problem of privacy and data control. These results seem to point at a lack of information and awareness, rather than a lack of interest.

To summarize, platforms like PIM can only be adopted once at least two conditions are met: 1) users are aware of privacy and data control issues; 2) the platform should provide a streamlined and smart data management solution.

V. Insights to potential business models

Taking into account all the elements presented, in this section we evaluate the business alternative for the introduction of the proposed paradigm into the market: how to monetize the possibility given to people to manage their personal data?

The first possibility, the most classical and obvious that we can call "the Telco option", is to introduce it as a complementary service side-linked with the access subscription. Telcos are already selling access to data networks, hence we can easily imagine that they could introduce additional options to present and secure the personal data. Nonetheless, apart from the feasibility of introducing such a service into the provisioning/billing information system, the Telco will bear the full cost and generate no obvious new value. The willingness to accept an increase by X percent of the price of subscription by the customers is quite uncertain. It can also be assumed that not all the Telcos will deploy the solution without some specific incentive to do it. Such incentives could come in form of pressure from the regulators and policy makers, but this remains to be seen. Moreover, specific applications justifying such a new investment should devised.

Conversely, the second possibility that we can call "the Startup option" consists into growing a specific service that will build a dedicated infrastructure and sell it at a certain price with the promise of simplicity of usage and safety of the data. Once again, the customer is expected to pay some specific cost in exchange for proper and substantial services. In this case, the key success factor from the customer point of view should be further evaluated. From the network side, i.e. Telcos and OTTs, the solution must be seamless and provided through public APIs in order to get a universal impact. We may assume that if any possibility to do it appears, the market will provide many different solutions.

Several other possibilities exist between these two extreme ones. For instance, we could discuss a mixed model, that could be labeled as "the Vertical option". In this model, a specific solution dedicated to one specific activity or interest, such as Health, Books or Banks is developed and provided by a startup, and syndicated by a professional organization, leading to a mixed model:

- a specific offer with a price paid by the final customer;
- a cost shared by the professional using the solution and more or less visible on their invoices to the final customer.

These three examples show that the business model spectrum for PIM is wide and open. We may hence conclude that it provides an incentive to the authors to propose a truly cooperative approach to set up the technical solution, its evaluation, and explore the value created for the different potential stakeholders.

VI. Limitations and Future Developments

The PIM platform presented in this paper aims at providing a technological facility to *break the data silos*, i.e. to allow the emergence of novel services in a horizontal fashion while establishing the customers' rights of ownership and control over the data they produce. In the current and preliminary implementation, data encryption is managed by the Telco operator – hence, the goal of providing customers with full control over their data is only partially met. Several solutions, such as the aforementioned Enigma [26], are in active development and may be considered for integration into the platform to fully achieve this goal.

Future developments of this work will focus on the extension of current functionalities in accordance to the insights derived from the user studies presented. In particular, much attention will be devoted to devise strategies for raising awareness of customers, and to the development and integration of machine learning capabilities needed to ease the burden of managing data access for the end users. Further user studies will adopt a mixed (quantitative and qualitative) methodology to investigate customers' acceptance of the alternative solutions proposed.

VII. Conclusion

In this paper, we described a first implementation of the Personal Information Management platform and discussed its motivations; we reported on a first sample service built on top of it, leveraging connected sensors jointly with mobile phone data, and discussed potential business plans centered on PIM as well as its limitations and further development.

The PIM aims at effectively breaking the data silos currently present in the Telco industry, allowing the deployment of novel services of societal and business value while providing customers with full control over the usage of their data. The presented solution, hence, represents a first step moving from vertical to horizontal solutions *within* a Telco operator, while granting its customers with rights and power on the data they produce. One goal of this paper is also to start a wider discussion *between* Telco operators on the possibilities opened by allowing customers to manage their own data: indeed, a solution like PIM can have a great impact in the Telco ecosystem if this vision is shared among operators. To this end, we plan to put efforts into establishing cooperation with other interested operators, in order to produce a first reference implementation to be adopted by more than one operator. In case of success, Emma, Charles and Léon (in the sample use case above) will be able to use and leverage the same service even in the case they are customers of different Telco companies.

References

[1] P. Russom *et al.*, "Big data analytics," *TDWI Best Practices Report, Fourth Quarter*, 2011.

[2] C. Zhao, Y. Wu, and H. Gao, "Study on Knowledge Acquisition of the Telecom Customers' Consuming Behaviour Based on Data Mining," in *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on*. IEEE, 2008, pp. 1–5.

[3] E. Nidelkou, M. Papadogiorgaki, B. Bratu, M. Ribiere, and S. Waddington, "User Profile Modeling and Learning," 2009.

[4] R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Clustering anonymized mobile call detail records to find usage groups."

[5] V. Frias-Martinez and J. Virseda, "On the relationship between socio-economic factors and cell phone usage," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. ACM, 2012, pp. 76–84.

[6] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Social computing, behavioral-cultural modeling and prediction*. Springer, 2013, pp. 48–55.

[7] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, and I. C. M. Simulation, "Human Mobility Modeling at Metropolitan Scales 2 . Spatial and Temporal Parameters for Mobility Modeling," *Acm*, pp. 239–251, 2012.

[8] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation*, vol. 42, pp. 597–623, 2015. [Online]. Available: http://link.springer.com/10.1007/s11116-015-9598-x

[9] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, "Friends don't lie: inferring personality traits from social network structure," in *ACM UbiComp*, 2012.

[10] M. C. González, C. a. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature06958

[11] A. Sevtsuk and C. Ratti, "Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks," *Journal of Urban Technology*, vol. 17, no. 1, pp. 41–60, 2010. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10630731003597322

[12] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10630731003597306{\#}.VaoNNaTtmko

[13] S. Isaacman, R. Becker, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People ' s Lives from Cellular Network Data 1 Introduction," *Pervasive Computing*, vol. 6696, pp. 133–151, 2011.

[14] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools, "Building a validation measure for activity-based transportation models based on mobile phone data," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6174–6189, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.03.054

[15] D. Maldeniya, S. Lokanathan, S. Lanka, A. Kumarage, and S. Lanka, "Origin-Destination Matrix Estimation for Sri Lanka Using 2 . the Four Step Model," no. May, pp. 785–794, 2015.

[16] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki, and C. Ratti, "Activity-Aware Map : Identifying human daily activity pattern using mobile phone data," *Human Behavior Understanding*, vol. 6219, pp. 14–25, 2010. [Online]. Available: http://www.springerlink.com/index/JJ21508881433584.pdf

[17] M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez, "Exploring the potential of phone call data to characterize the relationship between social network and travel behavior," *Transportation*, vol. 42, no. 4, pp. 647–668, 2015. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84930821078{\&}partnerID=40{\&}md5=c06377dc8f1b54834cc86cb75222807e

[18] D. Pastor-Escuredo, A. Morales-Guzmán, Y. Torres-Fernández, J.-M. Bauer, A. Wadhwa, C. Castro-Correa, L. Romanoff, J. G. Lee, A. Rutherford, V. Frias-Martinez *et al.*, "Flooding through the lens of mobile phone activity," in *Global Humanitarian Technology Conference (GHTC), 2014 IEEE*. IEEE, 2014, pp. 279–286.

[19] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki, "Prediction of Human Emergency Behavior and Their Mobility Following Large-scale Disaster," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 5–14. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623628

[20] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the Impact of Human Mobility on Malaria," *Science*, vol. 338, no. 6104, pp. 267–270, 2012. [Online]. Available: http://www.sciencemag.org/content/338/6104/267.abstract

[21] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez, "An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, Oct 2011, pp. 57–64.

[22] A. Bogomolov, B. Lepri, J. Staiano, E. Letouzé, N. Oliver, F. Pianesi, and A. Pentland, "Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics," *Big Data*, vol. 3, no. 3, pp. 148–158, 2015.

[23] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira, "Your browsing behavior for a big mac: Economics of personal information online," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 189–200.

[24] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe, "Money Walks: A Human-centric Study on the Economics of Personal Mobile Data," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: ACM, 2014, pp. 583–594. [Online]. Available: http://doi.acm.org/10.1145/2632048.2632074

[25] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, "openPDS: Protecting the Privacy of Metadata through SafeAnswers." *PLoS ONE*, vol. 9, no. 7, 2014.

[26] G. Zyskind, O. Nathan, and A. Pentland, "Decentralizing Privacy: Using Blockchain to Protect Personal Data," in *Security and Privacy Workshops (SPW), 2015 IEEE*. IEEE, 2015, pp. 180–184.

[27] R. Boutaba, "What's next on online social networking?" in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015.

[28] A. Filipowska, M. Mucha, and B. Perkowski, "Towards social telco applications based on the user behaviour and relations between users," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015, pp. 95–102.

[29] D. Estrin, "Small data, where n = me," *Commun. ACM*, vol. 57, no. 4, pp. 32–34, Apr. 2014. [Online]. Available: http://doi.acm.org/10.1145/2580944

[30] J. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Creating Evolving User Behavior Profiles Automatically," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 5, pp. 854–867, May 2012.

[31] M. De Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri, "The death and life of great italian cities: A mobile phone data perspective," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.

5.3 Towards Social Telco Applications Based on the User Behaviour and Relations Between Users

The goal of the section and the paper is to propose an approach benefiting from Call Detail Records to enable creation of socially-empowered and personalised applications. The applications should benefit from detailed user profiles, including behavioural profiles, to enrich user experience and sustain privacy policies. Such understanding of profiling and personalisation is in line with the restrictive law on telecommunication as well as recent General Data Protection Regulation [28]. The section contributes to the second of the detailed goals of the chapter: ”Proposing methods enabling for user profiling based on Call Detail Records data”.

The paper was accepted to 18th Conference on Innovations in Clouds, Internet and Networks, ICIN 2015, Paris, 17-19.02.2015 and published in the conference proceedings. The detailed bibliographic reference is as follows: Filipowska, A., Mucha, M., Perkowski, B., Szczekocka, E., Gromada, J., Konarski, A., 2015, Towards Social Telco Applications Based on the User Behaviour and Relations Between Users in: ICIN 2015: 18th International Conference on Intelligence in Next Generation Networks, IEEE, pp. 95-102.

Towards Social Telco Applications Based on the User Behaviour and Relations Between Users

Agata Filipowska, Michał Mucha, Bartosz Perkowski
Department of Information Systems
Poznan University of Economics
Poznan, Poland
agata.filipowska@kie.ue.poznan.pl

Ewelina Szczekocka, Justyna Gromada,
Adam Konarski
Orange Labs Poland
Orange Polska S.A.
Warsaw, Poland
ewelina.szczekocka@orange.com

***Abstract*—The paper presents an approach for creation of social and personalised applications that benefit from the Call Detail Records (CDR) data stream. These applications analysing the user-behaviour based on CDRs propose functionalities that enrich the user experience at the same time sustaining the privacy policies. The paper introduces My Social Connector application developed at Orange that applies the proposed paradigm.**

Keywords*—*user profiling, personalisation, Call Detail Records, behavioural modelling, Telco Social Graph

I. Introduction

Nowadays, a number of solutions, not only in the Telco industry, address challenges such as Big Data processing or personalisation of services to satisfy the needs of users and the individual user in particular.

It is important to differentiate between analytical solutions that use Big Data and the solutions that make the Big Data small, changing the perspective towards the single user's point of view. In this paper we focus on the second kind of systems. Our goal is to re-interpret the data in a way meaningful for a single user, extract its social dimensions and use it to supply socially-enriched applications. These applications may become personal assistants of a user e.g. supporting his social actions, recommending, searching, inviting, and sharing data etc. with a circle of his contacts and contacts of their contacts.

The challenge that appears in this case is a full privacy protection and data anonymisation (to a certain extent depending on the application scenario and privacy settings). The privacy is an important aspect of using the personal data. Different systems take care of a user consent at the same time processing the data to the full extent possible. And as it is concerned with Internet and mobile applications, users may know what can happen to their data by reading the privacy policy. After users agree to make their data available, they however may not be aware how and which data is used and where this data is processed. It is even difficult to control deleting the data from the system of the application provider.

The goal of this paper is to propose an approach benefiting from the Call Detail Records (CDR) data stream that enables creation of social and personalised applications. These applications analysing the user-behaviour based on CDRs propose functionalities that enrich the user experience at the same time sustaining the privacy policies.

The paper is structured as follows: the next section is devoted to the related work in modelling user relations, user profiles and privacy in information systems dealing with user data. It is followed by a description of approaches considered in the process of building evolving user profiles. After that, the concept of contextual services based on weak ties and rich access to user information is introduced. Finally, an example implementation is described, along with discussion on the privacy of user data in such systems.

II. Related work

A. Modelling relations between users

Many different disciplines of science define the notion of a relation (or a tie) between two objects. This applies also to the social networking (also computationally supported), where a tie is understood as a kind of interaction schema between two or more people. This relation may be tangible or intangible or may relate to a specific situation. Nowadays, one may obtain these relations inter alia based on the CDRs.

[1] defines the notion of a tie strength defining it as a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal services [1]. A tie between two individuals may be strong, weak or absent (even when two individuals know each other). The strength of a tie has various implications e.g. the stronger the tie, the more time these individuals spend together; the stronger the tie, the more similar the individuals are. The strong ties also support diffusion of information as close people tend to communicate more, however paradoxically weak ties may lead to integration into communities (strong ties lead to fragmentation) [1]. [2] confirm this assumption showing that people tend to disclose more personal information to someone in a strong relationship. [3] presents other implications showing that with individuals connected by strong ties we share our values, tastes and interests (this may be different for people living in the cities).

The weak ties, on the contrary, provide access to new information.

There are numerous approaches to quantify the strength of a relation, even for a graph retrieved from CDRs. The basic approaches encompass: sum of contacts, average of all contacts of an individual, weighted average of sum of all of mutual contact, etc. These measures take into account the data that is available. [2] proposed two other metrics (in this case for a Twitter-derived social network): the chain frequency (measuring the number of conversational chains between the dyad averaged per month) and the chain length (measuring the length of conversational chains averaged per month). The higher the values for both these metrics, the stronger the relationship is. [4] present other dimensions that need to be taken into account while quantifying the strength: intimacy, intensity, duration, social distance, services, emotional support and structure of a relation. The method that will be applied in a certain scenario depends on the application problem as well as the underlying data model.

B. *Modelling user behaviour based on CDRs*

Modelling the user behaviour is a vast field of research, increasing in complexity with the development of Big Data storage and processing capabilities. A great framework for modelling behaviour is provided in [5]. The paper is a result of a profound work on behaviour modelling, providing a structure for both the process of handling behaviour in informatics, as well as for modelling behaviour. A more specific path is taken in [6], where the authors show that Call Detail Records can be successfully used to infer personality traits.

In [7], the social network data is used to determine "activity centres" of users, which turn out to indicate where a user works and lives. Moreover, through semantic categorization of visited places, user profiles are enhanced with rule-based types. [8] propose a useful method for extracting mobility patterns from a perspective of an individual. The patterns of routine travels are a part of the profile and serve to match users with each other.

The authors of [9] approach the problem of dynamic computer user behaviour modelling. Behaviour is represented as sequences of actions, in a constantly Evolving Fuzzy Model suitable for Big Data applications. Merging CDR and Social Network data sets is explored in [10], with promising results.

[11] describe finding strongly related communities in data scrapped from the early-day Facebook database. The groups are then used to infer user profile attributes, based on homophily. A similar pursuit in [12] uses homophily-based inference to validate user-provided profile attributes, with good results.

Additionally, less related to the scope of research of our team, works such as [13]–[15] show the possibilities of content-based modelling of user tendencies and behaviour.

The important task of modelling the user behaviour can be accomplished with CDR datasets. Moreover, the relations between users also appear in such datasets. Together, these two form a foundation of our work.

III. Towards personalized and context-aware services

An unavoidable consequence of the widespread adaptation of “smart” devices, as well as routine use of digital services, is the ubiquity of data traces that are left by users and consumers. The data traces take many forms – clicks, shares, likes, calls [10], messages, check-ins [7], transactions, etc.

Consumer choices show that in most cases convenience is valued higher than privacy. The development of the Internet infrastructure, advances in processing power and improvement of data storage capacity facilitated a race to create systems that allow the most efficient personalized marketing strategies [14].

The recent revelations about the size of a cyber-espionage inclined many to care about privacy. In that light, a new vision of utilizing personal data becomes relevant. Viewing consumer data as the property of individual consumers, business companies can extend tools that allow consumers to receive beneficial information from their own data.

For that purpose, extensive user behaviour models can be built with the consent of users. The models do not need to consider any limits, because the end user will determine what data she wants to supply.

The richer the profile, the more accurate is the information about the user’s context. With more information, more possibilities to offer beneficial services arise [11].

User profiles are understood here as sets of attributes. The attributes are variables retrieved from mining data traces for information and patterns. The variables can be structured into five groups:

1. Statistical. Considering the example of Call Detail Records, such variables may range from the average call duration for a certain day, to the breakdown of how many times were certain Base Transceiver Stations used during a certain period. A wide variety of such measures can be calculated [6], meaning that particular applications of profiles determine the choice of variables.
2. Graph-based. Representing social networks as graphs turned out to provide valuable insight, through more complex measures such as betweenness centrality, to the simple ones such as a node degree [6]. To view the user in a social context, these variables are very important.
3. Sequential. Behaviour, seen as actions, is naturally forming sequences [5]. Noticing patterns in the sequences has shown to be a way of discerning among groups of users. Examples are sequences of calls and messages, visits at locations, etc.
4. Location-based. The physical presence of a user is one of the strongest and most important parts of what can be understood as a context. The variables could range from “activity areas”, through distance between users, to routine paths [7].
5. Content-based. This is the most prevalent form of profiling variables in networks such as Twitter, where the greatest amount of data is in the content of

messages. The variables can be sets of terms used, along with the frequency of usage, etc. [13]

The variables belonging to the listed categories may be used to build models that evaluate a vast variety of insights about the users. From judging psychological traits to determining types of relations, the pool of available applications is really broad.

When profiles are considered as part of an automated system that provides useful information to users who decide to allow access to their data, the question of maintenance appears. Since everything is continuously changing, the models need to evolve as well in accordance with new data. The information system running a model that regularly evaluates user profiles needs be able to update the values of variables, and reflect the changes in the context of applications and solutions [9].

User profiles that are possible to build in the described manner can serve as a component of a greater solution. For example, mixed with insights coming from a graph of social relations that allows the discovery of weak ties, this component forms a foundation for context-oriented services.

IV. Taking advantage of the social relations emerging from CDRs

One of the main features that social networks are based on is the strength of relationships between particular people, which describes how strongly people are connected with each other. Having that knowledge, a rich API can be provided to creators of social applications, recommendations algorithms, etc.

Telco Social Graph (TSG) is a component of Orange Ecosystem that connects Orange customers in a social network based on the data they introduce in the network through their mutual communications. The role of the TSG is to:

- collect data describing Orange clients and their behaviour from various sources of the Orange ecosystem,
- transform this data into ordered and connected information about users and their environment, and discover the social information (about Orange customers) such as:
    - How strongly are people connected with each other?
    - What kind of subgraphs/ subgroups do people create?
    - How can a particular person be described?
    - How do the relations change in time?
- Provide social information to social communication services and in this way enrich the Orange ecosystem with additional data.

TSG is not in disposal of an explicit declaration of relationships between people in contrary to e.g. Facebook. This knowledge needs to be measured based on the raw data gathered. According to [1], the relationships between Orange customers are divided into three types: absent ties, weak ties (acquaintances), strong ties (family & close friends). Granovetter pointed out the following factors that should be taken into account when calculating strength of ties:

- amount of time spent in relationship,
- the emotional intensity,
- the intimacy,
- the reciprocal services which characterize each tie.

When applying the above parameters to the data gathered and processed by the TSG, the following information has been identified to be used for relationships strength calculation:

- the amount of time that two Orange customers spend on communication with each other,
- communication channels used for this communication: calls, SMS, MMS, emails, other services. Some communication channels indicate more emotional intensity and intimacy than others (e.g. calls),
- a direction of the communication,
- similarities that can be found between particular users, e.g. the applications, services that they both use, places they visit, etc.,
- a duration of relationships – it is more probable that long lasting relationship is intimate and emotional,
- a way in which the relationship evolves in time – some deviations in strong and long lasting relationships do not affect the relationship but can ruin a weak relationship.

The TSG currently implements a simplified version of an algorithm for calculating the strength of relationships between particular Orange customers. A complex solution taking into account all presented factors is under development. The API provided to social services creators, is mostly based on this calculated value. The API is parameterized, so that the developers decide what relationships do they want to be retrieved from TSG. In many cases, strong relationships are interesting from their point of view, but according to [1], weak ties are the ones that are potentially the most meaningful.

In case of services built on the foundation of TSG, the

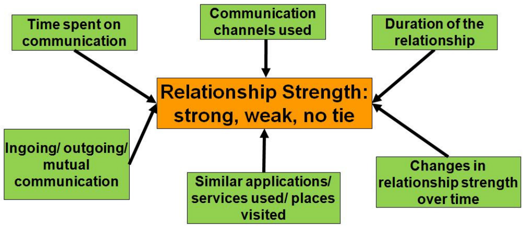

Time spent on communication
Communication channels used
Duration of the relationship
Relationship Strength: strong, weak, no tie
Ingoing/ outgoing/ mutual communication
Similar applications/ services used/ places visited
Changes in relationship strength over time

Fig. 1. Factors influencing strength of relationships between Orange customers in the TSG. Source: own study

following scenario is an example of the advantage that can be reaped from utilizing a method based on weak ties coupled together with a rich profile and contextual information:

- An Orange customer Bob communicates a lot via calls, SMS, MMS and emails.
- The TSG analyses Bob's data and divides people he contacts with into three categories: strong, weak and some contacts are not important at all so there is no tie marked here. Most of his business contacts are marked as weak.
- Bob goes far away from his usual place, like abroad for business affairs, e.g. to some local unit in big multi-country enterprise. He has some business relations with people right there while he contacts them on an every day basis (weak ties).
- Being abroad, Bob should contact local managers or people cooperating with his "business friends" from this location. For instance to receive additional opportunities of cooperation or explanation of local specifics that influences former results of cooperation and that will help him to better adjust his solution to a local environment, etc.
- A personalized address book (a social service built on the TSG foundation) helps Bob to reach people he does not know yet. It displays his strong contacts but also weak contacts from the present location (current context) and contacts of these weak contacts when their profiles indicate that contacting them might be useful in this situation. Thanks to that, Bob can automatically see e.g. his local managers (he may already know them) but also e.g. local legal unit or people working on specific project locally, that he wants to cooperate with in his premises.

The TSG provides various APIs based on relationships strength that can be used by services developers. It is presented in more details in the next section.

V. Telco Social Graph: applications making use of the Telco data

A. *Algorithms and methods*

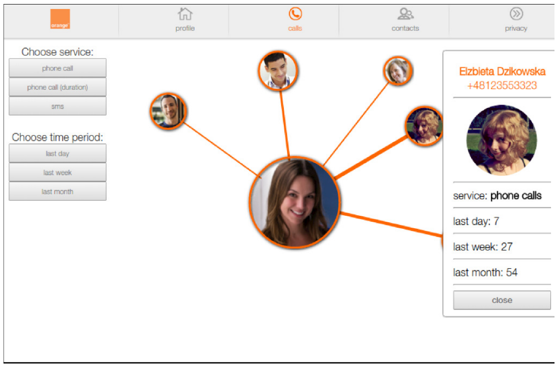
Many real-world data, e.g. telecommunication networks or social networks, can be represented as graphs that can be analyzed further to explore the properties of said networks. These properties are mostly a statistical evaluation of characteristics that many networks have in common. One of such properties is related to the presence of communities [16]. A community is defined as a subset of nodes within the graph, such that connections between these nodes are denser than connections with the rest of the network [17]. The communities can be identified by graph-based clustering methods. The two most popular approaches for identifying communities within a network are: graph partitioning and modularity scoring. Graph partitioning approach uses methods that partition different nodes into groups that share common features, however it produces communities that do not overlap. On the other hand, modularity-based algorithms propose a cluster derived from the topological structure of the network and finally use a modularity score optimization to produce high quality communities.

One of the most popular approaches that uses modularity to measure the quality of the network is implemented by Newman and Girvan. The CNM algorithm is a bottom-up agglomerative clustering which continuously finds and merges pairs of clusters trying to maximize the modularity score [18]. The method focuses not on removing the edges between the pairs of nodes with the lower similarity, but on finding edges with the highest betweenness. Betweenness can be described as a measure that favours edges that lie between communities and disfavours those that lie inside communities. For example, if two communities are connected by only a few edges, then all paths between the nodes from one community to the other must pass through one of those edges. Authors distinguish few types of betweenness measures, which can be useful in defining the strong and week ties between nodes. The first one, *shortest-path betweenness*, is based on the number of all shortest paths that lie between each pair of nodes. In other words, we can think of signals that travel through the network from the source to the destination. However, the authors assumed that the signal does not have to travel along geodesic paths, but can perform a random walk. For this purpose, the *random-walk betweenness* measure was described, which calculates the expected number of times that a random walk between particular pair of nodes will pass down a particular edge and sums over all pairs of nodes. Once the edges with the highest values of betweeness are identified, the proposed method removes these edges and recalculates betweenness in order to form new communities. When the communities are identified, it is important to know which of these divisions the best for a given network are. This is where the modularity measure is used, which measures the internal connectivity of the community.

Betweenness measure is also used in splitting communities through the graph theory approach. It is one of the algorithms implemented to identify articulation points in the graph, which are crucial to communication. We define an articulation point when all paths between certain nodes pass through this point. Removal of this point causes an increase in a number of biconnected components [19].

Described methods are applied to single-labelled and unweighted social networking datasets. In case of an overlapping structure of the network, the fuzzy clustering over the weighted undirected graph can be implemented. This technique leverages the weights on the edges and tries to validate *bridgeness* by using multi-labelled data. For the optimal fuzzy clusters for the given graph, the bridgeness measure quantifies the degree to which a given vertex is shared among different clusters. However, in order to identify the true bridges in a network, this measure has to be paired with the degree measure [16].

The identification of communities in graphs creates a challenge of how to evaluate the intensity of relations that bind users, and how they facilitate communication and the spread of information. These aspects have been extensively studied in



Elzbieta Dzikowska
+48123553323

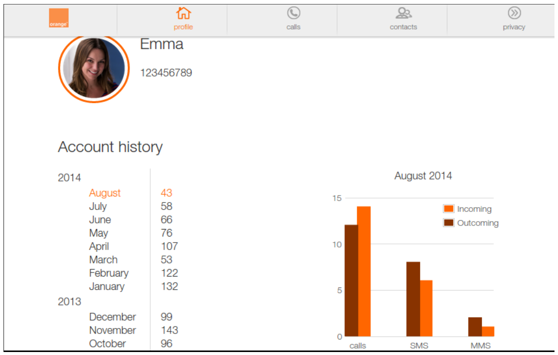
service: phone calls

last day: 7

last week: 27

last month: 54

Fig. 2. My Social Connector - the graph of user's contacts. Source: screenshot



Emma

123456789

Account history

2014		
	August	43
	July	58
	June	66
	May	76
	April	107
	March	53
	February	122
	January	132
2013		
	December	99
	November	143
	October	96

Fig. 3. My Social Connector – user's profile. Source: screenshot

social sciences, under the framework of the strength of weak ties theory [1]. This theory can be extended to online social networks, suggesting the use of information on user interactions for the purpose of predicting the tie strength. This requires the knowledge about the topology of a social network. In [20] authors describe few methods for the identification of weak ties. A *shortcut bridge* can be defined as a link that connects a pair of nodes, whose deletion would cause an increase of the distance between them. The method of identifying shortcut bridges uses the shortest-path betweenness described earlier. However, to avoid the computational challenges, to each edge of the network a value of strength is assigned, strength that defines the distance between two nodes not in the number of hops required to connect these nodes, but as the cost of the lightest path connecting them. Weak ties can also be defined as links that connect pairs of nodes belonging to different communities. The important characteristic of weak ties is that bridges create more, and shorter, paths. As described earlier, an articulation point was the key player in the graph that connects many users, however, from a community structure perspective, the deletion of a weak tie would be more disruptive. For this reason weak ties are proven to be very effective in the diffusion of information.

B. *My Social Connector*

The methods and algorithms used to describe structural properties of the networks, users and their relations are important not only for statistical purposes, but also helpful in providing a tool to support users' calls management. The results can be adapted to present to the user call's log in a legible and understandable manner. Telco operators try to satisfy their subscribers' needs with online and mobile applications that simplify the management of their data, especially related to the service usage, connections, billing data and personal data. The proposed solution called My Social Connector is a responsive Web application developed to visualize data about user's connections and support social activities. With a constantly growing popularity of mobile devices, the application has to support various screen resolutions and provide a unified layout on any Web browser. The user's data is visualized in the application in the form of charts and graphs. To supply these forms of visualization, we use an API that returns (from a neo4j database) the user's data packed in JSON objects. In order to fulfil these requirements, an application was created based on HTML5 and JavaScript. The decision was made based on the popularity of HTML5 (with CSS3), which supports many visual effects and allows to personalize the appearance of each element on the website. The selection of JavaScript was dictated by the need of parsing data returned from neo4j and libraries used to generate graphs and charts in the application, that are basically written in JavaScript.

The application consists of four tabs: profile, calls, contacts and privacy. The last one is strictly related to the privacy settings required for the personal data processing. In Fig. 3 the user is presented with general statistics about his usage of services divided monthly. The main screen is divided into two parts, with the selection of month on the left, for which we want do display the statistics. As a result, a chart of services used is shown, which are aggregated by the type of a service and distinguished between incoming and outgoing.

The second module, named "calls", provides the visualization of the most popular user contacts. The user can manage his/her relations with each contact based on the type of services used for communication.

