

# **Geographical Places as a Personalisation Element**

**Extracting Profiles from Human Activities and Services of Visited  
Places in Mobility Logs**

**A thesis submitted in partial fulfilment  
of the requirement for the degree of Doctor of Philosophy**

**Ahmed N. Makki Al-Azzawi**

**October 2011**

**Cardiff University  
School of Computer Science & Informatics**

UMI Number: U585512

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U585512

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.  
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against  
unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

## Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... ~~.....~~ ..... (candidate)

Date ..... 27/1/2012 .....

## Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... ~~.....~~ ..... (candidate)

Date ..... 27/1/2012 .....

## Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... ~~.....~~ ..... (candidate)

Date ..... 27/1/2012 .....

## Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... ~~.....~~ ..... (candidate)

Date ..... 27/1/2012 .....

*To L. S.*

# Abstract

Collecting personal mobility traces of individuals is currently applicable on a large scale due to the popularity of position-aware mobile phones. Statistical analysis of GPS data streams, collected with a mobile phone, can reveal several interesting measures such as the most frequently visited geographical places by some individual. Applying probabilistic models to such data sets can predict the next place to visit, and when. Several practical applications can utilise the results of such analysis. Current state of the art, however, is limited in terms of the qualitative analysis of personal mobility logs. Without explicit user-interactions, not much semantics can be inferred from a GPS log.

This work proposes the utilisation of the common human activities and services provided at certain place types to extract semantically rich profiles from personal mobility logs. The resulting profiles include spatial, temporal and generic thematic description of a user.

The work introduces several pre-processing methods for GPS data streams, collected with personal mobile devices, which improved the quality of the place extraction process from GPS logs. The thesis also introduces a method for extracting place semantics from multiple data sources. A textual corpus of functional descriptions of human activities and services associated with certain geographic place types is analysed to identify the frequent linguistic patterns used to describe such terms. The patterns found are then matched against multiple textual data sources of place semantics, to extract such

terms, for a collection of place types. The results were evaluated in comparison to an equivalent expert ontology, as well as to semantics collected from the general public. Finally, the work proposes a model for the resulting profiles, the necessary algorithms to build and utilise such profiles, along with an encoding mark-up language. A simulated mobile application was developed to show the usability and for evaluation of the resulting profiles.

# Acknowledgements

I can not emphasise enough that I would have never been able to do this work alone, neither I can thank enough the people who made it happen.

Nabeel and Manal Alazzawi, the tireless guards of me. Sulaiman and Noora Alazzawi, my beloved brother and sister. Thank you very much indeed.

I would also like to thank the two people who not only provided all their time, technical support and knowledge to this work, but also were kind to me all the way – something which I quite appreciate. My supervisors; Alia and Chris. Thank you very much.

Many of the people here at Cardiff School of Computer Science & Informatics, friends, colleagues and staff, made life easier for me. Phil, Mark, Florian, Vitally, Fahad, Ehab, Diego, Hmood, Sultan, Ralph, Alic, Mat, Konrad, Rob, Rob, Omer and Mona. Thank you all.

Jenny, from the Cardiff University Students Support Centre, and Helen, the School's administrator, were exceptionally kind and supportive. Thank you very much.

Finally, I would like to acknowledge the help of Vanessa.

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Contents</b>	<b>vi</b>
<b>List of Publications</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xvi</b>
<b>List of Algorithms</b>	<b>xviii</b>
<b>List of Acronyms</b>	<b>xx</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.1.1 Research problem . . . . .	2
1.1.2 Scope and applications . . . . .	3