The relations with all contacts are visualized in the form a graph as it is shown in Fig. 2.

In order to build a dynamically changing graph of user contacts, the application has to be supplied with specific data aggregates. Each contact stores the information about the quantity of phone calls and short text messages, the average duration of phone calls and the percentage of connections with a certain user in relation to all connections. Having such data prepared, we can apply various modifications to the graph based on this data, e.g. we can change the structure of the graph based on the service type used. Once we select a service type, the graph is generated and displayed with some specific characteristics. The size of each contact and the distance between that contact and the user is calculated based on the percentage of connections. The higher percentage of connections, the larger is the image of a contact, and the smaller the distance. Other attributes of contacts are related with the thickness of each connection, e.g. the higher number of phone calls, the thickest stroke that connects the two nodes. In order to ensure the readability of the graph, each value in the

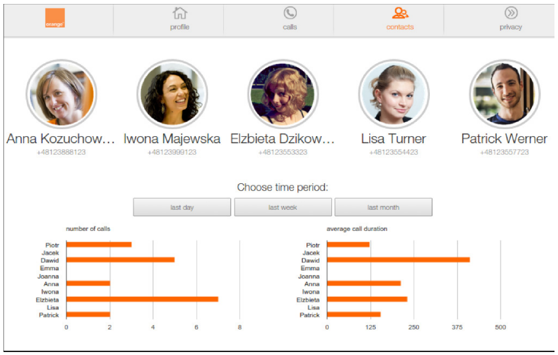


Fig. 4. My Social Connector - overall statistics of all user contacts. Source: screenshot

graph is normalized through calculating the logarithm of the value. For all types of services, the user can analyze his relations based on three periods of time: last day, last week and last month. For each, the graph dynamically changes the whole structure by adding more contacts (if occurred) and by redrawing the thickness of the edges. Finally, the user can display the detailed information for a specific contact just by clicking on the image.

The third module – contacts - displays the summary of all contacts and provides the possibility to compare all contacts with just one click. The module is split into two parts as shown in Fig. 4. The bottom part stores charts that allow a quick comparison of services used for connecting with each contact. The data can be dynamically changed according to the time period selected. The top part of the module consists of all user contacts and their online statuses. When the contact is online, the user can open a chat box and send a message through the application.

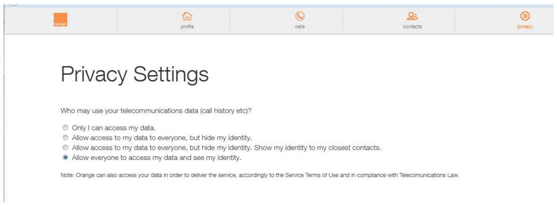
The application shows the potential in providing users with their personal data in a graphic form. The user can manage contacts by himself, and gather information about the most popular contacts and services used. Moreover, as it is developed with a responsive layout, it can be launched not only on a desktop browser, but also on a tablet or a smartphone. The visualizations and the usage of mobile devices bring the user account management to a higher level and provide users with a sense of greater possibilities and confidence in contacts and relations management.

C. *Privacy of My Social Connector*

Main assumptions for the opt-in privacy management system for My Social Connector relate to the four levels of access to private data that are proposed:

Level 1 - User does not allow using information on his communication activities. In this case, his data will not be processed and presented in form of a graph of his connections.

Level 2 - User allows using information on his communication activities and contacts, but anonymously, i.e. with hiding his identity. It implies for instance in case of an application supporting a personalised search taking into account relations of a user in the social network that his connection to a certain person performing the search will be taken into account while delivering search results, but his identity will not be returned from the graph and will not be available for a person performing the search.



Privacy Settings

Who may use your telecommunications data (call history etc)?

○ Only I can access my data.
○ Allow access to my data to everyone, but hide my identity.
○ Allow access to my data to everyone, but hide my identity. Show my identity to my closest contacts.
● Allow everyone to access my data and see my identity.

Note: Orange can also access your data in order to deliver the service, accordingly to the Service Terms of Use and in compliance with Telecomunications Law.

Fig. 5. My Social Connector – the privacy panel. Source: screenshot

Level 3 - User allows the use of his communication data by the application, as well as showing his identity to close contacts.

Level 4 - User allows the use of his communication data by the application and full public access to his identity.

Furthermore, it is possible to define access level 1, 2 or 3 for groups (defined by the user), and separately for every application.

In My Social Connector, the user can click the checkboxes with different access levels. My Social Connector can therefore serve as an interface for the personal data management system.

In different applications accessed by the user, during the installation, there will be a notice about the utilization of the user's data, together with request to express consent before installing the app.

Fig. 5 presents the screen that allows the user to express his consent and mark one of the four possible levels of privacy

VI. DISCUSSION ON PRIVACY

One of the goals of our work is to enable the users to benefit from their personal data. However, we need to remember that dealing with personal data brings a lot of challenges.

One challenge is that people lose their trust and are not willing anymore to allow companies to use their data. There is a challenge in convincing users that in fact they are the persons who have the power over their data and to make them willing to share their data. Secondly, we need to ensure people that their data is well protected. We also have to attract people by showing them what they can earn by giving some of their personal data to be processed (in a secure and trusted manner) by an operator like Orange.

A major step to solve these challenges is to provide a system that includes the personal data management (on a physical level) and personal information management (on a logical level). In order to reach this goal and build personal data management and personal information management systems, we defined several requirements:

1. Mandatory: we should use opt-in choice to allow users to maintain their data, i.e. perform different actions registered by the system with visible results for the user, like accept data, display data, modify data, delete data. A user should be aware how his/her data will be used.
2. Mandatory: a user should have the possibility to maintain different levels of privacy for the data – i.e. depending on his/her circles of communication (different groups of persons he/she is related with) e.g. different for family, friends, work-/classmates, colleagues, known people, unknown people etc.
3. Mandatory: for communication purposes the Webcom (Orange proposal) framework should be used opt-in could be introduced through additional libraries made available to any third-party services.
4. Optional: personal data management system can involve user profiles in order to enable identifying circles of communication, and automatic identification of affiliation of a user (to a particular group).

It is an initial set of requirements that is going to be further detailed.

We also plan to extend the application and TSG methods towards a social context management system driven by users' profiles. The aim of this profiling is not segmenting customers (for instance for marketing offers), but rather to allow users to build their own personal social profiles, which are going to help them by their social activities. Moreover, these profiles should be dynamic (based on changing user behaviour and customs) not static, i.e. declared ones. Only with such profiles we will be able to deliver appropriate information, at the right time and the right place.

We envision the following phases of building a Personal Data Management system:

1. Simple functionality (screen) for confirming an agreement of a user that his/her data can be used for improving the service with a feature of displaying own data of a user:
    a. Screen for confirmation of data allowance,
    b. Feature of displaying data,
    c. Information how the data will be used (purpose).
2. Providing elementary functionalities like building a user profile.
3. Providing specific functionalities of Personal Data Management module.
4. Providing Personal Data Management operational system – advanced maintenance of user data in real time.

Our next step is to enhance the personal data management and personal information management systems. Then, an extensive customer social profile will be modelled, in order to be used dynamically in a specific social context. Basic examples of usage of a social profile addressing the user needs will be provided as well, as we are convinced that a real value can be delivered to the customer.

VII. Conclusions

This paper presented an approach for processing the data concerning user activities performed using the mobile phone and available in CDRs. This data is processed bearing in mind all privacy aspects that have to be applied by each Telco provider.

The challenges that relate to the development of similar applications are diverse. Technical ones relate to processing the data stream, integrating the data with the data that may be found on the Web and analysing the data to provide an added value for a single user. Business issues concern finding a proper business model that may make the application survive in a very competitive environment.

The research presented in this paper concerned the Telco Social Graph (TSG) - a component of the Orange ecosystem that connects Orange customers in a social network based on the data they introduce in the network through their mutual communications.

The paper also presented an overview of the My Social Connector application showing how the data may be applied to enrich the user experience. The application will now be tested with users who may influence the interfaces and methods. However, the current version of the application shows the directions one may use the data and therefore provides inspirations for users.

References

[1] M. S. Granovetter, "Granovetter - 1973 - The Strength of Weak Ties," *Am. J. Sociol.*, vol. 78, pp. 1360–1380, 1973.

[2] J. Y. Bak, S. Kim, and A. Oh, "Self-disclosure and relationship strength in Twitter conversations," pp. 60–64, Jul. 2012.

[3] E. Gilbert, "Predicting tie strength in a new medium," *Proc. ACM 2012 Conf. Comput. Support. Coop. Work - CSCW '12*, p. 1047, 2012.

[4] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *ACM Conference on Human Factors in Computing Systems*, 2009, pp. 211–220.

[5] L. Cao, "In-depth Behavior Understanding and Use: The Behavior Informatics Approach," *Inf. Sci.*, vol. 180, no. 17, pp. 3067–3085, Sep. 2010.

[6] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. de Vicuña, and N. Oliver, "Towards a Psychographic User Model from Mobile Phone Usage," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 2191–2196.

[7] Y. Qu and J. Zhang, "Trade Area Analysis Using User Generated Mobile Location Data," in *Proceedings of the 22Nd International Conference on World Wide Web*, 2013, pp. 1053–1064.

[8] R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti, "Mining Mobility User Profiles for Car Pooling," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1190–1198.

[9] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Creating Evolving User Behavior Profiles Automatically," *Knowl. Data Eng. IEEE Trans.*, vol. 24, no. 5, pp. 854–867, May 2012.

[10] A. Cecaj, M. Mamei, and N. Bicocchi, "Re-identification of anonymized CDR datasets using social network data," in *Pervasive Computing and*

Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on, 2014, pp. 237–242.

[11] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, “You Are Who You Know: Inferring User Profiles in Online Social Networks,” in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 251–260.

[12] S.-H. Park, S.-Y. Huh, W. Oh, and S. P. Han, “A Social Network-based Inference Model for Validating Customer Profile Data,” *MIS Q.*, vol. 36, no. 4, pp. 1217–1237, Dec. 2012.

[13] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing User Modeling on Twitter for Personalized News Recommendations,” in *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, 2011, pp. 1–12.

[14] E. Aktolga, A. Jain, and E. Velipasaoglu, “Building Rich User Search Queries Profiles,” in *User Modeling, Adaptation, and Personalization*, vol. 7899, S. Carberry, S. Weibelzahl, A. Micarelli, and G. Semeraro, Eds. Springer Berlin Heidelberg, 2013, pp. 254–266.

[15] J. Min and G. J. F. Jones, “Building user interest profiles from wikipedia clusters,” 2011.

[16] T. Saha, C. Domeniconi, and H. Rangwala, “Detection of Communities and Bridges in Weighted Networks,” in *Machine Learning and Data Mining in Pattern Recognition*, vol. 6871, P. Perner, Ed. Springer Berlin Heidelberg, 2011, pp. 584–598.

[17] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, “Defining and identifying communities in networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 2658–2663, 2004.

[18] M. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E*, 2004.

[19] M. Saravanan, G. Prasad, S. Karishma, and D. Suganthi, “Analyzing and labeling telecom communities using structural properties,” *Social Network Analysis and Mining*, vol. 1. pp. 271–286, 2011.

[20] E. Ferrara, P. De Meo, G. Fiumara, and A. Provetti, “On Facebook, most ties are weak,” *Communications of the ACM, in press*, 2014. [Online]. Available: http://www.academia.edu/850096/On_Facebook_most_ties_are_weak. [Accessed: 21-Oct-2014].

5.4 Conclusions

This chapter contributes to the part of the thesis aiming at showing benefits from application of profiling in vertical domains, with a focus on public utility companies. The goal of the chapter was to "develop a profile of a customer/subscriber to telecommunication services, enabling for personalisation and taking into account issues of privacy and trust". This goal was further translated into two secondary goals addressed by specific sections of this chapter, namely:

G4.1 Defining a solution that enables to manage personal information in telecommunication (targeted in Section 5.2).

G4.2 Proposing methods enabling for user profiling based on Call Detail Records data (covered by Section 5.3).

To address the goal G4.1, the paper dealt with the following challenge: proposing a set of methods for managing personal information to enable new application scenarios. The supplementary goal was to empower the user to update and manage his/her personal data. The paper included in Section 5.2 provides a description of the Personal Information Management platform and depicts methods, algorithms and tools allowing different usages based on a social context of a user. The Social Gardening application is presented to demonstrate a potential application of the developed concept. The research goal targeted in the paper was similar to the one addressed by the EGO - Virtual Identity project, however the focus was on proposing new services benefiting from a user profile, taking into account the aspects of privacy and trust.

The goal G4.2 was further translated into proposing an approach benefiting from Call Detail Records to enable creation of socially-empowered and personalised applications. The applications should benefit from detailed user profiles, including behavioural profiles, to enrich user experience and sustain privacy policies.

The paper presented in Section 5.3 presents insights to what extent CDR data may be useful to derive profiles of users, including relations these users have with their community. The statistics that may describe users based on CDR data are studied. The paper proposes also a concept of the Social Connector application that focuses on presenting the history of contacts of a subscriber (to propose actions or understand the relations with other users). An important part of the paper is devoted to the privacy issues. The privacy levels regarding the telecommunication data are proposed and discussed (from no data sharing towards all data sharing approach).

This chapter concludes the part on utilisation of profiling for public utilities. Chapters 6 and 7 cover horizontal applications that may be used in diversity of domains.

Chapter 6

Profiling for Authentication

6.1 Introduction

6.1.1 Motivation

Profiling has a great potential not only for a specific domain, but also in cross-domain scenarios e.g. for authentication. Especially, when research results show that at least 30% (on average about 50%) of devices are unprotected by any means of security (no PIN or password set) [42, 60, 162], profiling might provide a solution. The lack of protection is usually due to the following issues: user needs to remember many passwords, methods require user interaction and traditional biometrical sensors are uncommon and inaccurate. On top of that we should remember that current systems allow only for the binary authorization and don't allow setting privileges when sharing a device.

In this research we would like to exploit the fact that a user may be distinguished and authorised based not only on a password, but also based on his/her behaviour that greatly differs. We addressed the issue of authentication by indicating how much data on user behaviour is needed to properly profile a user for the needs of authentication.

The goal of the chapter is to develop a method enabling for authentication of a user based on Call Detail Record data. Call Detail Records concern billing data of a telecommunication provider that covers every action a user made using his mobile phone.

To achieve this goal, the following secondary goals were defined:

- Verifying, if the Call Detail Record data is sufficient for detecting anomalies in the behavioural user profile and therefore enable for applying the CDR-based profile in authentication scenarios.

- Researching how much data describing a user is needed to provide an efficient authentication solution.
- Developing and validating a method for description of a user for the needs of authentication. The supplementary goal is to develop a methodology for testing the behaviour-based approaches based on the Call Detail Records data.

6.1.2 Structure of the Chapter

The chapter consists of four sections including an introduction presenting relation to goals of the thesis and a summary that presents results that were achieved in relation to these goals. Section 6.2 refers to first of the goals mentioned, as well as proves how much data describing a user is needed to provide an efficient authentication solution. Section 6.3 focuses on development of a profile sufficient for the needs of authentication.

6.2 Application of Trajectory Based Models for Continuous Behavioural User Authentication through Anomaly Detection

The goals of the section are twofold: on one hand it is to prove that it is possible to develop an authentication solution based on CDR data, and on the other to provide a method based on user mobility patterns for the needs of anomaly detection and authentication using the Call Detail Records. This paper researches therefore two issues: sufficiency of Call Detail Record data for the needs of anomaly detection and amount of data needed to provide an efficient authentication solution.

The paper was presented and published at the Conference on the Scientific Analysis of Mobile Phone Datasets, 5-7.04.2017 Vodafone Theatre, Milan, Italy. Detailed bibliographic reference is as follows: Kałużny, P., Jankowiak, P., Filipowska, A., Abramowicz, W., 2017, Application of trajectory based models for continuous behavioural user authentication through anomaly detection, NetMob 2017: Book of Abstracts. Oral., pp. 92-94.

Application of trajectory based models for continuous behavioural user authentication through anomaly detection

Piotr Kałużny, Piotr Jankowiak, Agata Filipowska, Witold Abramowicz[1]

[1]Department of Information Systems, Poznan University of Economics and Business, Poland

***Abstract*—This paper describes how mobility patterns understood as trajectory based models can be applied for an anomaly detection and authentication using telecommunication data. The trajectory based mobility model utilizing stay-point extraction suited for the sparse CDR data is used to describe mobility patterns of a user. In this model, the observed activities are assigned with anomaly scores in three distinctive areas including: geographical, sequential and temporal dimensions. Activities with threat values exceeding the user confidence threshold are identified as anomalies. The model is tested on the sample of Poznan inhabitants. Evaluation of the model performance is based on the similarity classes and the results are presented within the paper.**

I. Introduction

In recent years, due to the ubiquity of cell phones and raise in their penetration rates, mobile phones became not only a basic tool for everyday use, but also a valuable source of information about their users. The value of data generated by those devices has risen significantly over the years [6].

Regardless of value of this data as perceived by a user, mobile devices suffer from a lack of proper protection against unwanted access to this data in case of a theft (of both the device or user identity). Proper authentication techniques and automatic systems are needed to ensure those devices are secure from theft and an unintended use. The traditional methods have their drawbacks caused mostly by users' negligence resulting in 40% of the phones not secured by any means [7]. Due to that fact and drawbacks of point-of-entry traditional approaches, currently existing methods are not enough [1]: sometimes lacking the required security, availability or usability. As a possible answer to this problem a new family of methods is introduced, named behavioural biometry. These methods address unique, non transferable, difficult to reproduce and hard to forget or loose characteristics derived from user behaviour rather than physical traits. The methods can rely on various factors and focus on different aspects of behaviour to find unique patterns of e.g. gait, signature or keystroke dynamics. Those patterns, created while the device is running, can be used to secure it. This allows the authentication process to work continuously and transparently without the user interaction needed. The use of those methods can provide an additional layer of security on top of existing methods without diminishing the usability, e.g. PIN or password would be used only when the behaviour analysis system is not sure about the user's identity. Within those methods there exists a subgroup, referred to as behavioural profiling, which: *identifies people based upon the way in which they interact with services of their mobile device* [13]. The user's identity is determined based upon the comparison of a sample of activities with his profile. If the sample matches the profile, the user will be granted with an access, otherwise he will be refused [17] or an additional proof of identity will have to be brought (e.g. PIN).

Studying behavioural patterns and especially the user mobility had proven to be quite successful in differentiating between users and identifying deviations from a user profile. This is possible due to the fact that cell phone traces closely resemble user trajectories, regardless of the mobility data source being CDR or phone-collected data. Also observations of the human movement conducted by the researchers confirmed the stability of those patterns [9]. The predictability of mobility patterns was proven to be high and stable given historical behaviour of a user [15], [18], [19] regardless of the daily distance travelled [3]. The proxy of BTS (Base Transceiver Station) labelled geographical information, derived mostly from CDR, was proven to be precise enough to study human mobility. Its applications included identifying patterns on a large scale confirming correlation with e.g. population density [2] or transport networks. Those traces also remained precise enough to capture mobility patterns to allow for an individual user analysis based on visited locations and travel models [4], [5], [14], also introducing methods for a better cell dwell time prediction [16], [21]. The users' profiles mostly utilized the semi-structured patterns that can be observed when analysing mobility in a weekly manner in hourly bins [8] even in more frequently generated phone data [12]. Such models can be a source of features for behavioural profiling approach for the anomaly detection model that utilizes multiple methods based on: probability of visiting a location at a given time [22] or sequential characteristics [20] of movement. Those methods gave highly satisfying results on frequent, phone generated data [7].

II. Dataset and approach

Our dataset is the database of Orange covering about 4 million of anonymized users over six months (from February to July) in the 2013. For each record we are given the following information: an anonymized user identification number, a

BTS id which are grouped to obtain locations covering distinct geographical areas, and a timestamp at the initial moment of the phone activity.

The mobility model that we propose utilizes the user focused stay point extraction model mentioned above, with a few assumptions:

- Activities in the same location separated by less than an hour are considered to contain enough information to assume that a user has been in the location during the period.
- Activities which have a *stay time*[1] larger than 30 minutes or are separated by more than 4 hours of inactivity[2] are labeled as stays. Stay locations are the places where users engage in some activity (contrary to the ”pass-by” locations).
- All this information is kept in a weekly calendar of a user - a structure divided into hourly bins (timeframes).

The model outcome - profile is treated as a pattern, which is a base for comparison of the activities loaded to detect potential anomalies and frauds. As a result of that comparison, threat scores (which are measures between 0 and 1) are assigned to each activity in three dimensions. They define how much each tested activity varies from a user pattern in the following areas: geography - each activity is tested against the user geographical profile where the geography of a location is compared with a distance to the closest location from a set of important locations[3] compared to the user daily travel range[4]. Time - each activity is assigned a threat metric based on the distance in (hourly) timeframes, when a user is present at a given location and the usual time he is at the location divided by 24 hours[5]. Sequence - by using the trajectories built upon the extracted stays, all of the passed-by BTS are used to construct a mobility trie (TrieRoute) that contains information about frequencies of stations visited by a user when travelling between point A and B on the learning data along with the order. Each activity is assigned a sequence threat depending on the probability of a given point appearing in a sequence.

After defining those measures we conducted an experiment choosing the inhabitants of a Poznan area which is shown in the Figure 1. Home locations were extracted from our model as the longest stay between 7 p.m. and 7 a.m.[6].

The following approach was applied on the data: firstly, the mobility model was built on 24 days of data from March 2013, then 7 day verification period was used to generate a typical threat level for a user. This was used to test how consistent the users are with their patterns, generating threats in three above mentioned perspectives. 90th percentile of those threat values was used to create user confidence interval for each of the target threats. Each tested activity having its threat level above the threshold value of the confidence interval is classified as an anomaly in this dimension. Next, based on those upper thresholds for every user, test data from 14 days of April was loaded to test how the model performed in the authentication scenario, evaluating whether the threats generated by other users' data can be used to differentiate between the profiled user and an impostor.

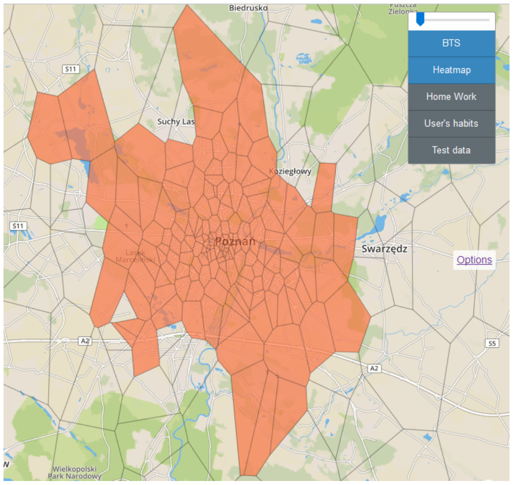


Figure 1. Poznan area - visualization made in Javascript (Mapbox, D3.js, jQuery) to showcase Voronoi cells of BTSs from the chosen area.

Due to the fact that for the evaluation of mobility authentication models, a random user case is not a valid use case scenario[7], we introduce a class based verification model that utilizes similarity classes. This idea can be compared to simulation of an uninformed and informed attacker case from the security domain (seen also in behavioural biometry cases e.g. [12]). Therefore, the test data is divided into five classes, listed considering the expected raising similarity to the base user and difficulty of the model adjustment: random user, user living in the same town, user having home in the same BTS, user having same home and work locations in respective BTS and finally the same user - with the data loaded from the testing period. Based on this comparison a model is run on the user base, without any prior filtering besides two aspects: all users from respective classes need to be found for the base user and each of the test class users needs to have at least 5 activities in the testing period.

III. Results

The experiment was run on a sample of 1000 users, for which we were able to find corresponding users in all of test classes. We excluded users with unstable mobility patterns or sparsity of the data within CDRs. This fraction accounted for about 0.8% of the sample. The average level of threat generated was highly dependent on the test class as it is shown

[1] The difference between a departure and an arrival time from a location based on actions.

[2] To avoid very long movement sequences spanning over multiple days due to the sparse activity data and the uncertainty period of this sparsity [16].

[3] Places that a user visits that comprise at least 5% of the model extracted stay time in any hourly bin during the period of comparison.

[4] The approach is based on [10].

[5] Which is a max. time distance at which threat equals one.

[6] Described in [11].

[7] Model detecting anomalies for a random user appearing in a different part of country may achieve high accuracy but be practically unusable.

in the Figure 2. This proved that such division is useful in evaluating method accuracy.

Next, an anomaly/impostor scenario was prepared. Each loaded set of consecutive activities (an activity window of 3 activities was used[8]) was classified as an anomaly or normal user behavior. The anomaly was detected when the activities threat values exceeded the user confidence threshold in at least two dimensions (as it was tested to have the best anomaly detection accuracy). Additionally, an algorithm iteratively decreasing the threat percentile to achieve best results was used to improve the accuracy of the method. The average fractions of properly detected anomalies over the sample were as follows:

- 98,97% for a random test user class,
- 91,1% for a test user living in the same town as a base user class,
- 53,32% for a test user having the same home location as a base user class,
- 31,84% for a test user having the same home and work location as a base user class.

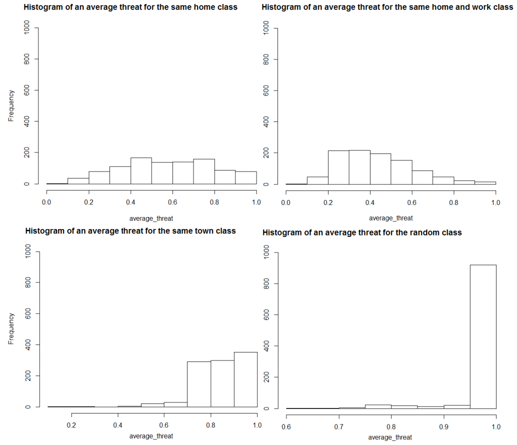


Figure 2. The distribution of average threat levels (considering three threats mentioned in the model) for all of the users.

The false rejection rate, interpreted as a portion of situations where valid user data from the test period was classified as an anomaly (compared to a profile), averaged to 13,71% over our sample. Presented results show that our model has a very high accuracy when applying a testing methodology used in the literature (random user) - about 99%. This showcases the usability of the model performance even on a sparse dataset like the event based CDR logs. However, we found that when comparing a user with another user having a very similar profile (living or working in similar areas), the accuracy is much lower. This indicates that a potential theft of a mobile phone by a thief, who has a similar mobility behaviour profile as the victim may be significantly harder to detect. The division in testing classes also introduces a new methodology for evaluation of methods' performance in both the authentication scenario (e.g. phone theft) and user pattern differentiation (distinguishing between patterns of similar users like e.g. family members).

References

[1] Beyond the password: The future of account security. https://www.telesign.com/wp-content/uploads/2016/06/Telesign-Report-Beyond-the-Password-June-2016-1.pdf. Accessed: 2016-09-10.

[2] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[3] J. P. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[4] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.

[5] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*, (2526):126–135, 2015.

[6] B. Fox, R. van den Dam, and R. Shockley. Analytics: Real-world use of big data in telecommunications. *IBM Institute for Business Value*, 2013.

[7] L. Fridman, S. Weber, R. Greenstadt, and M. Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. 2015.

[8] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behavior. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 550–555, 2013.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[10] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in Los Angeles and New York. *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011*, pages 88–93, 2011.

[11] P. Jankowiak and P. Kaluzny. Human mobility profiling based on Call Detail Records analysis. Bachelor thesis, Poznan University of Economics and Business, Poznan, 2015.

[12] H. G. Kayacik, M. Just, L. Baillie, D. Aspinall, and N. Micallef. Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. *arXiv preprint arXiv:1410.7743*, 2014.

[13] F. Li, N. Clarke, M. Papadaki, and P. Dowland. Active authentication for mobile devices utilising behaviour profiling. *International journal of information security*, 13(3):229–244, 2014.

[14] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools. Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174–6189, 2014.

[15] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3, 2013.

[16] M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez. Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, 42(4):647–668, 2015.

[17] H. Saevanee, N. Clarke, S. Furnell, and V. Biscione. Continuous user authentication using multi-modal biometrics. *Computers & Security*, 53:234–246, 2015.

[18] C. M. Schneider, V. Belik, T. Couronne, Z. Smoreda, and M. C. Gonzalez. Unravelling Daily Human Mobility Motifs. *Journal of The Royal Society Interface*, 10(84):20130246(1–8), 2013.

[19] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[20] G. Tandon and P. K. Chan. Tracking user mobility to detect suspicious behavior. In *SDM*, pages 871–882. SIAM, 2009.

[21] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, 2015.

[22] S. Yazji, P. Scheuermann, R. P. Dick, G. Trajcevski, and R. Jin. Efficient location aware intrusion detection to protect mobile devices. *Personal and Ubiquitous Computing*, 18(1):143–162, 2014.

[8] Meaning an average threat value for three activities was used for classifying whether the data belongs to a user or an impostor.

6.3 Large Scale Mobility-based Behavioural Biometrics on the Example of the Trajectory-based Model for Anomaly Detection

The most significant of the secondary goals to be addressed in this chapter aims at developing and validating a method for description of a user (profiling) for the needs of authentication. In line with the secondary goal, the goal of the paper is to propose a working model of a behaviour based authentication applying anomaly detection performed over the user's mobility patterns. The supplementary goal, a side effect as defined by [57], is to create a methodology for testing similar, behaviour-based authentication approaches.

The paper was published in the Journal of Universal Computer Science[1]. Detailed bibliographic reference is as follows: Kałużny, P., Filipowska, A., 2018, Large Scale Mobility-based Behavioral Biometrics on the Example of the Trajectory-based Model for Anomaly Detection, Journal of Universal Computer Science, 24 (4), pp. 417-443.