---

1.1.3	Current limitations . . . . .	5
1.1.4	Research hypothesis . . . . .	6
1.2	Research challenges . . . . .	7
1.3	State of the art . . . . .	8
1.4	Research objectives . . . . .	14
1.5	Research contributions . . . . .	15
1.6	Structure of the thesis . . . . .	17
<b>2</b>	<b>Background and related work</b>	<b>18</b>
2.1	Personalised systems . . . . .	19
2.1.1	Context-aware computing . . . . .	22
2.1.2	Personal profile . . . . .	24
2.1.3	Personal Gazetteers (Personal Gazetteer (PG)) . . . . .	27
2.2	Analysing human mobility . . . . .	27
2.3	Statistical analysis . . . . .	30
2.4	Semantic analysis . . . . .	33
2.5	Other techniques . . . . .	36
2.6	Limitations . . . . .	38
2.7	Extracting place affordance . . . . .	40
2.8	Summary . . . . .	43

---

<b>3</b>	<b>Improving the quality of PG extraction by pre-processing</b>	<b>45</b>
3.1	Introduction . . . . .	45
3.2	Personal Gazetteers from GPS data streams . . . . .	45
3.3	Problems of personal GPS data streams for building PG . . . . .	49
3.3.1	The loss of GPS signal . . . . .	49
3.3.2	Erroneous GPS readings . . . . .	50
3.3.3	Open-spaces detection . . . . .	50
3.3.4	Inaccurate GPS readings . . . . .	51
3.3.5	Other issues with GPS readings . . . . .	51
3.4	Improvements procedures . . . . .	52
3.4.1	Overshoot points filter . . . . .	53
3.4.2	Segmented tracks merger . . . . .	54
3.4.3	Noise filter . . . . .	56
3.5	Results . . . . .	56
3.5.1	Experiment 1 . . . . .	59
3.5.2	Experiment 2 . . . . .	61
3.6	Evaluation . . . . .	63
3.6.1	Analysis of experiment 1 . . . . .	63
3.6.2	Analysis of experiment 2 . . . . .	66
3.6.3	Discussion . . . . .	67
3.7	Application . . . . .	68
3.8	Conclusion . . . . .	71

---

<b>4</b>	<b>Extracting place-related services and activities</b>	<b>73</b>
4.1	Place Ontology . . . . .	73
4.2	Ontology building approach . . . . .	75
4.2.1	Candidate Pattern Mining (CPM) . . . . .	77
4.2.2	Candidate Pattern Expansion (CPE) . . . . .	79
4.3	Extracted patterns . . . . .	81
4.4	Results . . . . .	84
4.5	Evaluation . . . . .	87
4.5.1	User-based evaluation . . . . .	87
4.5.2	Comparison with an expert ontology . . . . .	90
4.6	Summary . . . . .	94
<b>5</b>	<b>Graph model of human mobility</b>	<b>95</b>
5.1	Introduction . . . . .	95
5.2	Modelling human mobility . . . . .	95
5.3	Personal Mobility Graph . . . . .	99
5.4	Personal Mobility Graph (PMG) Constructor Algorithm . . . . .	102
5.5	Personal Place Similarity Algorithm . . . . .	105
5.6	Conclusion . . . . .	110
<b>6</b>	<b>Personal gazetteers mark-up language</b>	<b>111</b>
6.1	Introduction . . . . .	111
6.2	Personal Gazetteers Mark-up Language (PGML) . . . . .	112

- 6.3 Motivating scenario . . . . . 113
- 6.4 Discussion . . . . . 114
- 6.5 Conclusion . . . . . 118
  
- 7 Profiles evaluation 119**

  - 7.1 Results . . . . . 119
    - 7.1.1 Pilot profile . . . . . 119
    - 7.1.2 Group profiles . . . . . 121
  - 7.2 Evaluation . . . . . 125
    - 7.2.1 Approach . . . . . 125
    - 7.2.2 Model design . . . . . 126
    - 7.2.3 Pilot profile . . . . . 127
    - 7.2.4 Group profiles . . . . . 130
  - 7.3 Conclusion . . . . . 134

  
- 8 Conclusions 135**

  - 8.1 Contributions . . . . . 136
  - 8.2 Discussion . . . . . 138
  - 8.3 Future work . . . . . 140

  
- Bibliography 144**

# List of Publications

Some of the work introduced in this thesis was published in the following publications.