[1] `http://www.jucs.org/`

Large scale mobility-based behavioral biometrics on the example of the trajectory-based model for anomaly detection

Piotr Kałużny, Agata Filipowska
Department of Information Systems
Faculty of Informatics and Electronic Economy
Poznan University of Economics and Business, Poland
{piotr.kaluzny, agata.filipowska}@ue.poznan.pl

Abstract: The paper describes an implementation of a behavioral authentication system, working on sparse geographical data generated by mobile devices in the form of CDR logs. While providing a review of state of the art w.r.t. sensors and measures that can be used when creating a system detecting anomalies in the user behavior, it also describes domain specific authorization methods focusing on the user mobility.

Thew trajectory based stay-extraction model is utilized to build user mobility patterns, upon which the anomaly detection model measures the repeatability of human behavior in dimensions of: geography, time and sequentiality. The goal is to measure the extent to which the geographical aspect of the human mobility can be used in behavioral biometrics systems i.e. in which scenarios geography may enable to describe (and differentiate between) user patterns – based on anomaly detection in cases resembling real life scenarios (phone theft or sharing between users). The research methods developed may be implemented on mobile devices to benefit from multiple sensors data in the authentication processes.

The model is evaluated on a large telecom dataset, with the use of similarity classes, what allows measuring the accuracy of the model in real-life scenarios and provides benchmarking guidelines for the future work on the topic.

1 Introduction

Nowadays mobile devices have become truly ubiquitous. Due to this fact, they became both a valuable source of information [16] and a concern to assure privacy of their owner's data. Due to reasons connected with a user's negligence, possibly caused by the usability barrier of the currently used authentication approaches [2], about 40% of mobile phones remain unprotected by any means [17]. The ease of use seems to be a significant factor in the adoption of new authentication methods [2, 39]. This enables behavioral biometrics to improve this process by utilizing multi-factor authentication and cover the drawbacks of the traditional authentication methods.

The goal of this paper is to propose a working model for behavior based authentication applying anomaly detection performed over

the user's mobility pattern. The issue was researched by application of the methodology described by Öesterle [28].

The structure of the paper is as follows. The description of the research problem is given in the introduction, analyzing issues of current authentication approaches and behavioral biometrics as a possible solution. Second chapter defines concepts used and provides an analysis of the literature - focusing on mobility as a base for behavioral authentication. The chapter also describes the advantages of the proposed approach over state of the art. In a third chapter a trajectory based model is described, and the anomaly detection method is presented along with a division of anomalies into dimensions of: geography, time, sequence and predictability. In the fourth chapter the model is evaluated on a large real world dataset. In the last chapter, the discussion on the comparability of the existing mobility-based behavioral authentication approaches is brought up along with some practical remarks. The future work is also discussed.

2 Related work

2.1 Traditional means of authentication

Traditional authentication factors have a few drawbacks. Among these drawbacks we may distinguish:

– **knowledge factors** represented by passwords, work as a point of entry mechanism which frustrates users [2] mostly due to the requirement of a user interaction and the issue of ”stacking up” [45, 8]. They are also often simple and easy to break.

– **possession factors** connected with token devices are a good choice for high security situations. Nonetheless, they are rarely used due to economical and usability reasons[1].

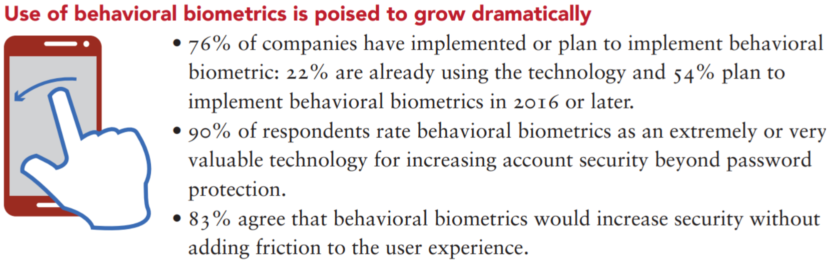
– **inherence factors** connected with traditional biometry, offer a family of high accuracy methods including fingerprint recognition or new examples of facial features biometrics. The main issue with these methods is that they are not available for all devices. Biometry adoption among the produced mobile phones achieved about 40%, but its penetration rates among companies and users are worse [1, 3]. These methods also can't work continuously due to the battery drain and/or characteristics of the methods used.

The family of traditional methods can be extended with the concept of **behavioral biometrics**. Behavioral biometrics includes a variety of methods, consisting of: gait [14], keystroke dynamics [40], voice recognition [27] and many

[1] Users are required to carry an additional device and interact with it to gain access. They also need not to lose or forget to take the device.

more. One of its fields covers the behavioral profiling, which tries to derive patterns from the user's behavior and interaction with a device, which are closely resembled by the data that is produced by the devices [4]. Behavioral profile model can consist of many aspects with a capture-able (quantifiable) regularity, where deviations from the observed behavior can lead to uncovering anomalies connected with a potential threat to user's data [10]. In some of those cases, domain specific algorithms can be used for capturing and comparing the patterns (e.g. voice recognition [31]).

This multi-aspect characteristics[2] allows for an **easy application of behavioral biometry models in multi-layer authentication**, widely adopted in tech companies [2]. Due to the fact, that those methods can be applied for a constant user authentication, they do not hinder the usability, while adding an additional layer of security. This makes the use of the behavioral system a good compliment to the password based or traditional biometric solutions (which do not work well in a multi-layer authentication [39]). These facts confirm a significant demand for the services among companies, as seen in Figure 1. Deriving insight from the behavioral patterns provides also information about the current context of the user behavior, which is important in domains where observation of a user is crucial e.g. patients, elderly people [15] in case of health care applications.



Use of behavioral biometrics is poised to grow dramatically

- 76% of companies have implemented or plan to implement behavioral biometric: 22% are already using the technology and 54% plan to implement behavioral biometrics in 2016 or later.
- 90% of respondents rate behavioral biometrics as an extremely or very valuable technology for increasing account security beyond password protection.
- 83% agree that behavioral biometrics would increase security without adding friction to the user experience.

Figure 1: Findings of a report on a potential adoption of the behavioral biometrics. Source: [2]

2.2 Behavioral profiling on mobile devices

The behavior (or behavioral) profiling is defined as it: *"identifies people based upon the way in which they interact with the services of their mobile device.*

[2] Meaning there can be multiple aspects of a behavioral profile which can be modeled with different methods and work in various scenarios.

In a behavior profiling system, user's current activities (e.g. dialing a telephone number) are compared with an existing profile (which is obtained from historical usage) by using a classification method (e.g. a Neural Network). The users identity is determined based upon the comparison result." [23]. The user's profile can include multiple aspects of his behavior [27]. Each of these aspects can be described by one or many measures (characteristics) that can be used for the pattern creation (presented in Figure 2). It is clearly visible, that multitude of these factors point to a non trivial tasks of pattern recognition. An exemplary aspect is mobility. Considering a range of user travels (geographical area) along with the sequence of visited cells (routes taken) and information connected with the repeatable nature of human behavior [19, 24], identifying patterns is not an easy task. Domain specific algorithms are required to create user's mobility behavioral profile and measure potential anomalies and deviations from these patterns.

2.3 Mobility models

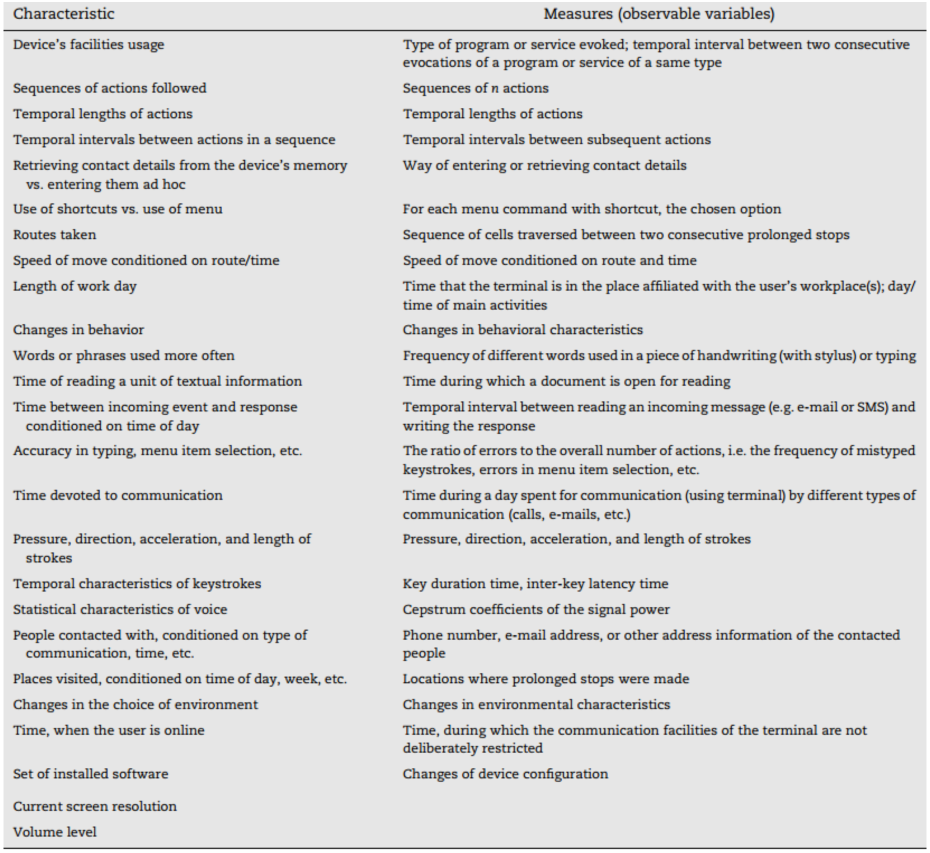
Use of data from various sensors connected with mobility and available on a device (GPS data, WiFi networks available or even IP address) is a broad field of study. In addition, the research around the usage of call logs (locally) or Call Detail Records (CDR) (on a server of a telco provider) is one of the most interesting areas due to the availability of this data on each phone. It was proven that geographical aspects of user whereabouts derived from CDR can be successfully used in modeling human mobility [21, 7].

Humans have stable mobility patterns and a significant tendency to return to a few often visited locations[3] [13, 19]. Despite the uncertainties, human mobility is predictable based on the historical behavior [34, 35, 25] regardless of the distance traveled [6]. Due to this fact, even using sparse data like CDRs we are able to get a good approximation of user movement patterns.

In case of using Call Detail Logs for the analysis of user mobility, only very brief moments of his whereabouts are known. They are related to calls or other services used by a user that were handled by BTS[4]. This estimation of a user's location is not ideal, but its accuracy can be measured based on the density of the towers. It proven to be sufficient to perform analysis of a human mobility on a small scale focusing on estimating temporal patterns of locations visited by a user and building a user's mobility profile [24, 12, 11]. This task can be performed by a family of trajectory-based methods (often relying on a stay-extraction) to estimate the dwell time in each place the user visits [43, 20, 42, 26]. The user's profile built mostly utilizes the semi-structured patterns that can be observed when analyzing the mobility in a weekly manner in hourly bins [29, 18, 5]. The

[3] Mostly identified as their home and workplace or their equivalents, like e.g. school.

[4] Base transceiver station.



Characteristic	Measures (observable variables)
Device's facilities usage	Type of program or service evoked; temporal interval between two consecutive evocations of a program or service of a same type
Sequences of actions followed	Sequences of *n* actions
Temporal lengths of actions	Temporal lengths of actions
Temporal intervals between actions in a sequence	Temporal intervals between subsequent actions
Retrieving contact details from the device's memory vs. entering them ad hoc	Way of entering or retrieving contact details
Use of shortcuts vs. use of menu	For each menu command with shortcut, the chosen option
Routes taken	Sequence of cells traversed between two consecutive prolonged stops
Speed of move conditioned on route/time	Speed of move conditioned on route and time
Length of work day	Time that the terminal is in the place affiliated with the user's workplace(s); day/ time of main activities
Changes in behavior	Changes in behavioral characteristics
Words or phrases used more often	Frequency of different words used in a piece of handwriting (with stylus) or typing
Time of reading a unit of textual information	Time during which a document is open for reading
Time between incoming event and response conditioned on time of day	Temporal interval between reading an incoming message (e.g. e-mail or SMS) and writing the response
Accuracy in typing, menu item selection, etc.	The ratio of errors to the overall number of actions, i.e. the frequency of mistyped keystrokes, errors in menu item selection, etc.
Time devoted to communication	Time during a day spent for communication (using terminal) by different types of communication (calls, e-mails, etc.)
Pressure, direction, acceleration, and length of strokes	Pressure, direction, acceleration, and length of strokes
Temporal characteristics of keystrokes	Key duration time, inter-key latency time
Statistical characteristics of voice	Cepstrum coefficients of the signal power
People contacted with, conditioned on type of communication, time, etc.	Phone number, e-mail address, or other address information of the contacted people
Places visited, conditioned on time of day, week, etc.	Locations where prolonged stops were made
Changes in the choice of environment	Changes in environmental characteristics
Time, when the user is online	Time, during which the communication facilities of the terminal are not deliberately restricted
Set of installed software	Changes of device configuration
Current screen resolution	
Volume level	

Figure 2: List of distinctive measures proposed by Mazhelis et al. for mobile masquerader detection. Source: [27]

trajectory-based models have the advantage of being an understandable representation of an approximated user mobility pattern and can have multiple uses in the analysis and uncovering human behavior patterns, in contrary to the often classification-heavy purpose of Machine Learning Algorithms (MLA). Nonetheless, due to some unpredictability of the human behavior mobility, pattern models require different learning periods depending on the data density and the task. They are also rather parameter heavy due to the fuzzy patterns users have - even having the perfectly sampled data, the upper threshold w.r.t. quality of prediction of user behavior is about 87% [33].

2.4 Anomaly detection in mobility

To detect anomalies in the user mobility patterns, a wide variety of methods can be applied. The basic methods are based on the Bayesian decision rule system to classify the conditional probabilities of visiting BTS stations and mean residence times [9]. Another family of methods focuses on the sequence of visited locations. An approach proposed by Sun et al. [36] built a Markov model, utilizing EWMA (Exponentially Weighted Moving Average) mobility tries, based on cells visited by the user. By building a probability-based model of the routes user followed, the model was able to detect anomalies in new sequences of locations that were unlikely - had a lower probability of the user's appearance than a design threshold (Pth). In this case also oscillations and errors in classification of locations should be considered [38]. A few recent methods described in Table 1 provide extensions of these basic approaches. These examples provide interesting insights and give a rough approximation on the expected accuracy of the model. Nonetheless, they work on well sampled and small datasets - their use on a large scale, real world and sparse datasets was not tested.

2.5 Advantage over state of the art

The approach proposed in the paper benefits from the findings on human mobility. The proposed model describing patterns of the user's mobility is created using trajectory based methods [43, 24, 20, 42, 26] and by clustering activities in weekly patterns with 1h discrete time windows [29, 18, 5]. The model considers characteristics of the sparse data and possible errors in the observed movement [38, 32]. The user's mobility profile is then used as a pattern for behavioral authentication based on anomaly detection, which utilizes a threshold method [37] based on 90th percentile of the normal behavior threat readings [46]. The model includes a novel approach based on the division of the mobility anomalies into different dimensions including: time, sequence (partially based on [36]) and a geographical area, along with the probability of a user visiting a given location. The proposed model is proved to be able to differentiate between the user patterns in a long term.

The paper, to the best of our knowledge, also presents the first large scale application of the mobility-based behavioral biometrics on sparse data (in this case CDR). The previous approaches focused on samples of: 76 [23], 100 [44] or 178 users [46]. This model was tested for 1000 users based on CDR logs. The respective test cases were chosen, based on the similarity metrics, from 252

[5] http://realitycommons.media.mit.edu/realitymining.html

[6] https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/

[7] http://realitycommons.media.mit.edu/realitymining.html

Table 1: Review of approaches for differentiation of user patterns, anomaly detection and authorization.

Publication	Dataset	Method used	Accuracy
Mobility-based anomaly detection in cellular mobile networks [37]	A simulated dataset showcasing a graph resembling the cellular mobile network. Call durations are the same for all users and exponentially distributed with a mean value of 3 minutes. The higher the mobility level, the more cells traversed with a given speed - set between 20 and 60 miles/hour for testing purposes.	High order Markov Model Exponentially Weighted Moving Average used for creating a profile - the probability of each route the user took. The design parameter Δ is based on the entropy of a current trace and is used for changing the detection threshold. Anomaly detection based on calculating the distance between the current trace and the EWMA-based mobility trie.	89% accuracy with 13% FRR
Mobi Watchdog: You Can Steal, But You Can't Run! [44]	Reality mining dataset[5] activities labeled with BTS cell id from 100 users, sampled every 30 minutes to showcase CDR granularity level. 30 days used to train the model and 30 to test the model performance.	HHMM (Hierarchical Hidden Markov Model). Decision is made after τ (design parameter) consecutive activities have been found anomalous (parameter in the model). Working authentication software raising alerts by requesting the device holder to reauthenticate himself when an observed mobility trace significantly deviates from the trained model.	Accuracy above 90%, for similar users between 50% and 70%. FRR about 13% for one anomalous activity window and 9% when using 3 activities.
Efficient location aware intrusion detection to protect mobile devices [46]	Geolife dataset[6] - GPS trajectories from 178 users with about 5 second sampling. Reality mining dataset - 68 users chosen with an average of 2.5 min sampling. 100 sample batches of x (5, 15, 30, 60 minutes) used for testing.	Trajectory based mobility model on frequently visited locations with 30 mins stay time and a confidence interval of 90% for anomaly detection (accepting 90% of the user's normal behavior based on the trace samples). Zero probabilities for visiting new locations.	94% accuracy in anomaly detection with FRR <= 10% within 15 minutes - about 6 activities.
Active authentication for mobile devices utilizing behavior profiling [23]	Reality mining dataset[7] - 76 users chosen. RBF tested on 20 users with the dataset divided in two halves.	Differentiating between user patterns (is this a user who he appears to be, based on other users' data). 7/10/14 used for learning, smoothing function applied to the tested activities for anomaly detection - up to 6 activities.	Best results: 9.8% EER with 10 days learning period and 6 activities smoothing. RBF neural network achieved 10,5% EER. Rule based approach - statistical occurrences 11% EER.

Source: own elaboration

174 inhabitants of Poznan area appointed by the home location detection algorithm. Also, a novel division of geographical similarity classes was introduced, transforming the approach described in the literature [22].

3 Trajectory-based model for the behavioral authentication scenario

3.1 Description of the dataset

The mobile phone data used for this work consists of more than 7 billion of anonimized records describing the activity of Orange SA clients in Poland for over 6 months between February and July 2013. This data is typical for publications dealing with the CDR processing [12, 33]. Each data record used in this work consisted of:

– anonymized **id** of a user initiating the call, being the client of Orange and a **receiver** of the service,

– type of a **service** (call, sms, Internet use) used along with associated **measure** e.g. duration in seconds,

– accurate time stamp with a date together with a BTS station data and location_id connected with it[8].

3.2 Trajectory-based model of the mobility

To be able to detect anomalies in the user's behavior, the mobility patterns need to be created to compare new activities against them. Our approach was to use the trajectory-based mobility model and evaluate it in a task of the constant event-based anomaly detection. The process of creation of the mobility profile consists of the following steps:

– extracting activity data,

– applying ABA method,

– creating movement blocks i.e. calculating stay time in a location,

– identifying important locations, passages and routes,

– creating dictionary of user's habits (user's mobility profile).

The process is depicted in the Figure 3. The details of our approach are presented in the following sections.

[8] Meaning a set of BTSs sharing the same coordinates to ease the geographical analysis.

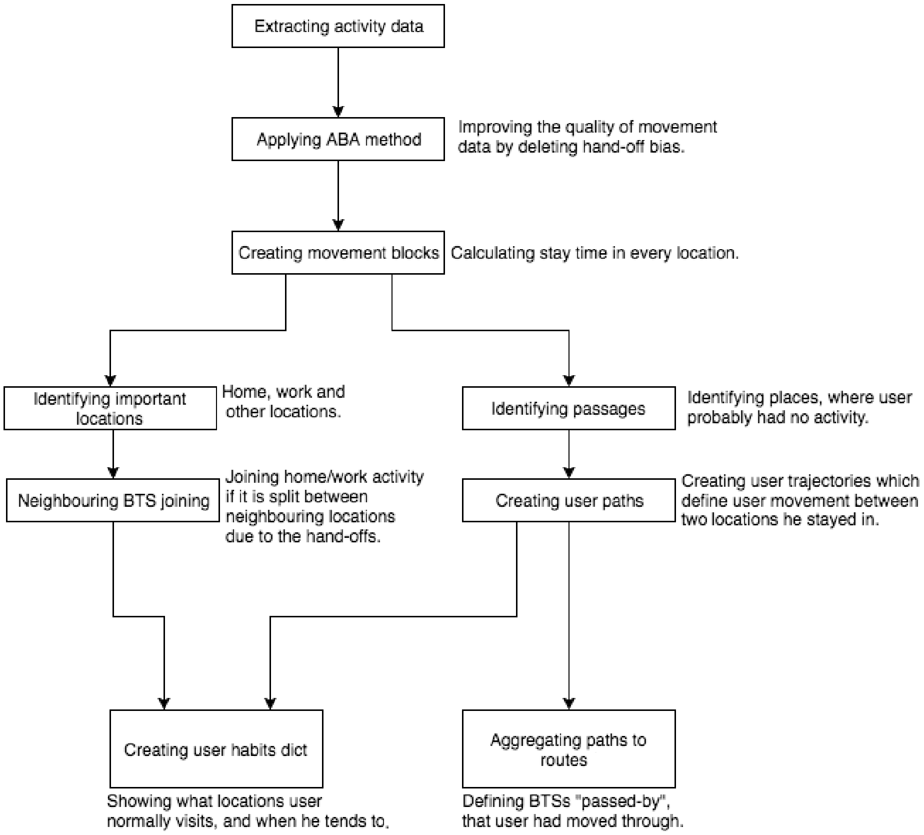

Extracting activity data
Applying ABA method
Improving the quality of movement data by deleting hand-off bias.
Creating movement blocks
Calculating stay time in every location.
Identifying important locations
Home, work and other locations.
Identifying passages
Identifying places, where user probably had no activity.
Neighbouring BTS joining
Joining home/work activity if it is split between neighbouring locations due to the hand-offs.
Creating user paths
Creating user trajectories which define user movement between two locations he stayed in.
Creating user habits dict
Showing what locations user normally visits, and when he tends to.
Aggregating paths to routes
Defining BTSs "passed-by", that user had moved through.

Figure 3: The process of creation of the mobility profile model used in this work. Source: own work

3.2.1 ABA method

Oscillations and quick location changes that appear between successive activities are often a case of the false displacement of a user [24, 38], caused by the traffic balancing or user position between the signal range of two or more stations. To address these problems, a method based on an approach used to clear shifting locations observed in transportation travel proposed by Schlaich is utilized [32]. Therefore, to eliminate these errors, events meeting the following pattern are corrected:

1. First, there is an activity from a location A.
2. Next activity is observed within the next x (10 is chosen based on the literature [20]) minutes from the location B.

3. Third subsequent activity within x minutes from the second activity is labeled with location A.

If errors like that are observed, they are fixed and the sequence of visited locations becomes AAA.

3.2.2 Consecutive activities and stay time

Based on the previous work in the field, considering consecutive activities from the same location that are temporally close to each other can lead to a mostly true assumption that a user stayed in a target location during the time of his activities. The probability of a user staying in a location declines together with the time passing and an upper threshold needs to be introduced. For this work 1h was chosen based on the previous research [20, 41, 26]. If activities are separated by a time less than $1hour$, we consider that the user had a constant stay time in a location.

Moreover, in contrary to the previous approaches, which consider single or temporally distant activities to have no influence on the pattern, different approach is proposed in this work. Due to the fact, that a user activity in a given BTS is considered as a certain information about his whereabouts in this period, we can assume he was there for at least a short period of time. This approach can be called **"weighted" activity labeling**. This method is similar to the time discretization mentioned in the recent literature [42] and based on our tests on the whole database, its use doesn't influence the structure of visited locations. The findings show that sparse activities that are separated by more than an hour are weighted and become at least 15 minutes long.

3.2.3 Identification of passages

With identification of locations with a significant stay time (derived from consecutive activities in a location), BTSs connected to the user movement need to be identified. Approach to distinguishing the "passed-by" locations, is as follows:

1. If a stay time in a location is longer than 30 minutes [24], it is a location where a user had some activity – it is a significant location and therefore a "non passed-by" location.

2. If the period between two consecutive activities is longer than 4 hours [30], the first activity is also labeled as a non "passed-by", to derive any trajectories from the sparse data and avoiding trajectories spanning for multiple days in case of rare activities.

From these trajectories, paths and routes are created that aggregate the user movement. A path is a vector of the user's movement with its start and end in locations with a significant stay time. The path may contain passed-by BTSs that a user moved through to get from the start to the end location, if any were identified. Each path is a trajectory of a user. The routes on the other hand aggregate all paths between points A and B. They are structures that describe the possible passed-by BTSs, when users moved from a point A to B giving probability values to a given sequence of passed-by BTSs. They will be referred to as probability tries later in this work.

3.2.4 Important locations - home/work

Based on the time a user spent in a given BTS station, the most visited, important locations can be distinguished. The home location is the BTS where a user spends most of his time between 19 and 7 in a week. The work location label is assigned to the location with the most time spent between 10 and 18 on weekdays, excluding the home location. To address an additional time spent in neighboring BTSs, the joining algorithm is used to negate the effects of possible hand-off errors in the data.

3.2.5 Data structure of the profile - the dictionary of habits

By identifying locations visited by a user along with the time he visits these locations, a user's mobility profile can be built. A model containing regularly visited locations and user movements between these locations, kept in a weekly-calendar data structure is called **user habits' dictionary**[9]. Each timeframe (1 hour is used in this work) is assigned with locations and routes a user took in the observed period along with their accuracy levels.

3.2.6 Accuracy of the model

Our model calculates the approximated user dwell time for each visited location. A ratio of the time spent in each cell of the habits dict (distinct pair of a day and an hour) compared to the sum or all locations in this timeframe can be calculated. This measure is independent of how active the user was[10], but rather indicates how much time a user spends in a given place during the time period in comparison to other locations that appear during this time. This structure becomes closer to the ground truth for active users[11]. The accuracy values split among locations in a timeframe tell us how predictable the user was in a given period.

[9] Also referred to as dict due to its programming dictionary-like structure.

[10] Very sparse activity with only one location in a timeframe gives it an accuracy of 1.

[11] Meaning the structure of visited locations is really close to the true time spent in these locations.

3.3 Anomaly detection model

In our model we define **anomalies in mobility as situations where a user appears (has an activity) in a location that is not present in his regular movements or the current movement varies significantly from his typical pattern (considering time, geographical area or sequence of places visited and probability of user being in a given location)**. Due to this fact, a model that includes these multiple dimensions of mobility needs to be introduced.

3.3.1 Time

To consider and study the time aspect of a user movement, a simple approach based on the fact that users tend to have distinct daily patterns is used. Each activity threat measure is equal to a number of time frames between observed and behaviors present in patterns in comparison to a max distance (achieving its max at 24h difference between the activities). Given the activity x in a timeframe t (x_t) and the maximal allowed difference in timeframes dt_{max}, considering the distance d in timeframes $d_t(x_t, T)$ between the activity x_t and all of the visits of a given location in other timeframes described in a set T, we can define the time threat as:

$$Threat(x) = min(\frac{d_t(x_t, T)}{dt_{max}}, 1) \tag{1}$$

3.3.2 Geography

Due to the fact that users tend to spend most of their time in already visited locations and their movement is highly predictable, the geographical aspect of a user movement plays an important role in the anomaly detection. Users also tend to move only within a small area of few kilometers around their habitat [6]. When a user is present at one of his "important"[12] locations, the geographical threat measure equals 0.

The geographical threat for a test activity x, equals 1 minus the distance in meters to a closest location from a set of important locations L, compared to the average distance traveled daily d_{daily}.

$$Threat(x) = min(\frac{d(l_t, L)}{d_{daily}}, 1) \tag{2}$$

[12] Regularly visited with more than 5% accuracy.

3.3.3 Sequence

With the added layer of the mobility information about user routes[13] a probability based model of the user movement can be built. It can utilize the built trie routes' model. By updating the routes and paths with counters that assign probabilities to certain trajectories the user took (based on the probability tries), we can extend the probability over the basic "stationary" model of accuracy. The model considers the weighted probabilities of a user following a given trajectory (ordered set of locations). This translates to utilizing an Markov Chain model on the sequence of n visited locations between the stay points extracted.

Reading a test activity x on a level i, means it is an *i-long* sequence[14]. Let $X = (X_1, X_2, ..., X_i)$ be a sequence of locations visited by a user, with a length ($|X|$) being equal to i, where the first place visited in the observed sequence is X_1 and the last is X_i. Then we define the set A that includes all sequences of length i.

$$\bigwedge X,\ if\ |X| = i,\ then\ X \in A \tag{3}$$

Based on this definition, a given test sequence X_t which is of length i and $X_t \in A$, we can define the threat as:

$$Threat(x) = 1 - P(X_t) \tag{4}$$

The probability $P(X_t)$ is calculated by comparing the number of times (C) this sequence appeared compared to the number of all sequences of this length.

$$P(X_t) = \frac{C(X_i|X_1, X_2, ..., X_{i-1})|_{X=X_t}}{\sum\limits_{X|A} C(X_i|X_1, X_2, ..., X_{i-1})} \tag{5}$$

3.3.4 Probability of visiting a location

Considering the mobility patterns of a user, we can focus on the probability of vising a place and distribution of the time spent there. The proposed approach involves creating a structure in which every location is assigned a probability of user's appearance based on the training data set. This probability gives a rough approximation of the time spent in this location as compared to the other locations in this period. It gives a rough approximation of user's movement pattern in a given time-frame and in our case is showcased by the accuracy parameter.

The interpretation of this measure is as follows: "How probable it is that a user is in this location in this timeframe (exact hour and day) compared to

[13] And the predicted accuracy of the BTS appearing in comparison to the routes in the pattern.

[14] E.g. for i=3 we consider all sequences that are of length 3, like: ABD, ABC, ACE.

the other places he visits". If a user visits a location that is present in the timeframe[15], including passed-by locations that match his currently traveled route, the uncertainty measure is calculated as follows. For a location l in a timeframe t, where the accuracy of an activity x in a location l and a timeframe t is denoted as $a(l,t)$:

$$Threat(x) = 1 - a(l,t) \tag{6}$$

The following sections will present the evaluation of the proposed method.

4 Evaluation of the method: using mobility in the behavioral authentication scenario

In order to verify the usability of the user's profile in the authentication and non binary authorization scenario, its outputs - namely threat levels, need to be tested to better describe everyday mobility behavior and differences between users.

4.1 Preparation of data

First a sample, consisting of users that shared a similarity in a geographical profile (being from Poznan area), was chosen to test the model in a scenario that would be close to real life applications of the model[16]. This also allowed to build a "hierarchy" of users based on the probable increase in similarity of mobility profiles to test model for different cases.

Based on the requirements of the model, home and work locations for all users that appeared in the Poznan area in March 2013 were calculated. The area was chosen based on the TERYT[17] mapping. This returned 173 distinct location_id's that were considered being in Poznan area as shown in Figure 4.

4.2 Division of users into classes

The studies mentioned in the literature did not set a stable testing environment, therefore a definition of such a testing approach was needed. This approach to testing methods on anomaly detection is novel and may be applied by other researchers in the field. Such an approach may enable future comparison of results between various approaches. We propose to evaluate similar methods addressing different levels of similarity to the tested user behavior, including:

[15] In our CDR dataset case - an hour.

[16] As it is obvious that selecting a random user for the anomaly detection will yield positive results in the anomaly detection but is not the case for most of the real life scenarios e.g. when the phone is stolen.

[17] The Polish administrative areas' territorial mapping.

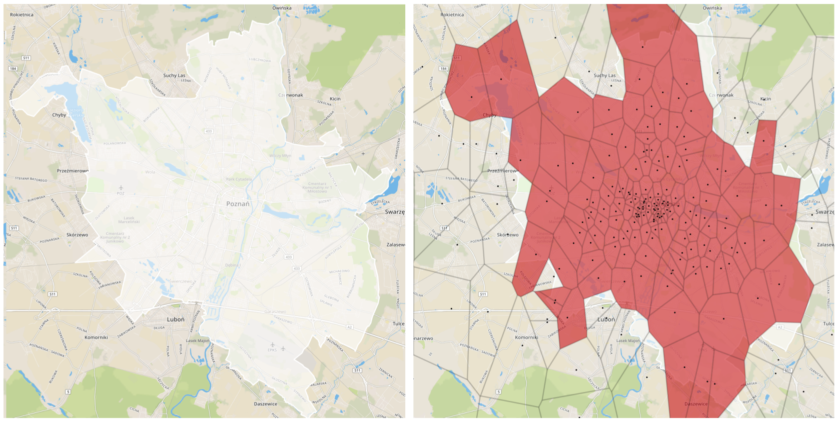


Figure 4: Visualization of Poznan administrative borders on a city level (on the left) along with the BTS stations laying inside this area with their Voronois colored (on the right). Source: own work

- **the same user** - choosing the unobserved new data of a user allows to test the extent of predictability of human patterns and sensitivity of the threat measures, while giving a clear answer about the **false rejection rate** for anomaly detection cases.
- **A random user** - similarly to the most of the approaches in the literature, a random user from the sample was chosen. This showcased how the trajectory model compares to the approaches in the literature.
- **A user from the same town** - by choosing a user that has a home location that falls into the same town, which more closely resembles a phone theft than a random user choice does.
- **A user with the same home location** - potential success in differentiation of these patterns would allow us to differentiate between e.g. family members sharing a phone.
- **A user with the same home and work location** - in this scenario the goal was to verify whether the characteristics of the mobility differs between just visiting the same locations (in sequences taken, time of visits etc.).

4.3 Identifying deviations in the mobility by measuring the activity threat levels

The evaluation of the model concerned checking, if the model is capable to differentiate between user patterns using threat values for new activities of the same and other users (from the similarity classes). The tests were performed on the sample of 1000 users from one month for whom corresponding samples in all of the test classes could be found. The statistics of this sample are shown in Table 2. The movement list is a structure that aggregates phone activities into a stay time labeled parts of trajectories - effectively aggregating activities that lengthen the stay in one location, meaning users have an average of 4 temporally distant activities daily (139,2 monthly). The distribution of a distance showcased a long tailed distribution, where a half of users has a daily distance shorter than 11.29 km. Home and work location accuracy levels state that users spend on average 70% of their night time at a home location, and a little above 50% of their work daytime at a work location.

Each user chosen for the test had his/her user profile built on the available data from one month. For each of them first **40 activities from the following month of their activity** were tested against the profile.

Table 2: Monthly statistics describing users who were the reference users in the class-based comparisons of the threat activity labeling.

	Activities in the movement list	Distinct locations visited	Average daily distance (km)	Home location accuracy	Work location accuracy
Min	2	1	0,003	0,05	0
Max	680	340	295,06	1	1
Mean	139,2	26	22,03	0,70	0,57
1st Quartile	77	11	5,6	0,51	0,34
Median	120	18,5	11,29	0,75	0,57
3rd Quartile	175,2	32	22,35	0,91	0,80
Std dev	87,79	26,77	33,76	0,24	0,27
Skewness	1,83	3,92	3,91	-0,51	-0,07

Source: own work

The results of the experiment prove that in the long run (with the average values for 40 activities) we can differentiate between user classes as shown in the Table 3 and Figure 6. Given enough data, the distinction between a user and someone very similar to him in terms of mobility is possible and the distinction is clearly visible in the average threat values. The distribution of average threat levels observed during the testing for each similarity class shows that classes influence the threat level distribution. High threat levels regarding the same town scenario also show a possibility to evaluate methods regarding fraud detection given much shorter timespan. The same user class threat distribution presented in Figure 5 depicts to what extent the user pattern is consistent over time on a sparse data (from CDR). On average, users show some level of unpredictability visible in the average threats generated by users, but this measure is not a normalized definition. No conclusions can be made just out of this fact, without deeper analysis of the variables influencing repeatability level of mobility patterns. **The higher the threat level presented in the table and in the pictures, the more the pattern measured differs from the user's pattern (the lower is the uncertainty).**

Table 3: Comparison of the average threat levels for user classes.

class type	Average geographical threat	Average sequence threat	Average time threat	Average uncertainty threat	Average of threats
same_user	0.06	0.26	0.32	0.69	0.33
home_work	0.14	0.39	0.45	0.77	0.44
home	0.26	0.58	0.62	0.84	0.58
town	0.47	0.96	0.98	0.99	0.85
random	0.95	1.00	1.00	1.00	0.99

Source: own work based on CDR data

4.4 Anomaly detection on Poznan sample

Based on the findings of the above experiment, the structure of threat levels was described depending on the similarity level to a user. These findings allow for identification of anomalies based on the threat level measure observed. **The uncertainty measure was omitted** in this classification due to the fact that it provides high threat values and could not be used for the threat threshold creation later. Nonetheless, it remains as an interesting characteristics of the movement as the more dense is the data, the more useful it would be due to the fact that with regularly sampled data (average sampling rate equal to time frame length) it closely resembles the real mobility and time spent in a location patterns of a user. The use of this measure for regularly sampled phone data would make this measure directly applicable in the model.

Since users vary heavily, when it comes to their mobility profiles and habits, the model which takes this phenomenon into consideration needs to be created. The model should minimize the false rejection rate for users with highly varying profiles, while also minimizing the false acceptance rate for users with more stable and predictable behavior. To address this challenge, an approach of calculating confidence intervals for the three threats (time, geography, sequence) is presented. For each of these threats, it is set as a 90th percentile of the corresponding threat values calculated on the validation data set of user activities [46]. To check if the tested activity is an anomaly, we analyze all three threats for this activity and if at least for one of them a confidence interval for the target threat is exceeded, the model marks this activity as an anomaly in this dimension. Whether one or more scores need to exceed the threshold to classify

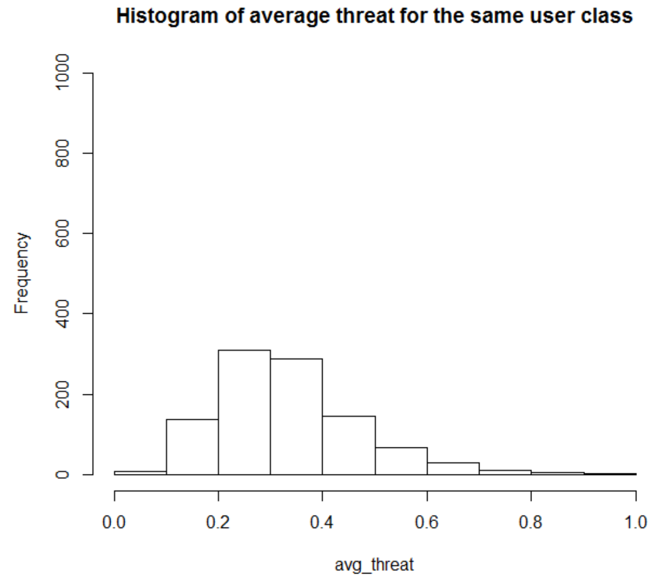


Figure 5: The histogram showcasing the values of the average user threat levels (of 40 test activities) for 'the same user' scenario, x axis indicates the threat levels and y describes a number of occurrences. Source: own work based on CDR data

an activity as an anomaly remains a matter of future work, and is a parameter in the proposed model. Setting this value high can cause higher false acceptance rate for tested activities, resulting in less anomalies detected. The approach that is proposed for the model learning and anomaly detection is presented in Figure 7.

To create confidence intervals for all three measures:

- user's mobility profile needs to be created from the learning data period,
- target threat values for all activities from the validation data set need to be assigned,
- smoothing function is applied by using the moving average on the threat values,
- 90th percentile of the above mentioned moving averages is defined as a target threat threshold for each of the threats.

Smoothing function that treats a number of successive activities as one event was introduced to deal with the mobile users inconsistent and variable usage

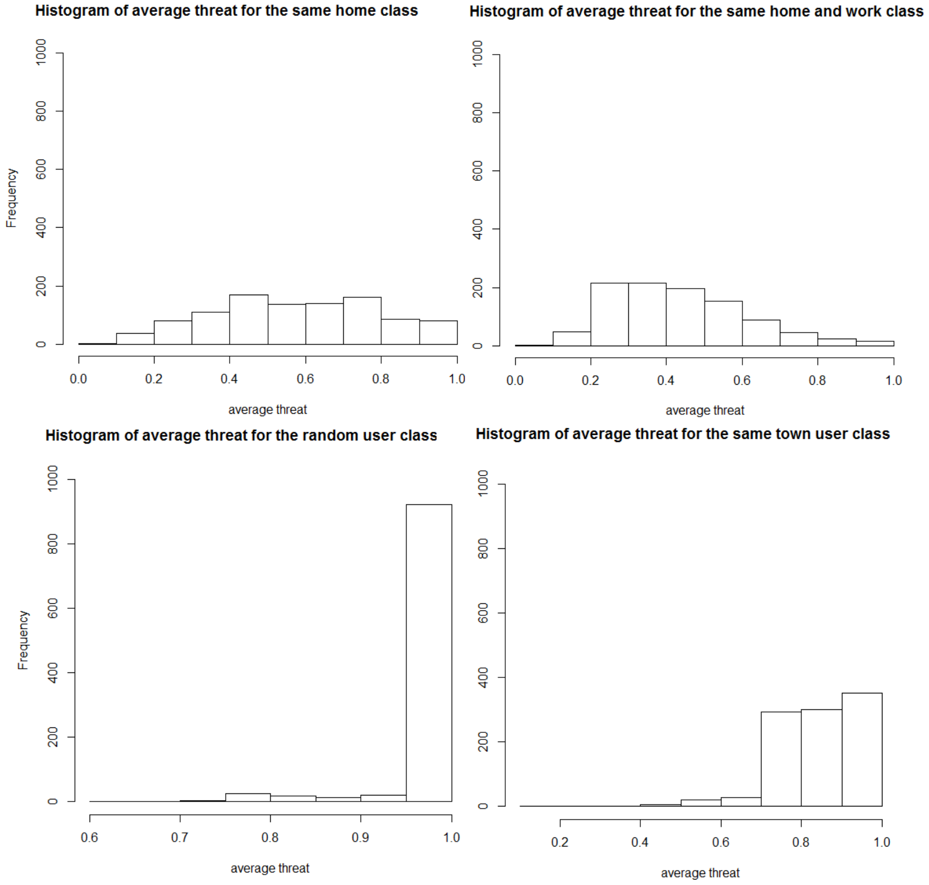


Figure 6: The histogram showcasing values of the average user threat levels (of 40 test activities) in the test classes scenario, x axis indicates the threat levels and y describes a number of occurrences. Source: own work based on CDR data

behavior. Therefore, a decision is made based upon the combined events rather than a single occurrence [23]. In our approach an **interval size equal to 3** is used. The moving average is calculated in each respective threat value to smooth it. This is a parameter in the method which can be adjusted e.g. based on how active the user is to achieve the lowest number of false positives, while detecting an anomaly for a user in an acceptable timespan.

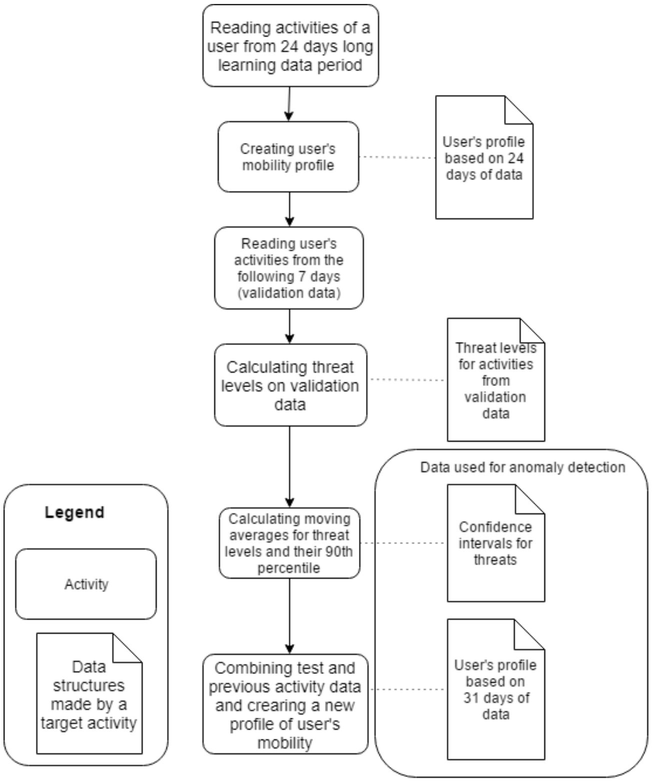

Reading activities of a user from 24 days long learning data period
Creating user's mobility profile
User's profile based on 24 days of data
Reading user's activities from the following 7 days (validation data)
Calculating threat levels on validation data
Threat levels for activities from validation data
Data used for anomaly detection
Legend
Activity
Data structures made by a target activity
Calculating moving averages for threat levels and their 90th percentile
Confidence intervals for threats
Combining test and previous activity data and crearing a new profile of user's mobility
User's profile based on 31 days of data

Figure 7: The approach used to define threat intervals and enable application of a user's profile for anomaly detection. Source: own work

4.5 Results

Based on the confidence intervals, an anomaly detection experiment was carried out. Its results are shown in Table 4. By exclusion of users having all of the threat intervals equaling 1[18], about 7% of the sample was removed. Each excluded user had his/her threshold levels equaling 1 in all three dimensions. This group of users did not have a stable pattern overall and could not be used for the model based checking of anomalies in the behavior.

To increase the method's performance, an improvement was tested that decreased the values of anomaly thresholds iteratively by one percentile below 90 for all users that had unpredictable patterns. This resulted in the reduction of percentage of excluded users to 0.8% and improved the model accuracy.

The FRR in the scenario was equal to the fraction of situations, where the valid user data from the test period was classified as an anomaly measure -

[18] Meaning no anomaly could be ever be detected using this model.

accuracy in same user class. On the other hand, the remaining accuracy values indicated a number of situations, where activities of users from another class were properly labeled as anomalies - this related to the effectiveness of the algorithm in detecting a change of the user.

Table 4: The results of an anomaly detection method (FRR) with the use of 3 activities batch length and 2 measures classified as anomaly. Source: own work

Approach	Static 90'th percentile		Iterative		
% of users rejected due to the unstable pattern	7%		0.8%		
Number of measures needed to classify an anomaly	1	2	1	2	3
Results					
Class	Measure (% of anomalies classified in)				
random class	99,33%	85,16%	99,58%	99,57%	97,17%
same town class	88,50%	72,32%	96,65%	91,63%	70,84%
same home location class	60,40%	44,18%	70,17%	53,64%	37,29%
same home and work location class	43,80%	26,04	53,21%	32,03%	20,09%
same user class (FRR)	20,80%	8,83%	32,68%	13,79%	6,33%

The results of the method, while using 3 measures show that when we define anomalies as a travel beyond the user's geographical area, we can achieve authentication methods with an accuracy similar to the approaches in the literature (the proposed model achieved about 97% accuracy, with the FRR staying close to 6% and FAR to 3% in the random class). This also clarifies why simple probability based methods achieve good results based solely on coordinates, visited locations or area of movement. The best outcome of the model was achieved using 2 measures for the anomaly classification[19]. The model proved to be effective in detecting anomalies in same town class scenario (an example of probable theft), while still remaining to be quite effective in differentiating people living in the same area (same home location class). The differentiation between similar users still seems to be an issue based solely on mobility and maybe other behavioral features would be the most successful when applied in these scenarios.

It is worth to underline, that the model achieved an accuracy similar to the approaches presented in Table 1, while working on much more sparse and unevenly sampled dataset. This proves that CDR derived authentication methods

[19] For clarification, this means that two distinctive measures out of three (geography, sequence, time) needed to exceed their respective thresholds for the model to consider an activity batch to be anomalous and and not belong to an original possessor of the device.

can achieve accuracy similar to the presented methods working on a device level data. Due to the fact, that characteristics of the data and the parametrization of the models (e.g. time window used for anomaly detection) plays a great role in the accuracy of the model. Therefore, the methodology for future comparisons of authentication methods needs to be discussed.

5 Discussion

The performance of the anomaly detection algorithms relies highly on the characteristics of the dataset and the model. Due to this fact, comparing the results of the methods working on different datasets may prove difficult - in our case we were able to achieve the same accuracy with more sparse data, but the scenario of the same town class is more useful to assess the quality of the model than the random class used in the literature. This makes the detailed comparison of result's metrics a good area for further work, which would not fit into the scope of the paper.

Based on the findings of this work, the requirements for benchmarks of algorithms in the future should include:

- **Spatial homogeneity of the dataset** (proposed approach: describing the area of study) – an area of the study should concern users of very similar patterns (like students/employers of a university) what may prove to be more challenging than just comparing random CDR users and influence the results.
- **Spatial homogeneity of a model compared to the test data** (proposed approach: division of the results in the comparison classes) – comparing a test profile with a random user always produces a high accuracy, the task becomes harder when comparing users that share locations visited with the base user. This also allows for building pattern differentiating methods that would be able to distinguish between family members sharing a phone and could be applicable not only on mobility data.
- **Sampling frequency** (proposed approach: calculating inter-event time or an average number of activities/day per user) – the activity based data like CDR varies in its characteristics depending on the users and their inclinations to more often phone activities. GPS frequently sampled data remains at a very different resolution and may provide significantly better results even with the use of naive methods.
- **Learning period length** (propose approach: stating the length of time period used for the model learning) – due to the fact that human mobility differentiates between days of the week, and pattern stabilizes only about

after two weeks, the length of the dataset remains important. Also very lengthy learning period may require model updates or recent data weighting.

- **Approach requirements** (proposed approach: stating number of activities (or time) needed for classification) – especially important in case of identification approach, where data from all users is compared to ensure uniqueness of the pattern. In case of anomaly detection, only extensive data on the user is needed. While pattern identification requires data for all (or many) of users to learn a model, an authentication approach uses only the user data as a one-class classifier.
- **Accuracy and type of the geographical label used** (proposed approach: stating or calculating geographical bias of used sensor, or providing density and average BTS area) – the average size of BTS area can be a good measure of accuracy for the CDR data. This also allows data to be comparable e.g. by introducing artificial bias when comparing results with more dense areas.
- **Other data used included in the model** besides of the tested aspect (proposed approach: measure the influence of other variables e.g. accelerometer readings on the performance of the model) - any other feature used besides the one tested (in this example geography) should be excluded from the base model to remain comparable to the current approaches.

6 Summary

In this paper we described advantages of using the behavioral authentication on mobile devices, along with the possible measures and methods that can be used. A trajectory based model was introduced along with the defined measures and definitions of anomalies in the dimensions of: geography, time, sequentiality and predictability. The model was tested on a large sample of CDR data and provided to be effective in dealing with sparse datasets. The results of the anomaly detection were satisfying in differentiating between the users and the model was proven to be effective in detecting the possible theft scenarios. What is worth to note, is that the modular design of the proposed solution allows for an ensemble of machine learning (or other domain based) methods to be easily utilized in the model. Nonetheless, the additional insight and findings concerning the repeatability of paths users traveled would not have been possible, if machine learning methods or probability based naive classifiers were used to detect anomalies.

The unpredictability of the user movement - captured by the FRR (6%) was similar to the studies in the literature and the accuracy of the model was also similar (97%). The model proved valuable in detecting a simulated theft scenario and provided insight into causes of good results of other methods. Differentiating between similar users proved to be difficult with the CDR data. Utilizing only mobility in this scenario may not be enough to differentiate between users living in a close proximity. Nonetheless, the mobility pattern may be of use in a more complicated system utilizing more behavioral factors.

The division on anomaly classes allowed to create a benchmark for the mobility based anomaly detection models considering the similarity of users. However, parameters used in the model (such as the length of the activity batch) and the dataset characteristics can also influence the outcomes.

6.1 Future work

The influence of the time and the number of activities needed for classification is one of the main areas for further testing of the model, along with the comparison on various datasets (including the whole CDR database). As the model was tested on a sparse dataset, considering the influence of this characteristics on the output of the presented methods, and testing the model on the phone generated data could provide accuracy scores more comparable to the other algorithms.

For directly improving the accuracy, developing methods that would calculate the **similarity of trajectories** and including **less rigorous thresholds on the time aspect** would definitely improve the performance of the model. Adding a **semantic aspect** on the visited places[20] could also improve the model but would significantly increase the complexity. Exchanging the methods used for the threat definition with machine learning algorithms that would be tailored and suitable for a given aspect of human mobility e.g. RNN (Recursive Neural Networks) would probably cope well with learning sequential patterns. Similarly SVM or density based methods would work well on estimating the geographical area that a user travels through based on the coordinates. The use of those methods could potentially help in achieving a higher accuracy (after the initial insight provided by this model has shown their potential areas of application). Nonetheless, it would greatly influence the computational complexity of the model and would require an ensemble of methods to utilize all dimensions.

After focusing mainly on one aspect of a behavioral biometry - mobility, the model could be also extended over different behavioral aspects like analysis of touchscreen interaction to better distinguish between users in high similarity classes.

References

1. Mobile devices secure or security risk? https://www.gartner.com/doc/2595417, 2013. Accessed: 2017-10-10.

[20] If a user is visiting a grocery store in a constant time period, being in an unobserved location where the grocery store is located should not generate a high level of threat, when we consider the semantics of the place.

2. Beyond the password: The future of account security. `https://www.telesign.com/wp-content/uploads/2016/06/Telesign-Report-Beyond-the-Password-June-2016-1.pdf`, 2016. Accessed: 2016-09-10.
3. Market research - biometric smartphone model list. `http://www.acuity-mi.com/BSP.php`, 2016. Accessed: 2016-07-19.
4. T. Aledavood, E. López, S. G. Roberts, F. Reed-Tsochas, E. Moro, R. I. Dunbar, and J. Saramäki. Daily rhythms in mobile telephone communication. *PloS one*, 10(9):e0138098, 2015.
5. N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, page 1473871615581216, 2015.
6. J. P. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.
7. R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Commun. ACM*, 56(1):74–82, jan 2013.
8. J. Bonneau and S. Preibusch. The password thicket: Technical and market failures in human authentication on the web. In *WEIS*, 2010.
9. R. Buschkes, D. Kesdogan, and P. Reichl. How to increase security in mobile networks by anomaly detection. In *Computer Security Applications Conference, 1998. Proceedings. 14th Annual*, pages 3–12. IEEE, 1998.
10. S. Buthpitiya. Modeling mobile user behavior for anomaly detection. 2014.
11. F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.
12. S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*, (2526):126–135, 2015.
13. B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.
14. R. Damaševičius, R. Maskeliūnas, A. Venčkauskas, and M. Woźniak. Smartphone user identity verification using gait characteristics. *Symmetry*, 8(10):100, 2016.
15. R. Damaševičius, M. Vasiljevas, J. Šalkevičius, and M. Woźniak. Human activity recognition in aal environments using random projections. *Computational and mathematical methods in medicine*, 2016, 2016.
16. B. Fox, R. van den Dam, and R. Shockley. Analytics: Real-world use of big data in telecommunications. *IBM Institute for Business Value*, 2013.
17. L. Fridman, S. Weber, R. Greenstadt, and M. Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. 2015.
18. B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behavior. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 550–555, 2013.
19. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
20. M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González. Development of origindestination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.
21. S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in Los Angeles and New York. *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011*, pages 88–93, 2011.

22. H. G. Kayacik, M. Just, L. Baillie, D. Aspinall, and N. Micallef. Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. *arXiv preprint arXiv:1410.7743*, 2014.
23. F. Li, N. Clarke, M. Papadaki, and P. Dowland. Active authentication for mobile devices utilising behaviour profiling. *International journal of information security*, 13(3):229–244, 2014.
24. F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools. Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174–6189, 2014.
25. X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3, 2013.
26. D. Maldeniya, S. Lokanathan, S. Lanka, A. Kumarage, and S. Lanka. Origin-Destination Matrix Estimation for Sri Lanka Using the Four Step Model. (May):785–794, 2015.
27. O. Mazhelis and S. Puuronen. A framework for behavior-based detection of user substitution in a mobile context. *computers & security*, 26(2):154–176, 2007.
28. H. Österle, J. Becker, U. Frank, T. Hess, D. Karagiannis, H. Krcmar, P. Loos, P. Mertens, A. Oberweis, and E. J. Sinz. Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20(1):7–10, Jan 2011.
29. S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, pages 14–25. Springer, 2010.
30. M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez. Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, 42(4):647–668, 2015.
31. D. Połap and M. Woźniak. The use of wavelet transformation in conjunction with a heuristic algorithm as a tool for feature extraction from signals. *Information Technology and Control*, 46(3):372–381, 2017.
32. J. Schlaich, T. Otterstatter, and M. Friedrich. Generating Trajectories from Mobile Phone Data. *Transportation Research Board 89th Annual Meeting*, pages 1–18, 2010.
33. C. M. Schneider, V. Belik, T. Couronne, Z. Smoreda, and M. C. Gonzalez. Unravelling Daily Human Mobility Motifs. *Journal of The Royal Society Interface*, 10(84):20130246(1–8), 2013.
34. C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.
35. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
36. B. Sun, Z. Chen, R. Wang, F. Yu, and V. C. Leung. Towards adaptive anomaly detection in cellular mobile networks. In *the IEEE consumer communications and networking conference*, volume 2, pages 666–670, 2006.
37. B. Sun, F. Yu, K. Wu, and V. Leung. Mobility-based anomaly detection in cellular mobile networks. In *Proceedings of the 3rd ACM workshop on Wireless security*, pages 61–69. ACM, 2004.
38. G. Tandon and P. K. Chan. Tracking user mobility to detect suspicious behavior. In *SDM*, pages 871–882. SIAM, 2009.
39. S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 159–168. ACM, 2012.
40. M. Ulinskas, M. Woźniak, and R. Damaševičius. Analysis of keystroke dynamics for fatigue recognition. In *International Conference on Computational Science and Its Applications*, pages 235–247. Springer, 2017.

41. P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.
42. P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, 2015.
43. R. Xie, Y. Ji, Y. Yue, and X. Zuo. Mining individual mobility patterns from mobile phone data. In *Proceedings of the 2011 international workshop on Trajectory data mining and analysis*, pages 37–44. ACM, 2011.
44. G. Yan, S. Eidenbenz, and B. Sun. Mobi-watchdog: you can steal, but you can't run! In *Proceedings of the second ACM conference on Wireless network security*, pages 139–150. ACM, 2009.
45. J. J. Yan, A. F. Blackwell, R. J. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security & privacy*, 2(5):25–31, 2004.
46. S. Yazji, P. Scheuermann, R. P. Dick, G. Trajcevski, and R. Jin. Efficient location aware intrusion detection to protect mobile devices. *Personal and Ubiquitous Computing*, 18(1):143–162, 2014.

6.4 Conclusions

The chapter aimed at describing a horizontal application for profiling techniques, proposing the profiling-based authentication method. The goal of the chapter was to "develop a method enabling for authentication of a user based on the Call Detail Record data". This goal was further translated into three secondary goals addressed by specific sections of this chapter. First two, concerned amount of data needed to assure building a precise user profile, in particular:

- to verify, if the Call Detail Record data is sufficient for detecting anomalies in the behavioural user profile and therefore enable for applying the CDR-based profile in authentication scenarios,
- to research how much data describing a user is needed to provide an efficient authentication solution.