## Journal article

- Ahmed N. Alazzawi, Alia I. Abdelmoty, and Christopher B. Jones. What can I do there? Towards the Automatic Discovery of Place-related Services and Activities. To appear in *International Journal of Geographical Information Science*, Taylor & Francis.

## Conference and workshop papers

- Ahmed N. Alazzawi, Alia I. Abdelmoty, and Christopher B. Jones. Toward a Common Model and a Markup Language for Personal Gazetteers. *ACM Hot-Planet 2010 The 2nd ACM International Workshop on Hot Topics in Planet-Scale Measurement*, held in cooperation with ACM MobiSys, San Francisco, CA, USA, June 15, 2010.
- Ahmed N. Alazzawi, Alia I. Abdelmoty, and Christopher B. Jones. An Ontology of Place and Service Types to Facilitate Place-Affordance Geographic Information Retrieval. *The 6th ACM Workshop on Geographic Information Retrieval*, held in cooperation with ACM SIGSPATIAL, Zurich, Switzerland 18-19 February, 2010.

- Hall, M. M., A. N. Alazzawi, A. I. Abdelmoty, and C. B. Jones. Improving the Quality of GPS-based Personal Gazetteers, *GI\_Forum 2009*: Herbert Wichman Verlag, Heidelberg, pp. 53-55, 2009.

**Poster and short paper**

- Alazzawi, A. N., A. I. Abdelmoty, and C. B. Jones. Spatio-temporal Personalised Web Search, Maynooth, GIS Research UK 15th Annual Conference (GISRUK-2007), Maynooth Ireland, pp. 457-459, 2007.

---

## List of Figures

1.1	Yahoo fire eagle platform . . . . .	4
1.2	Sample personal mobility based applications . . . . .	5
1.3	Research framework . . . . .	15
2.1	Architecture of a sample personalised Web search engine – simplified	21
2.2	A sample profile – (Newbould and Collingridge, 2003) . . . . .	24
2.3	A sample profile – (Searby, 2003) . . . . .	25
2.4	State of the art personal mobility analysis for the purpose of construct- ing personal gazetteers . . . . .	37
2.5	An example of a geographical place ontology employed in GIR on the Semantic Web (Abdelmoty et al., 2009) . . . . .	41
2.6	Action hierarchy derived from traffic code texts in (Kuhn, 2001) (re- produced) . . . . .	43
3.1	A segmented track due to the loss of a Global Positioning System (GPS) signal . . . . .	49
3.2	A track with overshoot points . . . . .	51
3.3	A track with inaccurate GPS readings . . . . .	52

---

3.4	Places which were visited in the second experiment – all of which are within a relatively small urban area; which lie within a circle of 1.2 kilometers diameter . . . . .	62
3.5	Detection of errors in urban canyons. The GPS signal was lost before reaching the destination places: Place 19, Place 23, Place 26. The places which were not correctly identified due to errors in the GPS readings are: Place 7, Place 45, Place 54 . . . . .	64
3.6	Place Activity Time data collection application architecture . . . . .	69
3.7	Place Activity Time application database schema . . . . .	70
3.8	Place Activity Time – Web interface snapshots . . . . .	71
4.1	Place ontology (RWPO). . . . .	75
4.2	System architecture for service types extraction and place ontology population . . . . .	76
4.3	A snapshot of RWPO in Protégé . . . . .	84
4.4	Survey statistics. . . . .	88
4.5	The average semantic relatedness between RWPO derived concepts and those identified in the questionnaire . . . . .	90
4.6	Sample terms comparisons using the semantic relatedness measures. The first is a term collected in the questionnaire and the second is a term in the RWPO ontology . . . . .	92
4.7	Average similarity measures of service concepts extracted for some place types compared with the benchmark set in the Ordnance Survey expert ontology . . . . .	93
5.1	Different levels of abstracting personal mobility . . . . .	97