In relation to these goals, the goal of the presented paper was to detect user mobility patterns for the needs of anomaly detection and authentication using the Call Detail Records, and check what would be the quality of such a model. The paper included in Section 6.2 proposes a model describing user's mobility that may be used for anomaly detection. The activities from Call Detail Records (taking into account all available data describing a single activity) were translated into a mobility dictionary for a user e.g. stay times at locations were quantified. The profile of a user was specified as: locations visited by a user, time that a user spends at a given location (probability that at a specific hour, a user is present at a specific place), sequence of BTS station visited (identification of routes of a user). Then evaluation of the approach was performed to prove accuracy of the approach (99% of model's accuracy for a so-called random user scenario). This proved that it is possible to build authentication method based on Call Detail Records, even taking into account sparsity of this data.

The third detailed goal addressed in this chapter concerned development and validation of a method for description of a user for the needs of authentication. The supplementary goal was to develop a methodology for testing behaviour-based approaches based on Call Detail Records data. The goal of the paper addressing this issue was therefore to propose a working model of a behaviour based authentication applying anomaly detection performed over the user's mobility patterns. The supplementary goal was to create a methodology for testing similar, behaviour-based approaches. The paper included in Section 6.3 presents a method (together with its validation) that enables development and verification of a mobility profile of a user and may be used for authentication. The method identifies and exploits anomalies in the user behaviour. These anomalies include e.g. checking how movements of a user differ from his typical behaviour (considering time, geographical areas, sequence of places visited or probability of a user being at a specific time, in a specific location). This approach is validated in the paper to prove its usefulness. In addition, the methodology of testing behaviour-based anomaly detection approaches is proposed. It is discussed why a random user comparison method does not make sense while researching authentication scenarios and how evaluation of authentication methods should be carried out.

Chapter 7

Profiling for Personalisation

7.1 Introduction to Personalisation

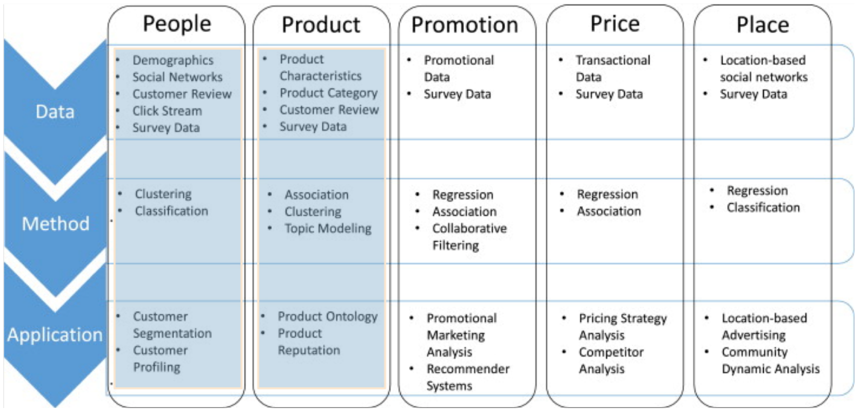
7.1.1 Motivation

Nowadays, profiling is often applied for the needs of personalisation of products and services. This is because the people's perspective within the Marketing Mix underlines the need for a detailed understanding of who is our customer and react accordingly [39]. Figure 7.1 presents data that should be collected on a customer, together with methods and applications targeting generation of business value. The data that a company would like to have about its customer includes (but is not limited to):

- demographic data of a customer and his location,
- socio-economic data of the client and his ecosystem,
- data on preferences,
- data on purchase history,
- trend analysis in the above data,
- payment behaviour.

The more data is acquired, the more detailed the profile and the better the recommendations of products and services for the customer. Better recommendations usually are directly related with an increase in sales. An example of the advanced recommendation may be found on Netflix[1], which besides of extending the list of features of movies, what impacted the number of genres

[1] http://blog.springtab.com/personalization-examples-netflix/?utm_campaign=quora+answer&utm_medium=social&utm_source=quora.com



	People	Product	Promotion	Price	Place
Data	Demographics Social Networks Customer Review Click Stream Survey Data	Product Characteristics Product Category Customer Review Survey Data	Promotional Data Survey Data	Transactional Data Survey Data	Location-based social networks Survey Data
Method	Clustering Classification	Association Clustering Topic Modeling	Regression Association Collaborative Filtering	Regression Association	Regression Classification
Application	Customer Segmentation Customer Profiling	Product Ontology Product Reputation	Promotional Marketing Analysis Recommender Systems	Pricing Strategy Analysis Competitor Analysis	Location-based Advertising Community Dynamic Analysis

Figure 7.1: Motivation for profiling clients and products. Source: [39]

(Netflix distinguishes about 80,000 micro-genres), applies the customer-centric approach. Netflix builds a customer profile from the very beginning – while registering to Netflix, a customer is asked about his/her preferences based on three simple movie choices. Then, he/she is assigned to a certain group with similar preferences, but what is important to note, he/she may belong to more than one group. Based on his/her interactions with the service (his/her behaviour), a user profile is detailed (based on micro-genres), and future recommendations are based on a result of a forecast of the film's rating by a user.

Other data important for profiling a customer may include[2]:

- day of the month of the transactions of a customer e.g. customers can shop in the first week of the month, because they have just received their salaries,
- day of the week: analysis of the customer's purchase history by examining days in which the customer made purchases. If we know that a customer usually buys on Sundays, we should send him a reminder about products he was looking for on Sunday, instead of bombarding him with notifications on other days.
- time of day: if a customer usually visits the site late in the evening, it can be assumed that a client is a working specialist. If we are to reach him/her, we should try to reach him/her during the late evening hours, because then the chance of customer interest increases.
- discount effect: understanding which customers respond positively to discounts, and what is the right percentage discount, helps us in positioning our products and at the same time balancing profit margins.

Though, profiling is important not only for the marketing and product placement. It may

[2] http://www.data-mania.com/blog/customer-profiling-and-segmentation-in-ecommerce/

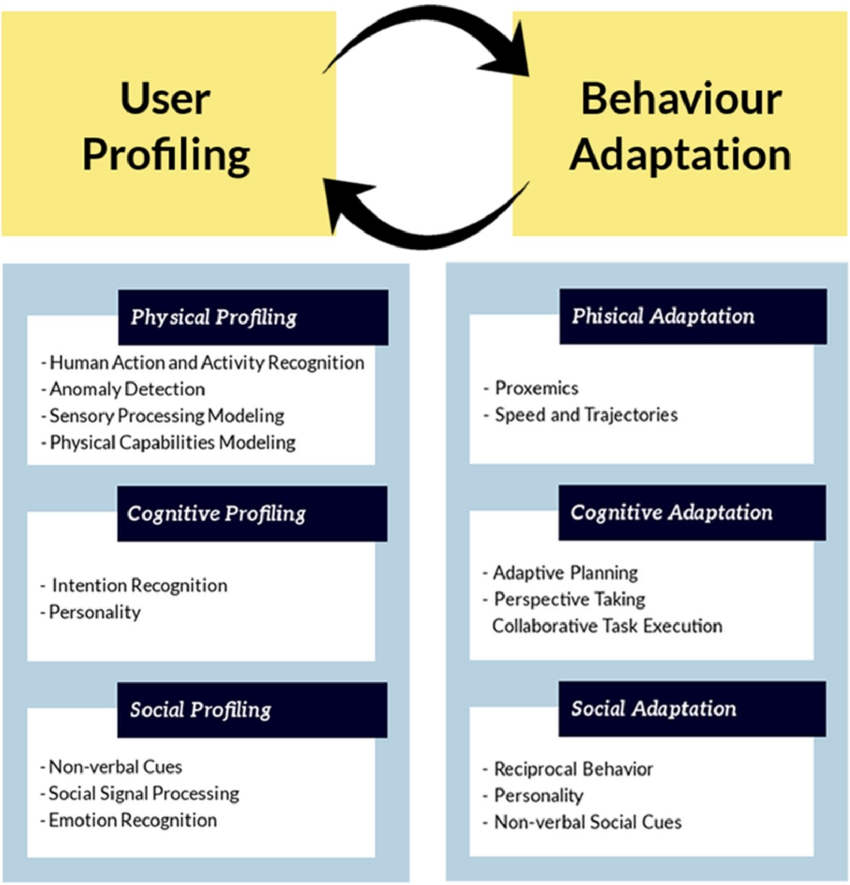


Figure 7.2: User profiling and behavioural adaptation. Source: [137]

be crucial also for acceptance of products in the field of robotics [137], especially in case of robots interacting or supporting people. Figure 7.2 presents aspects of profiling and respective adaptations that a robot should apply based on profiling results.

This chapter is to cover two fields for horizontal personalisation: personalisation for the need of content adaptation and understanding who is a customer of a shopping centre.

7.1.2 Goals

The goal of the chapter is to propose profiling methods that enable to describe a user in a way supporting personalisation of content or a service. The secondary goals that enable achieving the main goal are as follows:

- Proposing a personalisation method that improves the user experience of a solution, but does not impact the efficiency of the solution.
- Providing a method for user profiling that allows obtaining valuable insights from data that

was not collected for a specific purpose.

7.1.3 Structure of the Chapter

The chapter consists of four sections including an introduction presenting relation to goals of the thesis and a summary that presents results that were achieved in relation to these goals. Section 7.2 presents solution improving user experience applying personalisation and studies efficiency of proposed approaches. Section 7.3 describes a method for profiling a user that enables to obtain valuable insights from data that was not collected for a specific purpose.

7.2 Scalable Adaptation of Web Applications to Users' Behaviour

The goal of the paper is to propose and evaluate the efficiency of a method enabling a client-side adaptation of a Web interface. This goal contributes to achieving one of the secondary goals of the thesis, namely proposing a personalisation method that improves the user experience of a solution, but does not impact the efficiency of the solution.

The paper was published in Lecture Notes in Artificial Intelligence from 4th International Conference, ICCCI 2012, Ho Chi Minh City, Vietnam, November 28-30, 2012. Detailed bibliographic reference is as follows: Węcel, K., Kaczmarek, T., Filipowska, A., 2012, Scalable Adaptation of Web Applications to Users' Behaviour, Lecture Notes in Artificial Intelligence, 7654, pp. 79-88.

Scalable Adaptation of Web Applications to Users' Behavior

Krzysztof Węcel, Tomasz Kaczmarek, and Agata Filipowska

Department of Information Systems, Faculty of Informatics and Electronic Economy, Poznań University of Economics
k.wecel,t.kaczmarek,a.filipowska@kie.ue.poznan.pl

Abstract. In this paper we present a comparative study of performance of an adaptive web application supporting personalization either on a client or on a server side. Currently, modern applications being developed support various kinds of personalization. One of its types is changing behavior and appearance in a response to actions taken by a user. Not only existing rules should be applied but also new patterns discovered online and for different levels of events. Scaling such applications to a large number of users is challenging. First, the stream of events generated by users' actions may be huge, and second, processing of the adaptation rules per single user requires computing resources that multiply with the number of users.
This paper reports on the efficiency of the method enabling a client-side adaptation after moving adaptation logics from a server to a client.

1 Introduction

The adaptability and in particular development of adaptable user interface pose a number of challenges for application developers [7, 4] including, among the others, the issue of scalability.

The core of every adaptable system is the user model describing user preferences expressed explicitly or implicitly, derived from her behavior. This model should be updated, when new information is delivered, which is of particular importance when the user behavior is being traced. The user behavior patterns should be further transformed into adaptation rules. These rules have to be evaluated on the constant basis, as new events are caused by the user interacting with the application.

It is a challenge for most of the Web applications to provide many users with a personalization result instantly, because all processing is traditionally conducted on the server side, and the client (browser) is only responsible for rendering the final result. For large-scale applications, it is not feasible to evaluate dozens of rules for each user on the server side, even if efficient algorithms are applied. Therefore, the main problem is scalability of an adaptable Web application, which, as we are to show is achievable, if rule processing is moved entirely to the client side.

The example application on which we conducted our research is a modern e-banking application implemented in the Google Web Toolkit framework, with adaptability enhancements. We conducted several experiments that confirm the efficiency of rule based approach to adaptability on the client side and show the scalability of our solution to the rule execution problem.

2 Related Research

The problem of adaptability and in particular adaptable user interface (or intelligent user interface) has been studied for years in the context of regular applications ([7]) as well as hypertext ([4]). A goal of an adaptable system is to deliver content and experience that match best user's preferences, knowledge and experience level. The user may express the preferences regarding the look and behavior of the interface explicitly or implicitly (through interaction with the system and the events that the users generate). These preferences are used in the adaptation process, which may take a form of personalization or customization [1, 8]. Thanks to the development of Web technologies, the client-side scripting in particular, it is possible to capture detailed events generated by the user in the browser (such as mouse movement and individual keystrokes) [3]. This however, makes the stream of events denser and exerts greater pressure on the server to process it.

In order to mine the patterns of user behavior, it is necessary to apply the classic data mining methods (for example association rules [2]). The mined patterns are then converted to rules: this is the prevalent approach to date [10]. Depending on the approach, these rules may be event-based (series of events matching the rule head result in an action being executed) or state-based (rules are sensitive to the state change of the application) with several variants of how expressive the rule formalism is adopted ([5, 9]. More expressive formalisms, which are able to express sequences of events are said to handle a wider range of user requested adaptations [6]. Recently, semantic techniques are being adopted to modeling user preferences and adaptations [11].

3 Personalisation in the Web application

In this section we shortly describe main assumptions and ideas behind our personalisation method as well as solutions that were tested in the experiments. The e-banking Web application that we conducted our experiments on allows to conduct standard tasks (checking accounts balances, history of transactions, place money transfer orders and standing orders) and is implemented using Google Web Toolkit (GWT) framework which uses AJAX technology for asynchronous communication with server side of the application and allows for relatively easy and powerful scripting on the client side.

3.1 Personalisation types

The system supports two kinds of personalization, technical and semantic, that incur different changes within the system. Technical personalisation addresses issue of changing the graphical interface, e.g. adjusting placement of controls, changing their color, the font being displayed, adding the bounding or changing the size. The semantic analysis of user behavior on the other hand allows to suggest actions to the user, which are involved with application purpose rather than its technical side. The first group of semantic adaptations concerns providing content suggestions, which works like extended version of autocompletion. After entering data in one field other fields are filled in with appropriate (related) data, which is determined based on past behavior. For example, after a user provided a name of an organization that she would like to transfer money to the application may fill in other fields such as amount, address of this organization or transaction description.

We also adapt the functionality of application, i.e. user receives alerts and suggestions about potential actions. Alerts remind about actions that the user might be interested to perform, e.g. transferring money to a certain institution on a given day of the month, if the system discovered such a custom in the past. Suggestions propose to perform actions that are associated with other just performed action based on their co-occurrence in the past.

Guidelines regarding the adaptation of a user interface are expressed in a form of rules. The structure of a rule is independent from the type of personalization being performed. A rule consists of a body (antecedent) and a head (consequent). The body defines conditions required to fire an action defined in the head. The condition may be: a sequence of particular events (possibly interrupted by other events), a set of events (occurring without particular order) or a time-related event. The action may cause one of the effects (i.e. adaptations) in an application: filling given control with data, change of a style of a given control, display of tool tip for a given control, suggestion or alert for action, change of order of controls.

3.2 Rules Form and Lifecycle

Adaptation can take place in response to an event. Event is understood as an elementary manifestation of user behavior in the system. We distinguish three kinds of events, occurring on different levels: program events (fine grained mouse or keyboard events), logical events (change of contents in controls, e.g. typing in account number), semantic events (high level operations triggered by filling in forms, e.g. placing a transfer order).

The program and logical events are generated in the browser, while semantic events occur on the server. Program and logical events have to be transferred to the server for data mining purposes. As written previously, event stream is mined for patterns of user behaviour. However, as we learned, applying raw data mining algorithms bring a lot of noise and does not render useful user behaviour patterns. For example the most frequent associations between events are the following: “when the user **opens** transfer page, he **clicks** «send» button”. Such associations are the side-effect of technical organization of the application and have to be filtered out to find the actual user behaviour patterns. The behaviour pattern are transformed into adaptation rules, which are recorded in the user’s profile and updated with every run of the data mining subsystem. We were able to find from dozen to almost hundred such rules for a single user based on the exemplar data gathered from test user interaction with the e-banking application.

Since the rules are sensitive to program and logical events, which origin at the client, it is possible to move its enactment also to the client. The semantic events are also taken into account because they are associated with logical events that precede them (for example sending a transfer - a semantic event - is not possible without opening a transfer page - a logical event). The rules are transferred to the client upon application loading, where they can be enacted thanks to the scripting capabilities of modern browsers, as described in the next section.

3.3 Client-side Rule Evaluation

We implemented an efficient rule execution environment in the browser, which can be nowadays done conveniently thanks to modern application Web frameworks like GWT. The implementation is based on non-deterministic finite state automaton, which is a very efficient device for capturing multiple patterns at once, via a single pass over a stream of symbols (events in our case).

Each head of the rule passed to the client was added as a pattern to be recognized by the automaton. If the rule expected a sequence of particular events, possibly interleaved with other events, it was added as a sequential pattern. If the rule expected a set of events to occur, all the possible permutations of the event sequences were calculated on the client, and added as sequences with interleaving events to the automaton. Though it might seem inefficient, the system actually did not suffer from the explosion of permutations, since the maximal number of events in the rule body was 6, which gives 120 permutations of sequences only.

After creating the automaton, it was run over the stream of events, as they occurred. To handle the non-determinism of the automaton, a standard approach was taken, where the automaton was allowed to have several active states at once, corresponding to partially matched rules. Each of these active states was tracked independently, and new were activated as needed. Such approach proved to be efficient as expected with respect to computational time, and moderately heavy with respect to memory consumption. Total consumption of memory by the browser process did not exceed 500 MB, which was shared between all the components of the application and is comparable to opening several tabs with fairly standard Web pages.

4 Tests on Efficiency of Personalization

4.1 Research Methodology

In order to check the efficiency of methods we have prepared the scenario for navigating through the application and performing simple e-banking task. The scenario was scripted using iMacros tool, that automates execution of tasks in the client browser - this ensures repetitiveness of the experiment. To simulate the static Web pages we used HtmlUnit. The scenario was executed in several variants on a typical workstation supplied with Inter Core2 Duo, T8300 @ 2.4GHz and 3.5GB of physical memory. For remote tests, 100Mbit LAN was used. Our application implemented in Google Web Toolkit (GWT) ver. 2.0.3 was deployed on Glassfish server (v 3.0.1). As a database we used Postgresql ver. 8.4; application server used the following database library postgresql-8.4-701.jdbc3.jar. Test were executed in Mozilla Firefox.

4.2 Centralized Execution of Personalization

In this variant of the scenario the personalization is prepared on server. We employ HtmlUnit to simulate execution of the business logics on server, while typically it is run on client. The goal of the test was to estimate the mean time necessary to prepare a personalized view of the application on server side as a function of a number of concurrent users. HtmlUnit browser simulator was run on server in several threads; each thread simulated the behavior of one client.

Results. Each thread generated an independent instance of HtmlUnit and it influenced amount of required memory (up to 600MB for 15 concurrent threads). Time necessary to prepare the visualization increased as the number of threads grew. Table 1 presents detailed timing.

Table 1. Duration of scenario depending on the number of concurrent threads.

No of threads	Scenario duration in sec.
1	26,090
3	36,863
5	51,145
10	88,686
15	132,241

The dependency is also presented in Figure 1.

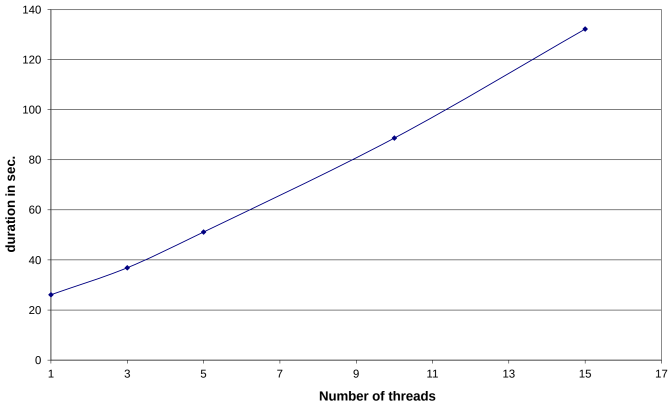


Fig. 1. Mean duration of scenario execution with HtmlUnit depending on the number of concurrent local threads.

Based on the figure we validated the hypothesis that the dependency is linear. Analysis of regression confirmed high dependence between variables. The coefficient of determination R^2, which measures proportion of variability in a data set that is accounted for by the statistical model, was 97,4%. We checked statistical significance of the regression using Fisher-Snedecor (F-test). On the test machine it took 19.51 sec. to prepare one personalized visualization from the scenario, and we need 7.97 sec. more for each additional client. Each test run resulted in almost 100% CPU utilization.

4.3 Distributed Execution of Personalization

In this variant of the scenario the personalization is prepared on clients. We employ web browser with appropriate scripting engine installed, so that execution can be started remotely and server load is investigated. The goal of the test was again to estimate the mean time necessary to prepare a personalized view of the application, but this time on the client side, when several clients connect at the same time. To preserve comparability of the results we again used HtmlUnit to simulate a web browser, but this time it was run on several machines connected in a local 100Mbit network at the same time, and each client executed the same scenario.

Results. Tests were run 8 times (8 sessions), using the following number of clients in respective sessions: 10, 10, 10, 5, 5, 5, 10, and 5. Mean execution times are within a narrow range (86 - 88 sec.), except for the first session (95 sec.) which can be attributed to necessary caching of files. Therefore, in further analysis we distinguish execution with first session (A) and without it (B). Client applications were run on identical machines, nevertheless, we have verified statistically that there is no statistically significant difference between duration of scenarios on different clients. We used univariate ANOVA analysis. For the first option (A) we have: F=0.45, p-value = 0.9 > 0.05; for the second option (B): F=1.19, p-value = 0.33 > 0.05. As p-value is substantially higher than 0.05, there is no ground to reject the hypothesis saying that all durations are equal.

The first step in analysis is the comparison of execution times depending on number of clients: 5 or 10 (two groups). Summary results are in table .

Group	Count	Sum	Avg	Var
‘5’	20	1726.2	86.309	0.451
‘10’	40	3559.2	88.980	14.775

Average execution times on both groups are similar. Analysis of variance, however, proofs that they cannot be assumed equal.

Source of variance	SS	df	MS	F	p-value	Test F
between groups	95.084	1	95.084	9.431	0.0032	4.007
within groups	584.780	58	10.082			
Total	679.864	59				

The F statistics equals F_{obs}=9.431 and is higher than a critical value F=4.007. The p-value = 0.0032 < 0.05, therefore at 95% significance level the null hypothesis saying that average duration times in both groups are identical should be rejected.

To make sure that results were not biased by caching, we verified also the set with 7 sessions (option B).

Group	Count	Sum	Avg	Var
‘5’	20	1726.186	86.309	0.451
‘10’	30	2606.834	86.894	0.385

Paradoxically, although average times do not differ much, ANOVA again pointed out that samples could have not been obtained from the same population. This is because of the small variances, where the tolerance margin is really small.

Source of variance	SS	df	MS	F	p-value	Test F
between groups	4.109	1	4.109	10.001	0.0027	4.043
within groups	19.721	48	0.411			
Total	23.830	49				

The F statistics equals F_{obs}=10.001 and is higher than a critical value F=4.007. The p-value = 0.0027 < 0.05, therefore at 95% significance level the null hypothesis saying that average duration times in both groups are identical should be rejected.

We therefore conducted analysis of regression in order to quantify dependency between number of machines and execution time of scenario.

The following equation was assumed:

$$y_2 = ax + b \qquad (1)$$

where: y_2 – duration of the scenario in seconds (the dependent variable), x – number of clients (the independent variable).

The coefficient of determination R^2 shows that just a fraction of variability of dependent variable was explained by independent variable - only 14%. Nevertheless, the regression is statistically significant (Fisher-Snedecor test F_{obs}=9.43). Values of t-Student statistics confirms that estimation of parameters a and b is also statistically significant. Therefore, we have:

$$y_2 = 0.534 * x + 83.639 \qquad (2)$$

It should be compared to the parameters obtained in the previous experiment which presents execution times for local setting. Figure 2 presents such comparison.

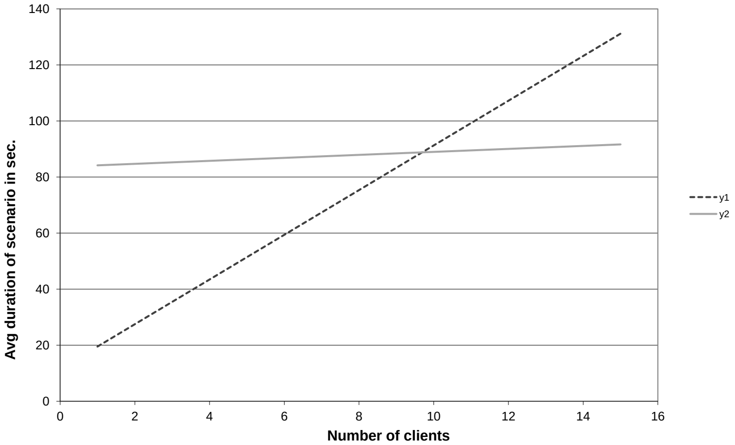


Fig. 2. Comparison of average execution times of scenario using HtmlUnit in local (y_1) and remote (y_2) setting.

An important benefit of distributed variant, where visualization is prepared on client side, over a centralized solution is just slight increase in processing time with increasing number of clients. For hypothetical number of 10,000 clients, which is an expected number of clients using a production grade e-banking system, the execution of the scenario on current hardware configuration would take respectively: 22 hours and 9 minutes for a centralized version and 1 hour and 30 minutes for the distributed version.

Conclusion: for the small number of clients the centralized solution is more efficient. With increasing number of clients a distributed approach is preferred, where the threshold in our experiment was estimated at 10 clients.

4.4 Execution of Personalization in Web Browsers

In the second experiment we used a browser which is not efficient. In this experiment we will focus on real efficiency of the system using a typical web browser instead of artificial interpreter like HtmlUnit. The goal of the test was to estimate the mean time necessary to prepare a personalized view of the application in web browser, when several clients connect to application server at the same time. We used Mozilla Firefox with iMacros to execute the scenario. The clients were run on twenty machines on a local 100Mbit network.

Results. Conducted experiments confirmed the efficiency of JavaScript engines built into web browsers. Execution times of scenario were significantly reduced: average time for a sample of size 72 was 3.66 sec. Shorter times can partly be attributed to more efficient caching mechanism of web browsers than of HtmlUnit. In order to measure the influence of parallelization on duration of scenarios, measurements were done in two variants:

- parallel – all clients connect to server at the same time
- sequential – clients connect independently, with some delays between connections.

The null hypothesis is that execution times in these two variants do not differ. Again, ANOVA analysis was used to verify hypothesis.

Group	Count	Sum	Avg	Var
‘parallel’	17	61.22	3.601	0.547
‘sequential’	55	202.11	3.675	5.376

Average times seem to be identical.

Source of variance	SS	df	MS	F	p-value	Test F
between groups	0.0705	1	0.0705	0.0165	0.8981	3.9778
within groups	299.0323	70	4.2719			
Total	299.1029	71				

The observed statistics F_{obs}=0.0165 is smaller than critical value F=3.9778 and at the same time p-value=0.898 > 0.05, therefore there is no reason to reject the null hypothesis.

The **conclusion** of this experiment is as follows: when using web browsers to connect simultaneously many clients in an experiment environment consisting of 20 machines, the concurrency (parallel vs. sequential) of connections does not affect the execution time of the scenario.

5 Conclusions and Discussion

For the implementation of the web application we used a very efficient Google Web Toolkit engine. As a way to optimize personalization method we proposed moving preparation place of the visualization to clients. This, however, implicated moving information usually available on server (e.g. business logics) to clients as well. Thus, server is not responsible for tasks related to graphical user interface, and merely provides the clients with the data necessary to prepare visualizations locally.

In this paper we verified to which extent the distributed solution excels centralized system. In the first variant, individual forms of corporate banking application and their adaptations were prepared in their entirety on the server. We utilized HtmlUnit, one of few possibilities to generate HTML pages on the server without modification of the web application. In the second variant, the GUI was generated on the client side. We observed certain overheads attributed to the transfer over the network and just when a number of clients exceeded ten, the distributed solution was more efficient than the centralized one. In the third variant, the web browsers were used as a client to generate the visualization and this solution was efficient as expected. The average time to execute the scenario was 3.6 sec. and was not significantly higher when 20 machines connected at the same time. The estimated time to execute the same scenario for 20 clients in the first variant is 171 sec. (47.5 times worst), and in the second variant – 94 sec. (26 times worst).

References

1. Adomavicius, G., Tuzhilin, A.: Personalization technologies: a process-oriented perspective. Communications of the ACM 48(10), 83–90 (2005)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487–499. VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994), `http://dl.acm.org/citation.cfm?id=645920.672836`
3. Atterer, R., Wnuk, M., Schmidt, A.: Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: Proceedings of the 15th international conference on World Wide Web. pp. 203–212. WWW '06, ACM, New York, NY, USA (2006), `http://doi.acm.org/10.1145/1135777.1135811`
4. Brusilovsky, P.: Adaptive Hypertext and Hypermedia. No. ISBN 978-0-7923-4843-6, Springer (1998)
5. De Virgilio, R., Torlone, R., Houben, G.J.: A rule-based approach to content delivery adaptation in web information systems. In: Proceedings of the 7th International Conference on Mobile Data Management. pp. 21–. MDM '06, IEEE Computer Society, Washington, DC, USA (2006), `http://dx.doi.org/10.1109/MDM.2006.16`
6. Gao, C., Wei, J., Xu, C., Cheung, S.C.: Sequential event pattern based context-aware adaptation. In: Proceedings of the Second Asia-Pacific Symposium on Internetware. pp. 3:1–3:8. Internetware '10, ACM, New York, NY, USA (2010), `http://doi.acm.org/10.1145/2020723.2020726`
7. Malinowski, U.: Adaptive user interfaces : principles and practice. Elsevier (1993), iSBN 978-0-444-81545-3

8. Mueller, F., Lockerd, A.: Cheese: tracking mouse movement activity on websites, a tool for user modeling. In: CHI '01 extended abstracts on Human factors in computing systems. pp. 279–280. CHI EA '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/634067.634233
9. Paskalev, P.: Rule based gui modification and adaptation. In: Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing. pp. 93:1–93:7. CompSysTech '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1731740.1731841
10. Paskalev, P., Serafimova, I.: Rule based framework for intelligent gui adaptation. In: Proceedings of the 12th International Conference on Computer Systems and Technologies. pp. 101–108. CompSysTech '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2023607.2023626
11. Wang, H., Mehta, R., Supakkul, S., Chung, L.: Rule-based context-aware adaptation using a goal-oriented ontology. In: Proceedings of the 2011 international workshop on Situation activity & goal awareness. pp. 67–76. SAGAware '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2030045.2030061

7.3 Profiling Visitors of Shopping Malls: the Meaning of Paths. Analyzing Mobility Patterns Based on CDR Data

7.3.1 Introduction

Nowadays, companies try to learn who are their customers to better suit their needs. These companies use diverse methods to collect data on customers including surveys, responses to direct marketing, loyalty programmes, etc. Data obtained these ways needs to be analysed to have value for a business entity. Such analysis may concern customer profiling and segmentation, prediction of the next best offer or a product for cross-selling. This section covers the first type of the analysis, namely profiling and segmentation.

A user profile is a set of characteristics of a user, including user related rules, settings, needs, interests, behaviours and preferences [30]. This information may be static e.g. native country or dynamic e.g. needs or preferences that change in time. Profiling of customers targets both of these groups of characteristics. It is important to know gender, home location or age of a customer. However, more important is to know his/her behaviour i.e. the dynamic profile. This research focuses on a dynamism (customer journeys) of a customer and his/her needs.

Understanding customer journeys has become one of the most important topics in the recent marketing literature [136]. It is undoubtedly crucial for companies to build informative insight about their customers and understand their holistic experience. In fact, Lemon and Verhoef conceptualized customer experience as "a customer's journey with a firm over time during the purchase cycle across multiple touchpoints" [97, p. 6]. Better understanding of customers can help companies to improve effectiveness of marketing through better design of customer touchpoints [136] and increase their behavioural response to promotional messages [176]. Up to date most research concerning customer journeys was done in a closed environment, i.e. with respect to a single brand or a single place (e.g. shopping mall). Recently many authors called for enhancing the perspective of customer experience and encouraged using larger scope for customer journey mapping [111, 121, 166]. We answer this call by modelling customer journey in between different shopping malls within one city (Poznań) in Poland.

The research goal is to profile customers of shopping malls based on Call Detail Records to learn customer preferences and improve marketing campaigns. The proposed method which utilizes mobility profiling to enable such an analysis is presented. The secondary research goal is to demonstrate that using data collected for a different purpose (billing of users of telecommunication services), it is possible to derive conclusions valuable for companies and various scenarios. This in turn would provide an example of directly utilizing profiling methods for building new services and extending the companies insight about their customers.

The section is structured as follows: subsection 7.3.2 presents related work regarding both profiling of customer journeys and Call Detail Record analysis. Following section presents the concept of the method for describing customer journey using CDR data. Then, the dataset is introduced and briefly discussed. Section 7.3.5 presents results achieved applying the method on the dataset. The section is summarised in conclusions[3].

7.3.2 Related Work

Mapping a Customer Journey

Customer journey mapping based on shopping trips is a known method that was utilized by companies to provide them with an insight on the localisation of their retail outlets. The first models of shopping trips were based on a gravity model. The assumption of the gravity model is that the probability to choose a certain destination for shopping decreases with a distance to the destination and increases with the size (attraction) of the shopping centre [77]. Since that time, shopping malls changed their meaning from a place where people buy goods towards a meeting or leisure place.

[124] defined shopping malls as closed, climate-controlled, lighted shopping centres with retail stores on one or both sides of an enclosed walkway. Apart from utilitarian benefits (e.g. parking), the atmosphere (or ambience) and other elements of customer experience, impact the choice of the shopping mall and the subsequent consumer behaviour. In contemporary retail, the overall experience generated out of tenant mix, facilities and atmosphere, stands for the competitive advantage.

Therefore, if a shopping centre is of diverse nature, it becomes more important to learn what makes a customer travel between different shopping malls. Introduction of further travel utility [92] led to development of a multi-purpose shopping trip model [9]. Hu and Jasper investigated consumer experiences on shopping malls and found convenience, choice, crowds, ambiance,

[3] Results described in this section are joint work with Piotr Kałużny (Poznan University of Economics and Business), Piotr Jankowiak (Poznan University of Economics and Business) and Piotr Kwiatek (American University of Middle East, Kuwait). The paper describing these results is currently being prepared as a journal submission.

parking and hedonic shopping orientation to be the most impactful [74]. In contrast, shopping mall choice studied by [174] uncovered assortment, atmosphere, convenience and quality to have the most impact. Other models were developed based on utility maximizing behaviour, choice heuristics and Markov chains, cellular automata and complexity theory [158, p.142-143].

This research focuses on loyalty of customers and their behaviour derived based on Call Detail Records.

Using CDR Data to Describe a Customer Journey

Nowadays, telecommunication service providers consider monetization of the big data their networks provide. As an example for this, ATT has plans to gain profit from selling aggregated user data to advertisers and marketing companies [13]. Similarly, Orange is participating in the ”Traffic Zen” experiment to create traffic forecasts. Telefonica’s Brazilian subsidiary Vivo is using cell data, demographics and predictive modelling to provide customers with services ”tailored for them” – created to match their needs and taking into consideration their lifestyle [79].

A detailed source of data on users of telecommunication services are Call Detail Records (CDRs). CDRs contain details of every billable action made using a phone. Actions included in CDRs concern calls, text messages, data transfer or usage of a specific service. Each action is described by s set of data i.e. date when an action took place, its duration, location of a BTS service, an initiator/target of an action, etc. Companies are obliged by law to keep their record (with the period depending on the country). There are many potential uses for CDRs. Previously preprocessed data can have an impact on various areas and can be used in multiple scenarios [19], including: urban planning [86, 171], tourism [4, 44], epidemiology, measuring carbon dioxide emission and traffic volume [15, 70].

Among multiple models of large and small scale profiling using CDRs, the literature provides only a few examples connected with uncovering physical customer paths. [16] used CDR data from Telefonica to model visiting patterns of mall visitors in Santiago de Chile. Their study targeted 16 shopping malls and over 1 million of unique devices that appeared in a neighbourhood of these malls. Visitors of the malls were assigned the Human Development Index of areas of their home locations. The authors proved that the choice of the mall is determined by two main factors: distance and the socio-economic level e.g. people from peripheral, poor areas tend to visit their local malls and do not travel much. The authors however did not perform analysis on flow of people between two or more malls, rather focusing on a single-mall analysis.

[91] researched daily lives of over a hundred thousand people in and around Los Angeles, using data from WeFI application monitoring location, data connections, application usage, etc. The data produced by the application is more dense, richer and of different nature than data contained in CDRs. Authors also augmented the data with census data based on home location of mobile phone users. Analysing the collected data they proved inter alia that female shoppers are seen at malls almost three times per week, while the males appear there less than twice a week. However, still no behaviour regarding customer journey was reported.

[26] studied how Call Detail Record (CDR) can be used to better understand the travel patterns of visitors. They developed Origin Destination Interactive Maps to map transportation information of tourists contained in CDRs. The goal of this study, carried out for Andorra, was to understand how tourists moved through the country to determine how connected cities were to one another in various months and holidays (and for different groups of tourists). They studied behaviour and paths, however using different tools and on a different level of granularity regarding visited locations.

This research uses Call Detail Records for studying behaviour of shopping malls' customers understood as customer journeys. The research targets citizens of a big Polish city (Poznań) and describes patterns on how malls within and outside the city are visited by customers.

7.3.3 Method of Analysing Customer Journeys

Analysis of Mobility using Call Detail Records

Geospatial information within CDRs can help uncovering previously unknown behaviour and meaningful patterns of user activity for a single user and on more aggregated level. CDR data includes information regarding locations in a form of the nearest BTS station (Base Transceiver Station) a user was connecting to while using telecommunication services. The accuracy of this information is limited due to technical issues to an average radius of 3 square kilometres, but considering urbanized areas (cities) a few hundred meters accuracy is possible [5, 70]. Based on CDR data e.g. home and work locations may be calculated with a precision of a district [19].

More complex is the analysis of user's mobility. Based on CDR data, daily travel paths of people may be identified [53], including travel on small distances (about 5 to 10 kilometres) on a daily basis [147]. Based in these paths, it is possible to build a user's mobility profile working in a small scale, focused on estimating user's visited locations and use it effectively [100]. This profile can be an understandable representation of a typical user movement usable in inferring commuting/movement statistics over the large area [44]. When using logs for the analysis of user mobility only very brief moments of his whereabouts are known (connected to the calls or other services performed that were handled by BTS). This estimation of location is not ideal, but its accuracy can be measured based on the density of cell towers.

7.3.4 Description of the Dataset

The CDR data used for this work consists of more than 7 billion records describing activities of Orange clients in Poland for over 6 months between February and July 2013. Each record consists of:

- an anonymised identifier of a user initiating the call (being the client of Orange),
- a type of a service (call, SMS, Internet use) along with its duration in seconds,
- a time stamp,
- a BTS station (location where the action was initiated),
- a location_id (grouping of a set of BTS stations that share the same coordinates).

From about 10 million of users represented in the dataset, a sample of about 100 thousand of people was created. The sampling procedure was as follows:

1. Firstly, all users who appeared in one of 173 location_ids in Poznan were selected. These location_ids were chosen based on their coordinates compared with GEO JSON from the Open Street Map indicating the boundaries of the city of Poznań. This way the data was limited to 766 948 690 events from 408 058 users.
2. Then, 408 058 users were assigned with their home and work locations:
   - home location is a place, where a user spends most of his nights and the time after work (in this case between 7PM and 8 AM),
   - work location is a place which a user tends to visit during the day, that is not his home location. To conduct the labelling, locations a user visited on working days (Monday to Friday) between 8AM and 6 PM were identified, and as a work location the one with the most stay time spent was selected.

   104 890 users had their home location in Poznan. Figure 7.3 presents home locations for all users who had at least one action assigned to one of BTS located in Poznan.
3. In the next step, call centers were excluded from the analysis. A call center was defined as a user having more than 24 hours of a daily activity (what is impossible for a single person). Such assumption resulted in a set of 123 'machines' that were excluded from the set of 104

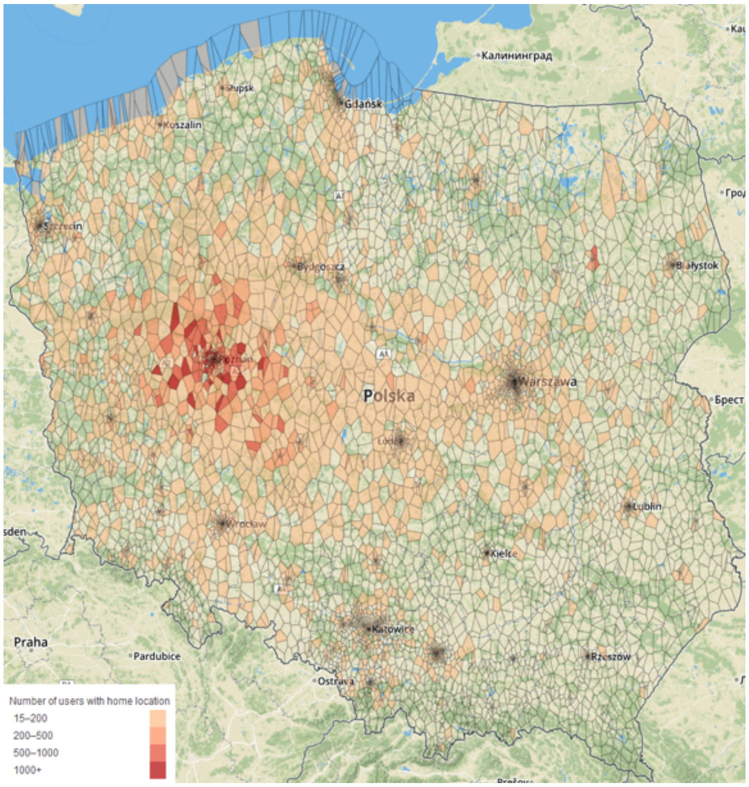


Figure 7.3: Distribution of home locations for a sample of all users that passed by Poznan area during the extent of the study. Source: own study

890 users.

4. Finally, 12 196 users, who had a home or work location in a spot (Voronoi cell[4]) assigned to one of the analysed shopping centres, were identified and excluded from the analysis regarding the number of visits and journeys between shopping centres. This was done in order not to skew the customer visits' results distribution for the users that spend most of their time close to the shopping malls (the section presents also some insights into the distribution of users before this step was applied on the data).

Details of the dataset are presented in Tables 7.1 and 7.2. Home accuracy (Table 7.2) is understood as the ratio of the amount of time spent at home during home hours (7PM - 8AM) relative to other locations in this time range. Similarly, work accuracy is the ratio of the amount of time spent at work during work hours (8AM - 6PM) relative to other locations in this time range).

[4] Based on BTS locations, a Voronoi diagram was created that divided the area of Poznań into rectangles. Please see Figure 7.3.

Table 7.1: Statistics of the dataset after processing of CDRs (Part 1).

Statistics	**Number of active days**	**Number of activities**	**Average number of visited locations**	**Unique BTS stations**
1st Quartile	118,0	446	1,381	14,0
Median	158,0	963	1,866	31,0
Mean	138,9	1547	2,181	43,29
3rd Quartile	173,0	1966	2,616	57,0

Source: own study

Table 7.2: Statistics of the dataset after processing of CDRs (Part 2).

Statistics	**Average max. daily distance**	**Home accuracy**	**Work accuracy**
1st Quartile	1,623	0,42711	0,2530
Median	6,090	0,60090	0,4325
Mean	15,218	0,61586	0,4620
3rd Quartile	16,051	0,80966	0,6642

Source: own study

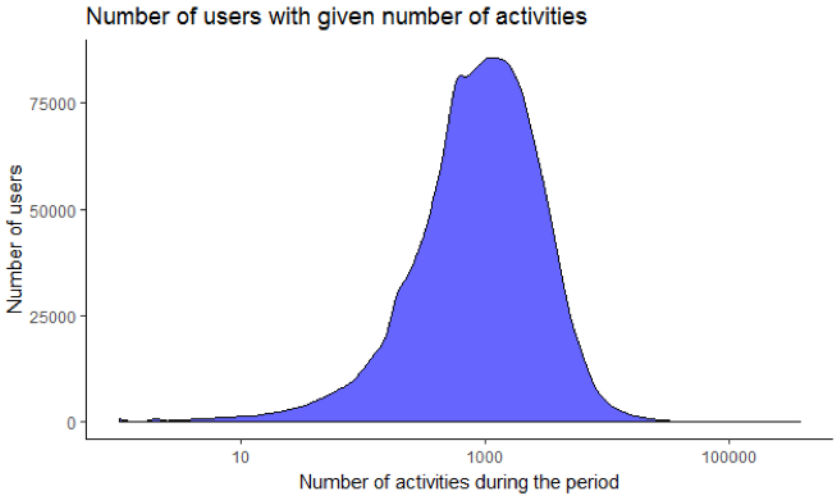


Figure 7.4: A number of activities of users from the sample (logarithmic scale - data is long tailed). Source: own study

Figure 7.4 presents distribution of the number of activities for people having a home location in Poznan. It is worth to mention, that the more activities, the better as it is possible to precisely describe customer paths (journeys).

Figure 7.5 presents a histogram showcasing the average daily distance travelled by users represented in the sample. Most of users do not commute on long distances and rather travel work/home within the city.

Figure 7.6 presents a histogram showcasing an average number of locations visited daily by users from the sample. It can be observed that the sample travels not only short distances, but also visits not more than 5 distinct locations daily.

This data was subject to analysis of customer journeys between shopping centres. Shopping centres for which the analysis was carried out are located in Poznań and include:

- Galeria Malta,
- Galeria MM,
- Green Point,
- Ikea Franowo,
- King Cross,
- Kupiec Poznanski,
- M1,
- Panorama,

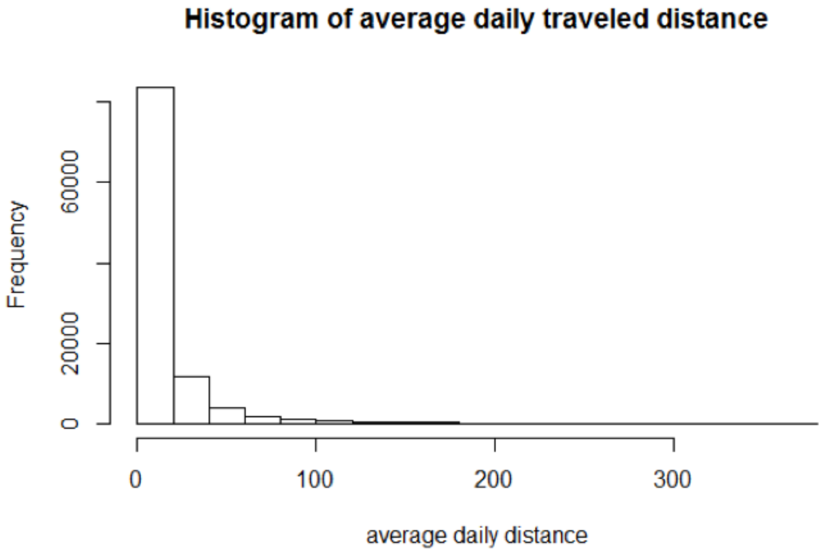


Figure 7.5: Users and their average travel daily distance (104k users from the sample of Poznan inhabitants). Source: own study

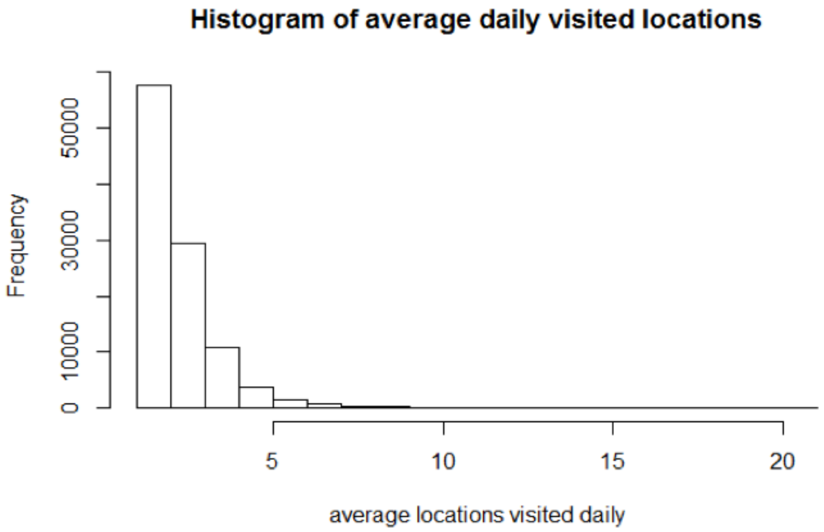


Figure 7.6: Users and their average number of locations visited daily (104k users from the sample of Poznan inhabitants). Source: own study

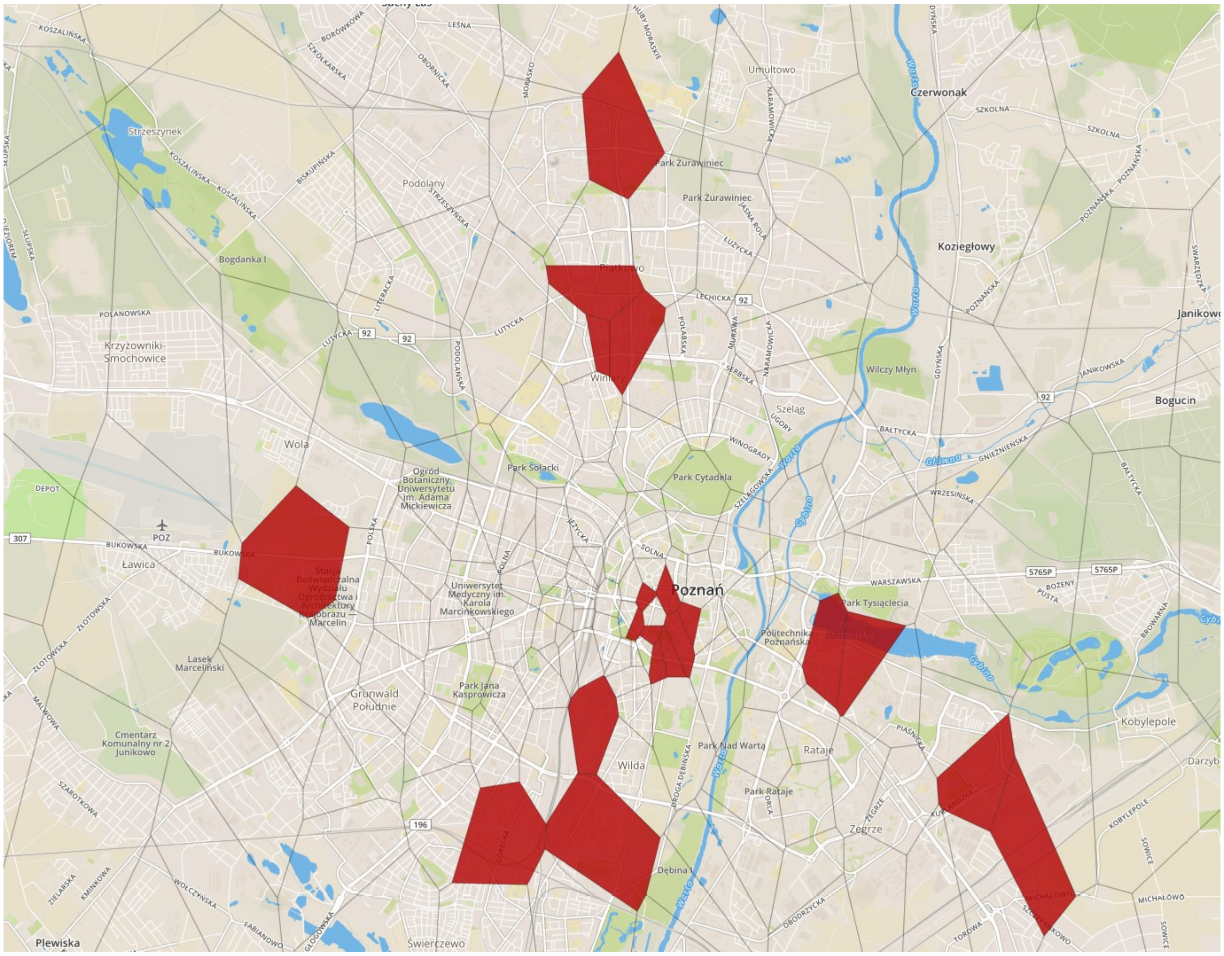


Figure 7.7: Locations of the shopping malls subject to analysis of customer paths. Source: own study

- Pasaz Rozowy,
- Pasaż Tesco os. Batorego,
- Plaza and Pestka (as these two shopping centres are situated one next to another and are covered by the same BTS stations),
- Stary Browar.

Figure 7.7 presents locations of these shopping centres in Poznań. It may be easily noticed that shopping malls are spread around Poznań city and if a customer visits two different shopping centres, he will be identified in two different locations.

7.3.5 Research Results

Firstly, in our research we aggregated data of users to see what shopping malls they visit and on which days. Figure 7.8 presents a number of visitors of chosen shopping centres. These numbers differ from a real number of visitors as not all people perform an action (phone call, sms) while being in the shopping mall and also we observe only the population of Orange subscribers[5]. It may be noticed that the number of visits of customers in shopping malls is stable during the weekdays and dropping on weekends (by a visit we define a stay at a shopping centre longer than 15 minutes). It should be highlighted that in our approach the estimated stay time from the trajectory model plays a big role. Due to the model assumptions, filtering visits with a significant stay time ($>$ 15 minutes) changes the structure of visits to the one more closely resembling real patterns. Based on multiple uses of the model in the literature, filtering single - accidental or insignificant activities (in the model being described by a low stay time estimated), leaves only visits that were intentional. Filtering data further (with a minimal stay time of 30 minutes) significantly limits the size of the sample of users that were observed, but could indicate users that spend considerably high portions of their time shopping. On the other hand utilizing all of the activities (even with small stay time) gives us a possibility to uncover situations where users also just passed by the location, which could be interesting for marketing purposes and indicate users who had the possibility to visit a shopping centre (meaning they were in the vicinity).

Analysing Figure 7.8, it needs to be mentioned that Green Point includes both: a shopping mall and company offices that do not open at weekends. This observation made us also to exclude people having home and work location in the same location as the shopping mall.

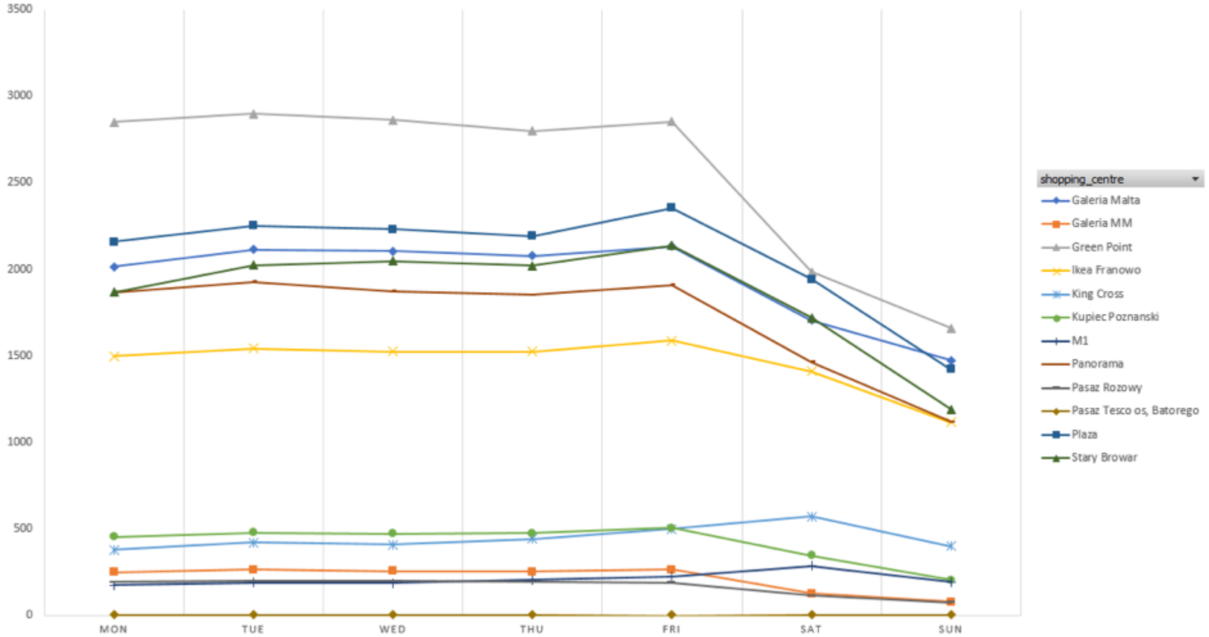
After excluding from the sample inhabitants having home or work location in the same location as the shopping mall, some changes in the numbers of visitors may be noticed (please see Figure 7.9). The number of visitors in Green Point decreases (as people having work location there are excluded from the sample), but popular shopping malls such as Plaza, Stary Browar and Malta seem to have a significantly bigger group of visitors than other shopping centres.

Figure 7.10 presents hourly distribution of visits in each shopping centre. It may be observed that malls in suburbs have different hours of activity than the ones situated in the city centre.

The total number of users from the sample who visited each of the researched shopping malls at least once is presented in Table 7.3.

Before discussing customer journeys between shopping malls, we may study from which area

[5] Orange SA in Poland has about 30% of all subscriptions.