---

5.2	Possible models of <i>PM</i> . . . . .	99
5.3	A sample PMG. Conceptual nodes are coloured in orange, spatial nodes and edges are coloured in green, and temporal nodes and edges are coloured in blue . . . . .	101
5.4	Place-affordance ontology – a snapshot. . . . .	103
5.5	GED Illustrated. . . . .	105
6.1	Partial XML encoding of the PMG for user 1 . . . . .	114
6.2	Partial XML encoding of the PMG for user 2 . . . . .	115
6.3	Partial XML encoding of the PMG for user 3 . . . . .	115
6.4	Partial XML encoding of the PMG for user 4 . . . . .	116
6.5	PGML XML schema diagram . . . . .	117
7.1	Partial visualization of the pilot profile. . . . .	119
7.2	Partial visualization of the PMG for user 1 . . . . .	122
7.3	Partial visualization of the PMG for user 2 . . . . .	123
7.4	Partial visualization of the PMG for user 3 . . . . .	124
7.5	Partial visualization of the PMG for user 4 . . . . .	125
7.6	Simulated mobile navigation guide showcasing the utilisation of PMG to adapt the selection of POI on the map . . . . .	130

## List of Tables

3.1	Possible problems of GPS data streams and their causes, data, and algorithmic effects . . . . .	48
3.2	Visited places types of collected GPS logs in Experiments 1 and 2 . . .	59
3.3	Results of constructing a PG without and with preprocessing . . . . .	61
3.4	Results with and without preprocessing and with and without filtering the candidate places . . . . .	63
3.5	Evaluation results with clusters of one candidate place filtered . . . . .	66
4.1	Data sources for place type definitions used as a construction corpus . . .	77
4.2	List of part of speech tags (POST) used for lexical analysis ((Santorini, 1990)) . . . . .	80
4.3	Sample of concepts extracted from individual data sources in the construction corpus . . . . .	85
4.4	Comparison of derived place concepts in the RWPO with an expert annotated ontology . . . . .	91
7.1	Visited places types of collected GPS logs . . . . .	121
7.2	Pilot profile concepts and their relative feedback . . . . .	129

---

7.3 Finding relevant POI in Oxford Street based on PMG. . . . . 131

# List of Algorithms

- 3.1 Overshoot points filter  
 Input: A set of tracks  $T$  and an overshoot threshold  $oT$   
 Output: Overshoot points are eliminated from  $T$  . . . . . 54
- 3.2 Segmented tracks merger  
 Input: A set of tracks  $T$  and a time threshold  $tT$   
 Output: Segmented tracks in  $T$  are merged . . . . . 55
- 3.3 Noise filter  
 Input: A set of tracks  $T$  and a noise threshold  $nT$   
 Output: Noisy tracks are eliminated from  $T$  . . . . . 56
- 4.1 Candidate Patterns Miner (CPM)  
 Input: Corpus of place types ( $pT$ ), where each  $pT$  has a set of string tokens ( $sT$ ), and each  $sT$  is annotated with its *POST* and *lemma*, no. of tokens threshold  $nT$ , and frequency threshold  $fT$   
 Output: A set of candidate patterns ( $cP$ ) . . . . . 78
- 4.2 Candidate Patterns Expansion (CPE)  
 Input: A set of annotated place types  $pT$  and a set of candidate patterns ( $cP$ )  
 Output: Each candidate pattern is expanded with its preceding *POST* in each token set in  $pT$  . . . . . 79