Day	Galeria Malta	Galeria MM	Green Point	Ikea Franowo	King Cross	Kupiec Poznanski	M1	Panorama	Pasaz Rozowy	Plaza	Stary Browar
Mon	973	99	1335	696	204	194	85	909	79	972	988
Tue	1017	103	1345	725	226	198	88	937	84	1009	1060
Wed	1006	99	1331	712	227	193	87	909	81	997	1071
Thu	1010	99	1294	712	246	195	106	896	79	990	1052
Fri	1059	103	1339	752	275	210	111	943	79	1084	1139
Sat	791	40	874	611	293	127	127	650	42	836	857
Sun	631	25	707	465	209	70	88	464	27	571	555

Figure 7.8: Average number of users per day in chosen shopping malls (104k of users). Source: own study

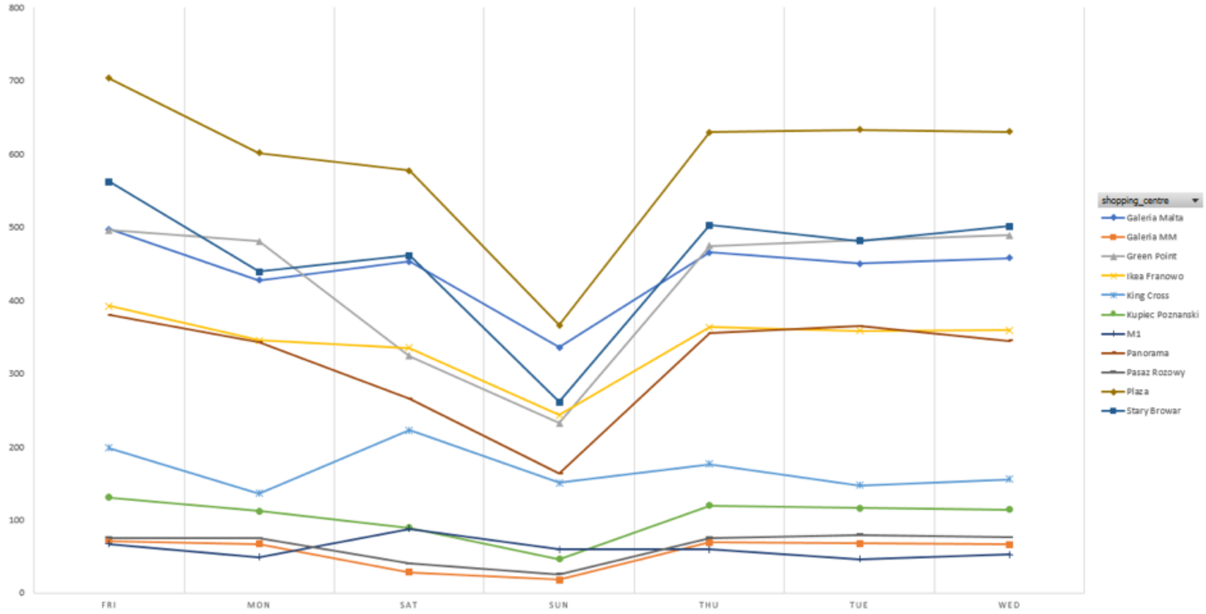


Figure 7.9: Average number of users per day in chosen shopping malls (92.5k of users). Source: own study

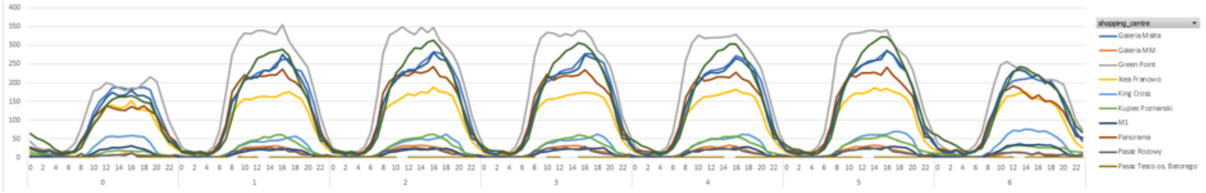


Figure 7.10: Average number of users per hour for each day (104k of users). Source: own study

Table 7.3: Unique users visiting shopping malls between February and July 2013.

No.	Shopping center	Number of visits
1	Plaza (+ Pestka)	27204
2	Stary Browar	23464
3	Galeria Malta	22323
4	Green Point	20952
5	Ikea Franowo	18808
6	Panorama	15508
7	King Cross	11136
8	Kupiec Poznanski	8929
9	M1	5321
10	Galeria MM	5210
11	Pasaz Rozowy	3054

Source: own study

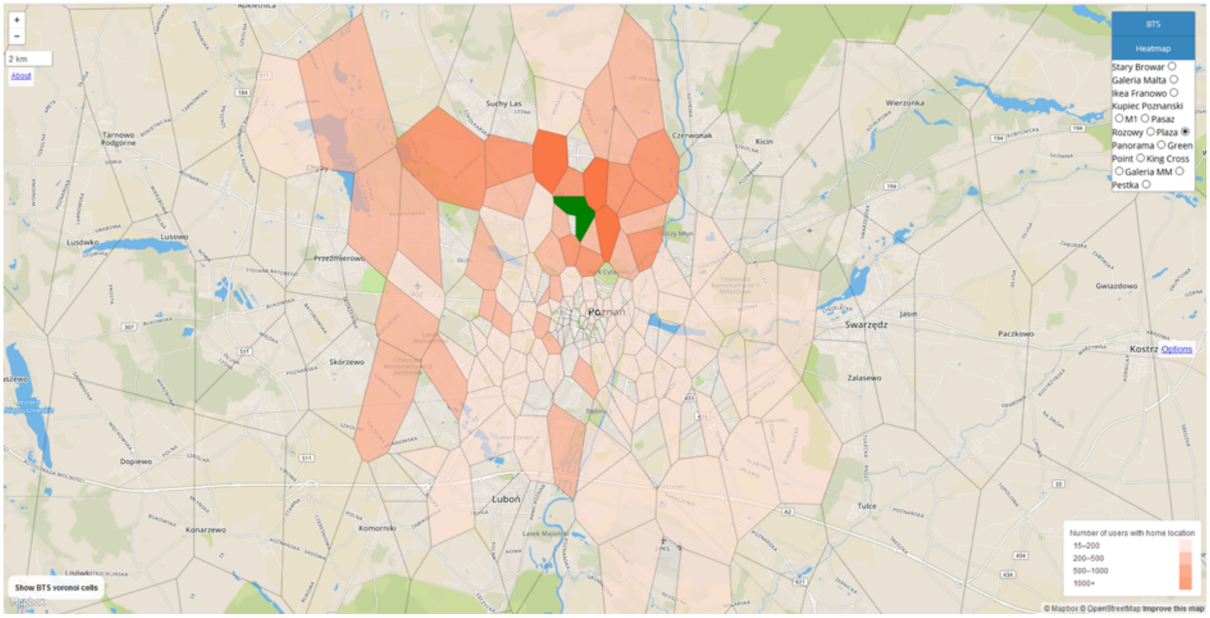


Figure 7.11: Home location of customers visiting Plaza shopping mall. Source: own study

the customers come. Figure 7.11 shows where do people visiting Plaza shopping centre live (the darker the colour, the more visitors from the spot). It should be noticed that the mall is mostly visited by the people living in a close distance, however also people from the north of Poznan (suburbs) come to visit the place. Plaza is however rarely visited by people from the south of Poznan, where other shopping centres are present. This may be due to the fact that there is no speciality of the shopping mall, and customers may find similar goods also in other malls. Similar analysis is possible for every other location.

The main goal of this paper was to study customer journeys between different shopping malls, so we studied how many of other shopping centres were visited by a customer of a given mall. Table 7.4 presents statistics on loyalty of customers of a given shopping centre. It may be noticed that the most loyal are the customers of Pestka and Plaza (who visit only about 2 other shopping malls), and the least loyal are people visiting M1.

Finally, we studied customer journeys (migrations) between different shopping centres and we figured out that there are two types of customers. One group includes visitors travelling from one mall to the other e.g. searching for a certain product, looking for occasions, etc. Second group concerns people being more loyal and visiting usually only one shopping centre (also these visits are more frequent, with a short average duration and small variation of duration). We visualised these migrations on chord diagrams (Figures 7.12 and 7.13). These diagrams show

Table 7.4: Average number of other shopping malls a visitor of a particular shopping centre visits.

Shopping centre	Number of other shopping centres a visitor of a particular shopping centre visits
M1	3,4587
Galeria MM	3,3601
Kupiec Poznanski	3,3391
Pasaz Rozowy	3,0917
Ikea Franowo	2,8135
Galeria Malta	2,7612
Stary Browar	2,7514
Panorama	2,7011
Green Point	2,6167
King Cross	2,5868
Plaza	2,116

Source: own study

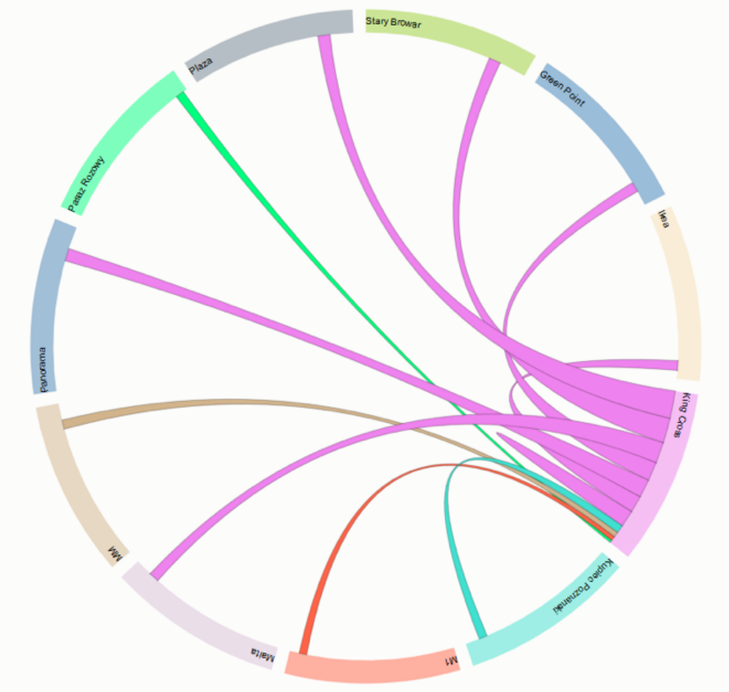


Figure 7.12: Chord diagram for customers visiting King Cross. Source: own study

what percentage of customers migrates between shopping centres.

Figure 7.12 presents flow of King Cross shopping centre customers. The colour is derived from the relation between incoming and outgoing customers. In case of King Cross, many customers tend to visit also other malls. What is worrying is that the percentage of customers coming from other shopping malls to King Cross is smaller than the percentage of King Cross customers who visit also other shopping malls.

Figure 7.13 shows the same analysis for Plaza. However, in case of Plaza the analysis has more colours and presents other situation. Simplifying, all other shopping centres share more users with Plaza than Plaza malls is giving to them (percentage-wise). Moreover, a lot of users visits only Plaza (gray area in the Figure).

This analysis may be further extended, however even the initial insights show us diversified profile of customers of different shopping malls in Poznań area.

7.3.6 Summary and Conclusions

In this section we presented an initial analysis of customer journeys between different shopping malls for the city of Poznan, Poland. The analysis carried out was based on data that was not collected for the needs of such analysis, but still enabled to draw some important conclusions,

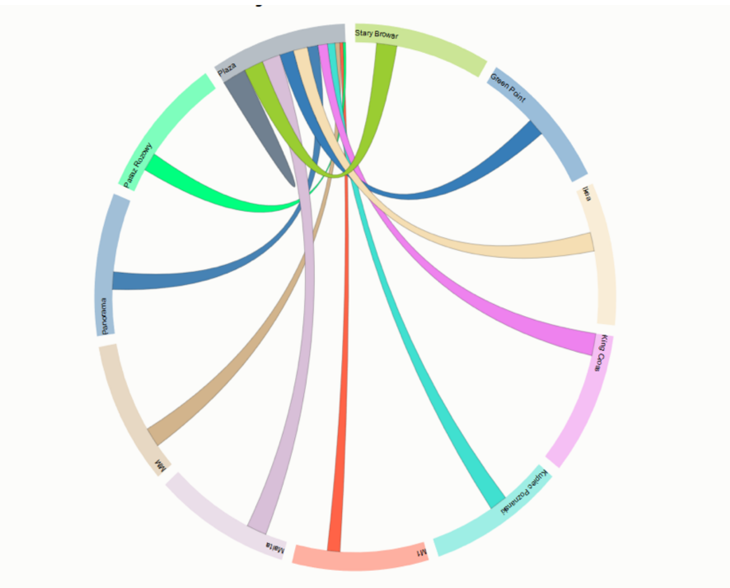


Figure 7.13: Migration for Plaza. Source: own study

namely:

1. It is possible to describe who are the customers of a shopping mall. In our case we were able to identify home/work location of the customers, but this data could be associated with census data for a given home/work area. Having more detailed profiles of customers, we may enrich segmentation and influence pricing or marketing campaigns. Additionally, utilizing mobility profiling can give us information about the areas those customers are travelling through daily. This may in turn prove valuable for marketing campaigns.
2. It is possible to define the geographical coverage of a shopping mall. Having geographical profile of a mall, we may target only selected group of customers (location-based) or focus on influencing customers also from other areas.
3. There are different loyalty levels concerning shopping malls. We identified shopping malls visited frequently by the same group of people and shopping malls which are addressed by customers as one of many spots.
4. We identified flows between shopping centres (which shopping malls share the same group of customers and which have their own visitors). This research was carried out on data covering all subscribers of Orange visible in shopping centres from the area of Poznan within half a year (in 2013).

Our approach shows that some important insights may be drawn from data sources that are not collected for a given purpose. Still, such results may have a significant business value. In scenarios concerning customer journeys, the second phase would be to study the motivation of customers to migrate between the shopping centres. Such analysis could be the future work in this domain. It is important to note however, that we may learn who is our customer without the need of surveying a customer in a traditional way.

7.4 Conclusions

The chapter aimed at development of a horizontal application scenario for profiling, targeting personalisation. Its goal was to ”propose profiling methods that enable to describe a user in a way enabling for personalisation of content or a service”. This goal was further translated into two secondary goals, namely:

G6.1 To propose a personalisation method that improves user experience of a solution, but does not impact the efficiency of the solution,

G6.2 To provide a user profiling method that enables to obtain valuable insights from data that was not collected for a specific purpose.

In relation to the goal G6.1, the goal of the paper was to propose and evaluate the efficiency of a method enabling a client-side adaptation of a Web interface. The paper included in Section 7.2 studies personalisation of Web interfaces including technical and semantic aspects (elements of the interface as well as content). The scenario studied concerns a banking system available online. The technical adaptations (personalisation) concern changing a graphical user interface, whereas semantic personalisation concerns analysing user behaviour and history of his/her transactions to suggest actions to the user. Ten, two approaches are studied (centralised and distributed computing) to perform personalisation and check challenges and limitations. In the paper, it is proved there is a threshold concerning the number of users below which the centralised approach works more effective, however if a large number of users is expected, the distributed approach is better for performing personalisation.

The G6.2 secondary goal was translated in the respective section as to: ”profile customers of shopping malls based on Call Detail Records to learn customer preferences and improve marketing campaigns”. Section 7.3 concerns studying journeys of customers of shopping malls. The experimental data concern city of Poznań and is expressed in the form of Call Detail Records. The loyalty of customers (and their preferences towards choosing a shopping center) is analysed. The paper presents a method of analysis that offers more knowledge on a customer than traditionally possessed by shopping malls. The results also enable to deliver new business models for telecommunication companies as data used in analysis was collected for billing purposes, but still has a huge value for other companies (it is worth to underline that this is aggregated data, so GDPR issues are not impacted).

Chapter 8

Summary

This chapter is to summarise the thesis and indicate what are the results achieved in relation to state of the art and goals envisioned. It needs to be underlined that research described in this document has been carried out for the last 8 years and therefore some of these results may not appear to be novel from today's perspective, but still remain significant. Personalisation and profiling are continuously a subject of numerous research projects and are of interest of many researchers and practitioners from multiple domains (including horizontal and vertical application scenarios).

The remainder of this chapter is structured as follows. Firstly goals of the thesis are reminded. Then, the section on contribution to state of the art follows. The contribution discusses results in relation to the goals stated and also shows which sections especially focus on achieving these goals. Finally, some insights into the future work are presented.

8.1 Goals

The research described in the thesis concerns profiling of people and things. The domain targeted is very broad and may be perceived from various perspectives. We tried to address only three of them, namely:

- Background and definition: the document presents definitions of profile and profiling as well as provides insights into methods that may be used while profiling people and things. When discussing approaches, also diverse data sources that may be used for profiling are studied e.g. regarding emotions or colour preferences (chapter 2 and 3).
- Vertical applications: chapters 4 and 5 present applications of profiling for public utilities

represented by domains of smart grid and telecommunication. Both of these domains define their own challenges for profiling, but it is demonstrated that they may benefit from similar methods. Nowadays, business utilises only some basic quantitative approaches to profiling, however enabling reasoning on relations between features as well as between entities may enable addressing new business challenges.

- Horizontal applications: profiling may be successfully used also for a broadly understood personalisation and authentication of users. The more specific the profile, the more aligned the recommendation, and this may directly influence e.g. sales. On the other hand, detailed profile enables quick identification of anomalies and therefore may be used for authentication purposes. Chapters 6 and 7 concern these issues and present how profiling may be used across domains, proving its usefulness.

The goals that were to be achieved within this thesis are summarised in Table 8.1.

Table 8.1: Goals addressed in the thesis.

Perspective	**Goals**	**Reference**
Background and definition	The goal is to define a profile of a person or a thing and identify features of a person or a thing that may be represented in a profile and are usable for diverse application scenarios.	B1
	The goal is to analyse profiling methods that enable for describing a user/a thing or relations between users.	B2
Vertical applications	The goal is to create a profile of a user or a thing that will be applicable for solutions enabling management of production and consumption of the electric energy in the smart grid.	V1
	The goal is to develop a profile of a customer/subscriber to telecommunication services, enabling for personalisation and taking into account the issues of privacy and trust.	V2

Table 8.1: Goals addressed in the thesis.

Perspective	**Goals**	**Reference**
Horizontal applications	The goal is to develop a method allowing authentication of a user based on Call Detail Record data.	H1
	The goal is to propose profiling methods that enable to describe a user in a way supporting personalisation of content or a service.	H2

Source: own study

8.2 Contribution over State of the Art

This section is to demonstrate advantage over state of the art of research described within this document. As prevailing part of sections of this thesis includes publications published in journals or at international conferences after a peer review process, the quality of specific chapters has already been confirmed by a significant group of reviewers. However, it is still worth to show the results and their importance for the domain, especially in relation to the previously stated goals. Table 8.2 presents contribution of the publications included in this thesis in relation to the previously presented goals.

Table 8.2: Research contribution.

Reference	**Contribution**
B1	The outcomes concern: analysis of the related work in the area of managing user data e.g. identities, profiles, etc.; proposing identity lifecycle (starting from creation, enabling its updates, merging different profiles, update by a user, querying, controlling access by different services, etc.) and proposal of an architecture of a system that enables providing a user with an invisible personalisation support while browsing the content (the system is to be working in the background). The system delivered was developed within the EGO - Virtual Identity project and is a working solution.

B1	The results concern a detailed review of state of the art in the domain of identity management. The notion of a digital identity was analysed from different perspectives and various definitions were provided. The paper also proposes a set of use cases for identity management systems. The use cases focus on improving user experience while utilising data included in the virtual identity. Then, a comparison of selected identity protocols, projects and initiatives taking into account the proposed use cases follows. This issue is further addressed in the Bachelor thesis of Adam Maćkowiak, which Agata Filipowska supervised, that concerned utilisation of a virtual identity based on previous activities of a user on the Web or his profiles on social media, in the process of advising financial instruments.
B1	The paper provides a method for building a profile (identity) of a user based on his activities on the Web. The profile is described using categories from Wikipedia. A method that was presented in the paper is developed for the Polish language and is based on keywords derived from papers mapped on Wikipedia categories that may be included in the profile/identity. The paper presents also a potential extension to DBpedia showing potential of profiling using Linked Open Data.
B2	The results described are based on data from a survey carried out among 144 respondents. Profiling concerned both: modelling personality traits of respondents and their colour preferences. Using regression and Apriori methods, links between BFI-44 personality and colours were identified. Relations between results achieved using two different methods were presented and analysed (there are colours important for a given personality type identified by two methods, but there are also some differences between outcomes received by application of these methods). This research shows also that given the same set of data, outcomes received by two profiling algorithms may significantly differ.

B2	The paper, addressing the goal, provides analysis of the related work in the area of modelling relations between users. Different types of relations are presented and typical approaches on describing these relations in social networks are discussed. Then, based on a survey carried out among 306 respondents, it was confirmed what features emerging from call logs should be taken into account while working on a description of a relation. The goal was not only to model strength of a relation, but to make this feature reflecting the real relation between two individuals.
V1	The paper studied challenges for the microgrid and proposed requirements for a system addressing these challenges. Following, a concept and an architecture of a system, enabling not only management, but also forecasting of production and usage of electric energy, was proposed (that supports also acquisition of data from the Web to enable for improved reasoning). The system was developed and is currently in operation.
V1	The paper explains definition of a profile and a stereotype in the microgrid. Then, a method for demand estimation taking into account features of a prosumer, including appliances used (profiles of things), is proposed. The application of the method in the smart grid management software is also demonstrated. The method was also implemented in the Future Energy Management System.
V2	The paper provides description of the Personal Information Management platform and depicts methods, algorithms and tools allowing different usages based on a social context of a user. The Social Gardening application is presented to demonstrate a potential application of the developed concept. The research goal was similar to the one addressed by the EGO - Virtual Identity project, however the focus was on proposing new services benefiting from a user profile, taking into account the aspects of privacy and trust.

V2	The paper presents insights to what extent CDR data may be useful to derive profiles of users, including relations these users have with their community. The statistics that may describe users based on CDR data are studied. The paper proposes also a concept of the Social Connector application that focuses on presenting the history of contacts of a subscriber (to propose user with actions or make him/her understand the relations with other users). An important part of the paper is devoted to the privacy issues. The privacy levels for the telecommunication data are proposed and discussed (from no data sharing towards all data sharing approach).
H1	The paper proposes a model describing user's mobility that may be used for anomaly detection in the process of authentication. The activities from Call Detail Records (taking into account all available data describing a single activity) were translated into a mobility dictionary for a user e.g. stay times at locations were quantified. The profile of a user was specified as: locations visited by a user, time that a user spends at a given location (probability that at a specific hour, a user will be present at a specific place), sequence of BTS station visited (identification of routes of a user). Then evaluation of the approach was performed to prove accuracy of the approach (99% of model's accuracy for a random user scenario).
H1	The paper that addresses the goal, presents a method (together with its validation) that enables for development and verification of a mobility profile of a user and may be used for authentication. Here, anomalies in user behaviour are identified and analysed. These anomalies include e.g. checking how movements of a user differ from his typical behaviour (considering time, geographical areas, sequence of places visited or probability of a user being at a specific time, in a specific location). This approach is validated in the paper to prove its usefulness. In addition, the methodology of testing behaviour-based anomaly detection approaches is proposed. It is discussed why a random user comparison method does not make sense while researching authentication scenarios.

H2	The paper in line with the goal studies personalisation of Web interfaces including technical and semantic aspects (elements of the interface as well as its content). The scenario studied concerns a banking system available online. The technical adaptations (personalisation) concern changing graphical interface, whereas semantic personalisation concerns analysing user behaviour and history of transactions to suggest actions to the user. Ten, two approaches are researched (centralised and distributed computing) to personalise and check challenges and limitations. It is proved, that there is a threshold concerning the number of users below which the centralised approach works more effective, however if a large number of users is expected, the distributed approach is better while delivering personalisation.
H2	The chapter concerns studying customer journeys of customers of shopping malls. The experimental data concern city of Poznań and is expressed in the form of Call Detail Records. The loyalty of customers (and their preferences towards choosing a shopping centre) is analysed. The paper presents a method of analysis that offers more knowledge on a customer than traditionally possessed by shopping malls. The results also enable to deliver a new business models for telecommunication companies.

Source: own study

All goals defined in the introduction were successfully realised and therefore the research questions stated were answered.

8.3 Future Work

In the upcoming era of the *personalised healthcare*, the health profiling is getting on importance [66]. A health profile may describe health of a specific population (group of people) and include important factors that influence health of individuals, such as environment, deprivation or educational attainment [66]. Applying terminology used in this thesis, we would define such a profile as a stereotype. In our understanding, a health profile would be defined as a set of features describing a specific person.

The purpose of health stereotypes is generation of high-quality indicators grouped into various health domains with a special geographical focus, and their presentation in a way that allows users to see the extent to which each indicator varies from the average. The criteria for the selection of indicators have been described in [66] and focus on: demonstration of an important impact on the health of population, providing support for the information needs of various healthcare entities, a comparison over time and between places, etc. Some results of health profiling are already provided e.g. by the European project I2SARE[1]. The project produces health stereotypes for each region of the EU to assist European, national, regional and local authorities in developing health policies.

Another example of the personalised healthcare and profiling is related with pharmaceutical companies, and refers to advances in diagnostics and pharmaceuticals aimed at tailoring medicine patients' needs[2]. This research focuses on the study of the effects of genetic differences between patients and their response to medicines.

In medicine four different meanings of personalisation and profiling are distinguished [69]:

- Delivery of highly individualised health management in a sense of prediction, prevention and treatment. Activities should be more tailored to the needs of a specific patient, regarding genetic, physiological or psychological characteristics, provided by personalised technologies applied in a person-specific way.
- Treating each individual separately and being respectful of his/her particular wishes, lifestyle and health status.
- Personalised treatment or management that aims to provide healthcare as a good in ways not dissimilar to other traded products or services that are offered in response to consumer demand.
- Personalisation that can arise from policies of 'responsibilisation', from individuals' choices to manage their healthcare. It means that more responsibility for health management arises from individuals or their carer rather than medical professionals.

Until now, the use of profiling in hospitals and healthcare has been limited. The patient profiling is useful in a variety of situations such as providing a personalised service based on a patient himself [140]. Identifying services that a patient requires allows to speed up patient's recovery progress. Disambiguating patient's diagnostic data based on patient's profile assists in matching doctor's specialisation to the right patient and in providing information about the

[1] `http://www.i2sare.org/`

[2] `http://www.roche.com/personalised_healthcare.htm`

patient on continuous basis for the doctors, so that tailored and appropriate care can be provided. The information about a patient, presented on continuous basis to the medical staff, satisfies the need of up to date information [153]. Therefore, a patient's profile can be described as a collection of data that can be used in a decision analysis situation. It can be divided into a **static profile**, when all information is kept in pre-fixed data fields, where the period between data field updates is very long, and a **dynamic profile**, which is constantly updated as per evaluation of the situation and the updates can be performed manually or in an automated manner [140]. An example of data in a static profile of a patient includes: first name and last name, address, contact data, but also, blood type, allergies, chronic diseases, medicines taken, smoking, etc. On the other hand, dynamic patient profile should store all data that changes over time, e.g. weight, blood pressure, body temperature, GFR, dietary protein restrictions, lipid metabolism, etc. In any case, it is necessary to have both, **static and dynamic patient profile**, to have a full view on the status of patient's health and the course of treatment and reaction of the organism to the procedures carried out.

The recent work of the author concerns collaboration with immunologists and nephrologists from the Pomeranian University in Szczecin and regards care over a patient with kidney-related issues. Regarding personalised healthcare, profiling and data science methods may provide a significant change while diagnosing and treatment of e.g.:

- **ESKD (end stage kidney disease)** – GFR (*gromerular filtration rate*) is an indication of the kidney's condition and describes the flow rate of filtered fluid through the kidney [72]. The problem to be targeted is the unknown origin of chronic kidney disease and the medium-term challenge is to predict the GFR loss [52, 141].
- **Haemodialysis** is a process of purifying the blood of a person whose kidneys are not working normally. When a kidney is in a state of failure, haemodialysis achieves the extracorporeal removal of waste products such as creatinine, urea and free water from blood. As a short-term challenge, two research challenges were identified: anaemia management optimisation [51, 61, 107] and an improvement of arterial-venous fistula management [73, 157]. For the middle-term research issues, mortality risk prediction including novel biomarkers is a desirable topic [7, 115, 160, 164] as well as research on hyoresponsiveness to HBV vaccination [144]. Moreover, also a long-term challenge can be specified and described as life quality improvement including mental well-being [43, 109].

- **Haemodialysis (qualified for transplantation)** – the organ transplant of a kidney into a patient with end stage kidney disease, that can be typically classified as deceased-donor or living-donor transplantation depending on the source of the donor organ. Based on this, a short-term challenge with prevalence of anti-HLA antibodies pre-transplant [127, 130] and a long-term challenge for living donation [29, 45, 82] can be identified.
- **Post-transplant** – the long-term success of a kidney transplant depends on regular examination and parameter monitoring, taking anti-rejection medications in a proper dose and at the right time, following the schedule for lab tests and clinic visits and following a healthy lifestyle including proper diet, exercise and weight loss, if needed. In this case, the short-term research challenge concerns studying pharmacogenomic approach towards immunosuppressants dosage optimisation [8, 37, 106]. For the medium-term challenges two were identified: long term renal allograft survival prediction [52, 168, 172] and donor specific antibodies (DSA) post-transplant prevalence with the precise allograft rejection diagnostics [59]. Also one long-term research challenge was specified as an allograft management for validated and applied biomarkers [32, 113].
- **Repeated transplantation** – each transplanted kidney has a limited lifetime, so after a certain period of time the patient has to undergo transplantation again. Based on literature a long-term challenge can be specified, regarding repeated transplantation in a highly immunised patient [95, 175].

These research challenges demand inter alia diverse profiling approaches to be put in place and are already tackled by the author. First publications showing importance of profiling for the needs of personalised treatment of patients with kidney problems that may influence the treatment are to be published this year.

References

[1] Witold Abramowicz. *Filtrowanie informacji.* Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań, 2008.

[2] accountlearning.com. Introduction to public utilities. `https://accountlearning.com/public-utility-meaning-characteristics-rights-duties/`, 2018.

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[4] Rein Ahas, Anto Aasa, Siiri Silm, and Margus Tiru. Daily rhythms of suburban commuters' movements in the tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies*, 18(1):45–54, 2010.

[5] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[6] Sebastian Ahrndt, Armin Aria, Johannes Fähndrich, and Sahin Albayrak. Ants in the ocean: Modulating agents with personality for planning with humans. In *European Conference on Multi-Agent Systems*, pages 3–18. Springer, 2014.

[7] A.A.D. Allawi. Malnutrition, inflamation and atherosclerosis (mia syndrome) in patients with end stage renal disease on maintenance hemodialysis (a single centre experience). *Diabetes Metab Syndr*, September 2017.

[8] L.M. Andrews, Y. Li, B.C.M. De Winter, Y.Y. Shi, C.C. Baan, T. Van Gelder, and D.A. Hesselink. Pharmacokinetic considerations related to therapeutic drug monitoring of tacrolimus in kidney transplant patients. *Expert Opin Drug Metab Toxicol*, pages 1–12, October 2017.

[9] Theo Arentze, Aloys Borgers, and Harry Timmermans. A model of multi-purpose shopping trip behavior. *Papers in Regional Science*, 72(3):239–256, Jul 1993.

[10] Söeren Auer. Potentials and benefits of linked open data (lod). `https://www.slideshare.net/lod2project/potentials-and-benefits-of-linked-open-data-17846142`, 2013.

[11] Mohd Alif Syami Bin Azmi, Nazrul Bin Mazli, Yusman Yusof, and Mohd Fadzil Hj Abu Hassan. Study of rgb color classification using fuzzy logic. In *ETERD'10 Proceedings*. ACM, 2010.

[12] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 965–974. ACM, 2014.

[13] Ari Banerjee. Big data & advanced analytics in telecom: A multi-billion-dollar revenue opportunity. *Huwawei Heavy Reading*, 2013.

[14] Florian Bauer and Martin Kaltenböck. Linked open data: The essentials. a quick start guide for decision makers. `https://www.reeep.org/LOD-the-Essentials.pdf`, 2012.

[15] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Commun. ACM*, 56(1):74–82, January 2013.

[16] M. G. Beiro, C. Cattuto, L. Ferres, E. Graells-Garrido, L. Bravo, and D. Caro. Understanding mall visiting patterns and mobility in santiago de chile with cdr data. In Francesco Calabrese, Esteban Moro, Vincent Blondel, and Alex Pentland, editors, *NetMob Book of Abstracts: Posters*, pages 10–12. NetMob.

[17] Tim Berners-Lee. Linked data. `https://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[18] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[19] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*, 2015.

[20] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.

[21] Adrian Bridgwater. The 13 types of data. `https://www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/#57b589953362`, 2018.

[22] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 3–53. Springer Berlin / Heidelberg, 2007.

[23] Rachel Cardell-Oliver and Travis Povey. Profiling urban activity hubs using transit smart card data. In *Proceedings of the 5th Conference on Systems for Built Environments*, BuildSys '18, pages 116–125, New York, NY, USA, 2018. ACM.

[24] Reshma Chaudhari and AM Patil. Content based image retrieval using color and shape features. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 1(5):67–72, 2012.

[25] Q. Chen, A. F. Norcio, and J. Wang. Neural network based stereotyping for user profiles. *Neural Computing & Applications*, 9(4):259–265, Dec 2000.

[26] Nai Chun Chen, Jenny Xie, Phil Tinn, Luis Alonso, Takehiko Nagakura, and Kent Larson. Data mining tourism patterns. call detail records as complementary tools for urban decision making. In P. Janssen, P. Loh, A. Raonic, and M. A. Schnabel, editors, *Protocols, Flows and Glitches, Proceedings of the 22nd International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2017*. The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA).

[27] Alan Cooper, Robert Reimann, and Hugh Dubberly. *About Face 2.0: The Essentials of Interaction Design*. John Wiley & Sons, Inc., New York, NY, USA, 1 edition, 2003.

[28] Council of European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (EU) no 5419/2016, 2016. `http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`.

[29] R.Z. Croft and C. Maddison. Experience of directed living donor kidney transplant recipients: a literature review. *Nurs Stand*, 32(3):41–49, September 2017.

[30] Ayse Cufoglu. Article: User profiling - a short review. *International Journal of Computer Applications*, 108(3):1–9, December 2014. Full text available.

[31] Michael C. Daconta, Leo J. Obrst, and Kevin T. Smith. *The Semantic Web : A Guide to the Future of XML, Web Services, and Knowledge Management*. Wiley, 2003.

[32] Richard Danger, Birgit Sawitzki, and Sophie Brouard. Immune monitoring in renal transplantation:the search for biomarkers. *Eur. J. Immunol*, 46:2695–2704, 2016.

[33] DATA.GOV. Welcome to the green button! `https://www.data.gov/energy/welcome-green-button`, 2018.

[34] Deloitte. Utility 2.0. winning over the next generation of utility customers. `https://www2.deloitte.com/content/dam/Deloitte/us/Documents/energy-resources/us-e-r-utility-report.pdf`, 2017.

[35] David Dubin. The most influential paper gerard salton never wrote. *Library Trends*, 2004.

[36] Editors of Encyclopaedia Britannica. Public utility. `https://www.britannica.com/technology/public-utility`, 2018.

[37] L. Elens and V. Haufroid. Genotype-based tacrolimus dosing guidelines: with or without cyp3a4*22? *Pharmacogenomics*, November 2017.

[38] European Commission. Building a european data economy. `https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy`, 2018.

[39] S. Fan, R.Y.K. Lau, and J.L. Zhao. Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research*, 2(1):28–32, 2015. cited By 44.

[40] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, pages 7–10. ACM, 2015.

[41] Tom Flaherty, Norbert Schwieters, and Steve Jennings. 2017 power and utilities trends. `https://www.strategyand.pwc.com/trend/2017-power-and-utilities-industry-trends`, 2016.

[42] Lex Fridman, Steven Weber, Rachel Greenstadt, and Moshe Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and GPS location. *CoRR*, abs/1503.08479, 2015.

[43] S. Fukuma, S. Shimizu, A. Shintani, T. Kamitani, T. Akizawa, and S. Fukuhara. Development and validation of a prediction model for loss of physical function in elderly hemodialysis patients. *Nephrol Dial Transplant*, September 2017.

[44] Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, and Salvatore Rinzivillo. Analysis of GSM calls data for understanding user mobility behavior. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 550–555, 2013.

[45] J.C. Gander, E.J. Gordon, and R.E. Patzer. Decision aids to increase living donor kidney transplantation. *Curr Transplant Rep*, 4(1):1–12, March 2017.

[46] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. pages 54–89. 2007.

[47] Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. Personalizing search results using hierarchical rnn with query-aware attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 347–356, New York, NY, USA, 2018. ACM.

[48] Andrea N Geurin-Eagleman and Lauren M Burch. Communicating via photographs: A gendered analysis of olympic athletes' visual self-presentation on instagram. *Sport Management Review*, 19(2):133–145, 2016.

[49] Divya Ghorawat and Ravina Madan. Correlation between personality types and color shade preference. *The International Journal of Indian Psychology*, 1(04), 2014.

[50] Stamatios Giannoulakis and Nicolas Tsapatsoulis. Evaluating the descriptive power of instagram hashtags. *Journal of Innovation in Digital Ecosystems*, 3(2):114–129, 2016.

[51] D.T. Gilbertson, Y. Hu, Y. Peng, B.J. Maroni, and J.B. Wetmore. Variability in hemoglobin levels in hemodialysis patients in the current era: a retrospective cohort study. *Clin Nephrol*, 88(11):254–265, November 2017.

[52] O.N. Goek, C. Prehn, P. Sekula, W. Römisch-Margl, A. Döring, C. Gieger, M. Heier, W. Koenig, R. Wang-Sattler, T. Illig, K. Suhre, J. Adamski, A. Köttgen, and C. Meisinger. Metabolites associate with kidney function decline and incident chronic kidney disease in the general population. *Nephrol Dial Transplant*, 28(8):2131–2138, August 2013.

[53] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[54] Gavin Graham. How to create customer profiles with template and examples. `https://fitsmallbusiness.com/customer-profile-template-examples/`, 2017.

[55] Green Button Alliance. Button colors. blue button, green button, and now orange button... what do they all mean? `https://www.greenbuttonalliance.org/button-colors`, 2018.

[56] Green Button Alliance. The history of green button. `https://www.greenbuttonalliance.org/about#history`, 2018.

[57] Shirley Gregor and Alan R Hevner. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2):337–356, 2013.

[58] Synh Viet-Uyen Ha, Nhan Thanh Pham, Long Hoang Pham, and Ha Manh Tran. Robust reflection detection and removal in rainy conditions using lab and hsv color spaces. *REV Journal on Electronics and Communications*, 6(1-2), 2016.

[59] M. Haas. Chronic allograft nephropathy or interstitial fibrosis and tubular atrophy: what is in a name? *Curr Opin Nephrol Hypertens*, 23(3):245–250, May 2014.

[60] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 213–230, Menlo Park, CA, July 2014. USENIX Association.

[61] T. Hasegawa, J. Zhao, D.S. Fuller, B. Bieber, J. Zee, H. Morgenstern, N. Hanafusa, and M. Nangaku. Erythropoietin hyporesponsiveness in dialysis patients: Possible role of statins. *Am J Nephrol*, 46(1):11–17, 2017.

[62] T. Heath, M. Hepp, and C. Bizer. Special issue on linked data, international journal on semantic web and information systems (ijswis). `http://linkeddata.org/docs/ijswis-special-issue`, ..

[63] Alan Hevner and Samir Chatterjee. Design science research in information systems. In *Design Research in Information Systems*, volume 22 of *Integrated Series in Information Systems*, pages 9–22. Springer US, 2010.

[64] Alan Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

[65] Alan R Hevner. A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2):4, 2007.

[66] A. Hill, K. Balanda, L. Galbraith, J. Greenacre, and D. Sinclair. Profiling health in the uk and ireland. *Public Health*, 124:253–258, 2010.

[67] Sarah Hippold. Enter the age of analytics. `https://www.gartner.com/smarterwithgartner/enter-the-age-of-analytics/`, 2018.

[68] Nadav Hochman and Raz Schwartz. Visualizing instagram: Tracing cultural visual rhythms. In *Proceedings of the workshop on Social Media Visualization (SocMedVis) in conjunction with the sixth international AAAI conference on Weblogs and Social Media (ICWSM–12)*, pages 6–9, 2012.

[69] Christopher Hood, editor. *Medical profiling and online medicine: the ethics of personalised healthcare in a consumer age.* Nuffield Council on Bioethics, 2010.

[70] Teerayut Horanont. A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources. *The University of Tokyo Center for Spatial Information Service, CSIS Discussion Paper*, (115), 2010.

[71] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*, 2015.

[72] J.M. Hougardy, P. Delanaye, A. Le Moine, and J. Nortier. Estimation of the glomerular filtration rate in 2014 by tests and equations: strengths and weaknesses. *Rev Med Brux*, 35(4):250–257, September 2014.

[73] H. Hu, S. Patel, J.J. Hanisch, J.M. Santana, T. Hashimoto, H. Bai, T. Kudze, T.R. Foster, J. Guo, B. Yatsula, J. Tsui, and A. Dardik. Future research directions to improve fistula maturation and reduce access failure. *Semin Vasc Surg*, 29(4):153–171, December 2016.

[74] Haiyan Hu and Cynthia Jasper. Consumer shopping experience in the mall: Conceptualization and measurement. In Dheeraj Sharma and Shaheen Borna, editors, *Proceedings of the 2007 Academy of Marketing Science (AMS) Annual Conference*, pages 8–8, Cham, 2015. Springer International Publishing.

[75] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. What we instagram: A first analysis of instagram photo content and user types. In *Icwsm*, 2014.

[76] Yuxia Huang and Ling Bian. A bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet. *Expert Syst. Appl.*, 36(1):933–943, 2009.

[77] David L. Huff. A probabilistic analysis of shopping center trade areas. *Land Economics*, 39(1):81–90, 1963.

[78] IBM. 10 key marketing trends for 2017 and ideas for exceeding customer expectations. `https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl12345usen-20170719.pdf`, 2017.

[79] IBM Institute. Analytics: Real-world use of big data in telecommunications. *IBM Institute for Business Value Executive Report*, 2013.

[80] McKinley Stacker IV. Using customer behavior data to improve customer retention. `https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/`, 2015.

[81] Christina A Jackson and Andrew F Luchner. Self-presentation mediates the relationship between self-criticism and emotional response to instagram feedback. *Personality and Individual Differences*, 2017.

[82] K.J. Jager, V.S. Stel, P. Branger, M. Guijt, M. Busic, and M. Dragovic. The effect of differing kidney disease treatment modalities and organ donation and transplantation practices on health expenditure and patient outcomes. *Nephrol Dial Transplant*, July 2017.

[83] Laura James. Defining open data. `https://blog.okfn.org/2013/10/03/defining-open-data/`, 2013.

[84] Nursuriati Jamil et al. Automatic image annotation using color k-means clustering. In *International Visual Informatics Conference*, pages 645–652. Springer, 2009.

[85] Jin Yea Jang, Kyungsik Han, Patrick C Shih, and Dongwon Lee. Generation like: comparative characteristics in instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4039–4042. ACM, 2015.

[86] Shan Jiang, Joseph Ferreira, and Marta C González. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore. *ACM KDD UrbComp'15*, 2015.

[87] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[88] Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. User modeling with personas. In *Proceedings of the 2005 Latin American Conference on Human-computer Interaction*, CLIHC '05, pages 277–282, New York, NY, USA, 2005. ACM.

[89] Sumitkumar Kanoje, Sheetal Girase, and Debajyoti Mukhopadhyay. User profiling trends, techniques and applications. *CoRR*, abs/1503.07474, 2015.

[90] Judy Kay. Lifelong learner modeling for lifelong personalized pervasive learning. *IEEE Trans. Learn. Technol.*, 1(4):215–228, 2008.

[91] S. Kirkpatrick, R. Bekkerman, A. Zmirli, and F. Malandrino. Mining the Air – for Research in Social Science and Networking Measurement. *ArXiv e-prints*, June 2018.

[92] Ryuichi Kitamura. Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B: Methodological*, 18(1):67 – 81, 1984.

[93] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets semantic web – how the bbc uses dbpedia and linked data to make connections. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl, editors, *The Semantic Web: Research and Applications*, pages 723–737, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[94] Nora Koch. *Software Engineering for Adaptive Hypermedia Systems*. PhD thesis, Ludwig-Maximilians-Universität München, 2000.

[95] P. Koefoed-Nielsen, I. Weinreich, M. Bengtsson, J. Lauronen, C. Naper, M. Gäbel, S.S. Sørensen, L. Wennberg, A.V. Reisaeter, B.K. Møller, Nordic Kidney group, and the Tissue Typing group in Scandiatransplant. Scandiatransplant acceptable mismatch program (stamp) a bridge to transplanting highly immunized patients. *HLA*, 90(1):17–24, July 2017.

[96] Sanford Labovitz. Criteria for selecting a significance level: A note on the sacredness of. 05. *The American Sociologist*, pages 220–222, 1968.

[97] Katherine N. Lemon and Peter C. Verhoef. Understanding customer experience throughout the customer journey. *Journal of Marketing*, 80(6):69–96, 2016.

[98] K. W. Leung and D. L. Lee. Deriving concept-based user profiles from search engine logs. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):969–982, July 2010.

[99] Lightpetal. Design and Creative Direction. Vouchercodes persona decals and cards. `https://www.lightpetal.com/vouchercodes-persona-decals-and-cards/`, 2018.

[100] Feng Liu, Davy Janssens, JianXun Cui, YunPeng Wang, Geert Wets, and Mario Cools. Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174–6189, 2014.

[101] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen Ebrahimi Moghaddam, and Lyle H Ungar. Analyzing personality through social media profile picture choice. In *ICWSM*, pages 211–220, 2016.

[102] Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu. *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, volume 9471. Springer, 2015.

[103] Steve Luong. The top 6 challenges utilities face—and how they can beat them. `https://www.kony.com/resources/blog/top-6-challenges-utilities-face-and-how-they-can-beat-them/`, 2017.

[104] Katerina Lup, Leora Trub, and Lisa Rosenthal. Instagram# instasad?: exploring associations among instagram use, depressive symptoms, negative social comparison, and strangers followed. *Cyberpsychology, Behavior, and Social Networking*, 18(5):247–252, 2015.

[105] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM, 2010.

[106] M. Maier, T. Takano, and R. Sapir-Pichhadze. Changing paradigms in the management of rejection in kidney transplantation: Evolving from protocol-based care to the era of p4 medicine. *Can J Kidney Health Dis*, January 2017.

[107] S.W. Majoni, J.A. Ellis, H. Hall, A. Abeyaratne, and P.D. Lawton. Inflammation, high ferritin, and erythropoietin resistance in indigenous maintenance hemodialysis patients from the top end of northern australia. *Hemodial Int*, 18(4):740–750, October 2014.

[108] Market Business News. What is a public utility? definition and meaning. `https://marketbusinessnews.com/financial-glossary/public-utility-definition-meaning/`, 2018.

[109] R. Matsuzawa, B. Roshanravan, T. Shimoda, N. Mamorita, K. Yoneki, M. Harada, T. Watanabe, A. Yoshida, Y. Takeuchi, and A. Matsunaga. Physical activity dose for hemodialysis patients: Where to begin? results from a prospective cohort study. *J Ren Nutr*, September 2017.

[110] C. McCauley, C.L. Stitt, and M. Segal. Stereotyping: from prejudice to prediction. *Psycho Bulletin*, 1980.

[111] Janet R. McColl-Kennedy, Anders Gustafsson, Elina Jaakkola, Phil Klaus, Zoe Jane Radnor, Helen Perks, and Margareta Friman. Fresh perspectives on customer experience. *Journal of Services Marketing*, 29(6/7):430–435, 2015.

[112] John P. McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, and Vladimir Andryushechkin. The linked open data cloud. `https://lod-cloud.net/`, 2018.

[113] Madhav C. Menon, Barbara Murphy, and Peter S. Heeger. Moving biomarkers toward clinical implementation in kidney transplantation. *J Am Soc Nephrol*, 28:735–747, 2017.

[114] Rui Min and HD Cheng. Effective image retrieval using dominant color descriptor and fuzzy support vector machine. *Pattern Recognition*, 42(1):147–157, 2009.

[115] K. Nagai, M. Matsuura, K. Tsuchida, H.O. Kanayama, T. Doi, and J. Minakuchi. Prognostic factors for mortality in middle-aged and older hemodialysis patients: a 5-year observational study. *J Artif Organs*, September 2017.

[116] Chihiro Ono, Mori Kurokawa, Yoichi Motomura, and Hideki Asoh. A context-aware movie preference model using a bayesian network for recommendation and promotion. In *UM '07: Proceedings of the 11th international conference on User Modeling*, pages 247–257, Berlin, Heidelberg, 2007. Springer-Verlag.

[117] Open Gov Data. The 8 principles of open government data. `https://opengovdata.org/`, 2007.

[118] Open Knowledge International. What is open data? `http://opendatahandbook.org/guide/en/what-is-open-data/`, 2018.

[119] Open Knowledge International. Why open data? `http://opendatahandbook.org/guide/en/why-open-data/`, 2018.

[120] Hubert Österle, Jörg Becker, Ulrich Frank, Thomas Hess, Dimitris Karagiannis, Helmut Krcmar, Peter Loos, Peter Mertens, Andreas Oberweis, and Elmar J Sinz. Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20(1):7–10, 2011.

[121] Amy L. Ostrom, A. Parasuraman, David E. Bowen, Lia Patrício, and Christopher A. Voss. Service research priorities in a rapidly changing context. *Journal of Service Research*, 18(2):127–159, 5 2015.

[122] Boris Otto and Hubert Österle. *Corporate Data Quality: Prerequisite for Successful Business Models*. epubli, Berlin, 2015.

[123] Boris Otto and Hubert Österle. Relevance through consortium research? findings from an expert interview study. In Robert Winter, J. Leon Zhao, and Stephan Aier, editors, *DESRIST*, volume 6105 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2010.

[124] Ajay Pandit, Barton Weitz, and Michael Levy. *Retailing Management (Eighth Edition)*. Tata McGraw-Hill Education Pvt. Ltd,, 2012.

[125] Adam D Pazda. *Colorful Personalities: Investigating the Relationship Between Chroma, Person Perception, and Personality Traits.* PhD thesis, University of Rochester, 2015.

[126] Michael J. Pazzani and Daniel Billsus. *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[127] J.P. Perasaari, L.E. Kyllönen, K.T. Salmela, and J.M. Merenmies. Pre-transplant donor-specific anti-human leukocyte antigen antibodies are associated with high risk of delayed graft function after renal transplantation. *Nephrol Dial Transplant*, 31(4):672–678, April 2016.

[128] Jenna Perkins, Suzie Dutson, Rachel Quinn, Yevi Greene, Teuila Nautu, and Emilie J Davis. Is every picture worth 1,000 likes?: A content analysis of the images and messages on popular instagram accounts. 2017.

[129] Guangyuan Piao and John G. Breslin. Inferring user interests in microblogging social networks: A survey. *CoRR*, abs/1712.07691, 2017.

[130] I. Pirim, M. Soyoz, T.K. Ayna, A.O. Kocyigit, B.C. Gurbuz, C. Tugmen, Y. Kurtulmus, and B. Ozyilmaz. De novo produced anti-human leukocyte antigen antibodies relation to alloimmunity in patients with chronic renal failure. *Genet Test Mol Biomarkers*, 19(6):335–338, June 2015.

[131] Matthew Pittman. Creating, consuming, and connecting: examining the relationship between social media engagement and loneliness. *The Journal of Social Media in Society*, 4(1), 2015.

[132] Robert Plutchik. *Emotion: A psychoevolutionary synthesis.* Harpercollins College Division, 1980.

[133] Tejaswi Potluri and Gnaneswararao Nitta. Content based video retrieval using dominant color of the truncated blocks of frame. *Journal of Theoretical and Applied Information Technology*, 85(2):165, 2016.

[134] PWC. Flipping the switch on disruption to opportunity: Top 6 focus areas in 2016. `https://www.pwc.com/us/en/industries/power-utilities/library/flipping-the-switch-on-disruption-2016.html?_ga=2.85518673.952090783.1543092826-2015122441.1490945510`, 2016.

[135] Ian Rives. Five challenges facing the utilities industry in 2018. `https://www.easi.com/en/insights/articles/5-challenges-facing-the-utilities-industry-in-2018`, 2018.

[136] Mark S. Rosenbaum, Mauricio Losada Otalora, and Germán Contreras Ramírez. How to create a realistic customer journey map. *Business Horizons*, 60(1):143–150, 2017.

[137] Silvia Rossi, François Ferland, and Adriana Tapus. User profiling and behavioral adaptation for hri: A survey. *Pattern Recognition Letters*, 99:3 – 12, 2017. User Profiling and Behavior Adaptation for Human-Robot Interaction.

[138] Mor Rubinstein and Katelyn Rogers. Saludos - health and open data in uruguay and argentina. `http://opendatahandbook.org/value-stories/en/latam-health/`, 2018.

[139] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[140] Surendra Sarnikar, Dorine Bennett, and Mark Gaynor, editors. *Cases on Healthcare Information Technology for Patient Care Management*. IGI Global, 1st edition, December 2012.

[141] P.J. Saulnier, E. Gand, G. Velho, K. Mohammedi, P. Zaoui, M. Fraty, J.M. Halimi, R. Roussel, S. Ragot, S. Hadjadj, and SURDIAGENE Study Group. Association of circulating biomarkers (adrenomedullin, tnfr1, and nt-probnp) with renal function decline in patients with type 2 diabetes: A french prospective cohort. *Diabetes Care*, 40(3):367–374, March 2017.

[142] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[143] Ryan Singel. Meet the company that records your calls for quality assurance, 2011.

[144] D. Sit, B. Esen, A.E. Atay, and H. Kayabasi. Is hemodialysis a reason for unresponsiveness to hepatitis b vaccine? hepatitis b virus and dialysis therapy. *World J Hepatol*, 7(5):761–768, April 2015.

[145] K. Slaninová. User behavioural patterns and reduced user profiles extracted from log files. In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 289–294, Dec 2013.

[146] Lauren Reichart Smith and Jimmy Sanderson. I'm going to instagram it! an analysis of athlete self-presentation on instagram. *Journal of Broadcasting & Electronic Media*, 59(2):342–358, 2015.

[147] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[148] Sergey Sosnovsky and Darina Dicheva. Ontological technologies for user modelling. *Int. J. Metadata Semant. Ontologies*, 5:32–71, April 2010.

[149] Flávio Souza, Diego de Las Casas, Vinícius Flores, SunBum Youn, Meeyoung Cha, Daniele Quercia, and Virgílio Almeida. Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram. In *Proceedings of the 2015 ACM on conference on online social networks*, pages 221–231. ACM, 2015.

[150] STATISTA. Media usage in an internet minute as of june 2018. `https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/`, 2018.

[151] A. Stevenson and C.A. Lindberg. *New Oxford American Dictionary, Third Edition*. Oxford reference online premium. OUP USA, 2010.

[152] Sunlight Foundation. Ten principles for opening up government information. `https://sunlightfoundation.com/policy/documents/ten-open-data-principles/`, 2018.

[153] David Taniar. *Mobile Computing: Concepts, Methodologies, Tools, and Applications*. IGI Global, 1st edition, November 2008.

[154] Baiping Tao, Shaofang Xu, Xin Pan, Qianqian Gao, and Wei Wang. Personality trait correlates of color preference in schizophrenia. *Translational Neuroscience*, 6(1), 2015.

[155] Roberto Tedesco, Peter Dolog, Wolfgang Nejdl, and Heidrun Allert. Distributed bayesian networks for user modeling. Hawai, USA, oct 2006.

[156] The Open Data Institute. Making aid more effective in nepal. `http://opendatahandbook.org/value-stories/en/effective-aid-in-nepal/`, 2018.

[157] M. Thomas, C. Nesbitt, M. Ghouri, and M. Hansrani. Maintenance of hemodialysis vascular access and prevention of access dysfunction: A review. *Ann Vasc Surg*, 43:318–327, August 2017.

[158] Harry Timmermans. Applied geography: A world perspective, geojournal library. Springer Netherlands, 2004.

[159] J. M. Torres and A. P. Parkes. User modelling and adaptivity in visual information retrieval systems. `http://www.comp.lancs.ac.uk/computing/research/mcg/papers/torresParkessurrey2000.pdf`, 2000.

[160] M.T. Tsai, H.C. Liu, and T.P. Huang. The impact of malnutritional status on survival in elderly hemodialysis patients. *J Chin Med Assoc*, 79(6):30, June 2016.

[161] D. Vallet, I. Cantador, and J. Jose. Personalizing web search with folksonomy-based user and document profiles. *Advances in Information Retrieval*, pages 420–431, 2010.

[162] Dirk Van Bruggen, Shu Liu, Mitch Kajzer, Aaron Striegel, Charles R. Crowell, and John D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 10:1–10:14, New York, NY, USA, 2013. ACM.

[163] Jan-Willem van Dam and Michel van de Velden. Online profiling and clustering of facebook users. *Decision Support Systems*, 70:60 – 72, 2015.

[164] J. Varas, R. Ramos, P. Aljama, R. Pérez-García, F. Moreso, M. Pinedo, J. Ignacio Merello, S. Stuard, B. Canaud, A. Martín-Malo, and ORD Group. Relationships between iron dose, hospitalizations and mortality in incident haemodialysis patients: a propensity-score matched approach. *Nephrol Dial Transplant*, July 2017.

[165] John Venable, Jan Pries-Heje, and Richard Baskerville. Feds: a framework for evaluation in design science research. *European Journal of Information Systems*, 25(1):77–89, Jan 2016.

[166] Clay M. Voorhees, Paul W. Fombelle, Yany Gregoire, Sterling Bone, Anders Gustafsson, Rui Sousa, and Travis Walkowiak. Service encounters, experiences and the customer journey: Defining the field and a call to expand our lens. *Journal of Business Research*, 79:269 – 280, 2017.

[167] W. Wahlster and A. Kobsa. *User Models in Dialog Systems.* Springer, 1989.

[168] L. Walsh and R. Dinavahi. Current unmet needs in renal transplantation: a review of challenges and therapeutics. *Front Biosci*, 8:1–14, January 2016.

[169] Peng Wang, Dongqing Zhang, Gang Zeng, and Jingdong Wang. Contextual dominant color name extraction for web image search. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 319–324. IEEE, 2012.

[170] Sophie F Waterloo, Susanne E Baumgartner, Jochen Peter, and Patti M Valkenburg. Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. *New Media & Society*, page 1461444817707349, 2017.

[171] Yihong Yuan, Martin Raubal, and Yu Liu. Correlating mobile phone usage and travel behavior – A case study of Harbin, China. *Computers, Environment and Urban Systems*, 36(2):118–130, 2012.

[172] G. Zaza, S. Granata, P. Tomei, A. Dalla Gassa, and A. Lupo. Personalization of the immunosuppressive treatment in renal transplant recipients: the great challenge in "omics" medicine. *Int J Mol Sci*, 16(2):4281–4305, February 2015.

[173] Changtao Zhong, Hau-wen Chan, Dmytro Karamshu, Dongwon Lee, and Nishanth Sastry. Wearing many (social) hats: How different are your different social network personae? *arXiv preprint arXiv:1703.04791*, 2017.

[174] Guijun Zhuang, Alex S.L. Tsang, Nan Zhou, Fuan Li, and J.A.F. Nicholls. Impacts of situational factors on buying decisions in shopping malls: An empirical study with multinational data. *European Journal of Marketing*, 40(1/2):17–43, 2006.

[175] H. Zielińska, G. Moszkowska, M. Zieliński, A. Debska-Ślizień, B. Rutkowski, and P. Trzonkowski. Algorithm to manage highly sensitized kidney transplant recipients in poland. *Transplant Proc*, 43(8):2903, October 2011.

[176] Peter Pal Zubcsek, Zsolt Katona, and Miklos Sarvary. Predicting mobile advertising response using consumer colocation networks. *Journal of Marketing*, 81(4):109–126, 2017.