5.1 PMG Constructor

Input: STPM and data sources of place semantics

Output: *PMG* . . . . . 102

5.2 Personal Place Similarity

Input: PMG and a set of places  $P'$

Output: places in  $P'$  are sorted based on their GED with PMG . . . . 107

# List of Acronyms

**PG** Personal Gazetteer

**GPS** Global Positioning System

**A-GPS** Assisted Global Positioning System

**PMG** Personal Mobility Graph

**PGML** Personal Gazetteers Mark-up Language

**LBS** Location Based Services

**PDA** Personal Digital Assistant

**GIS** Geographic Information Science

**SW** Semantic Web

**MN** Markov Network

**GIR** Geographic Information Retrieval

**GED** Graph Edit Distance

**GML** Geographic Mark-up Language

**UserML** User Model Mark-up Language

**IDYP** Internet Derived Yellow Pages

**ANN** Artificial Neural Network

**GUMO** General User Model Ontology

**UMML** User Model Mark-up Language

---

# *Chapter 1*

## **Introduction**

Personalised systems aim to present only the relevant information to their users based on the user information requests and preferences. This trend of information processing is developed in response to the vast amounts of data currently available on the Web – the commonly known problem of information overload. Personalised systems on the Web have their well established architectures, algorithms, and applications. With the current developments in mobile communications, new challenges are emerging for developing personalised systems. One of the major challenges is the utilisation of the continuously changing personal mobility data.

In the context of mobile applications, the position of a person is one of the main data dimensions which are employed to personalise the available data and services to that person. The spatial coordinates of the current position as well as the historical positions of a person define the context of that person (among other elements). Personalising information based on the position of the user is what mainly differentiates mobile information personalisation from other forms of personalised systems. Mobility data streams can be clustered to extract the visited places of an individual. It can also be analysed – along with their associated temporal data – to build probabilistic models to predict the mobility of a person. However, to extract a more semantically rich profile of an individual's mobility; a profile which includes more than the spatio-temporal history or the visited places, other data sources and different analysis techniques are needed.



This work aims to produce a semantically rich profile of a person which contains spatial, temporal, and thematic information about that person, based on only the spatio-temporal history of that person. We propose an approach to build such a profile that is based on the common human activities occurring in certain geographical place types. This work involves developing the necessary algorithms to mine for these functional descriptions of places. A model, algorithms, and an encoding scheme to produce the resulting profiles are introduced. The applicability of the approach and the resulting profiles are also demonstrated.

In section 1.1 we begin by defining the research problem in subsection 1.1.1, followed by selected applications showing the scope of our research in subsection 1.1.2, limitations of the current state of the art are summarised in subsection 1.1.3, and the research hypothesis given in subsection 1.1.4. Research challenges are described in section 1.2, and an overview of the state of the art is given in section 1.3. The research objectives are given in section 1.4, followed by the research contributions in section 1.5. Finally, in section 1.6 we give an overview of the remainder of this thesis.

## **1.1 Motivation**

### **1.1.1 Research problem**

The research problem we are investigating in this work is the feasibility of extracting a semantically rich profile of a person by using only the spatio-temporal mobility history of that person. We are proposing to overlap this history with the typical services and activities performed in the visited places, to get some understanding about that person. It is valid to assume that collecting and using other data related to an individual may lead to more granular semantics (in terms of their level of detail). However, relying only on the personal mobility data has some advantages. Mainly, from a personalised system perspective, achieving a general semantic profiling and an acceptable level of

personalisation from the spatio-temporal data alone eliminates the need to log any additional data. For example, by relying on the concepts associated with visits to a of place type *university* such as *learning* or *teaching*, there is no need to log the Web searches a person takes so that these searches are to be analysed to extract the same concepts of interest to that person. This in turn overcomes many of the limitations of the other scenarios of building semantically rich mobility profiles, as will be shown on the course of this thesis. To show the scope of this problem, the next subsection overviews some selected applications related to the work of this thesis.

### 1.1.2 Scope and applications

Positional data is a key to several applications which are important in everyday life, in particular as position-aware mobile devices are becoming cheaper, and hence widely used. This subsection introduces some selected personal mobility based – or related – applications.

Some social network applications provide their users the capability to expose their current places to others. A user has to manually register the visited place, either by its address or its spatial coordinates, fetched from a Global Positioning System (GPS) receiver on a mobile phone. For example, Facebook<sup>1</sup> places API allows systems developers to build applications which utilise this feature. Location sharing platform FourSquare<sup>2</sup> is another example where users can share their visits to places with other users as well as with other social network applications such as Facebook and Twitter<sup>3</sup>. The users can also annotate the places with their comments, ratings, and so on.

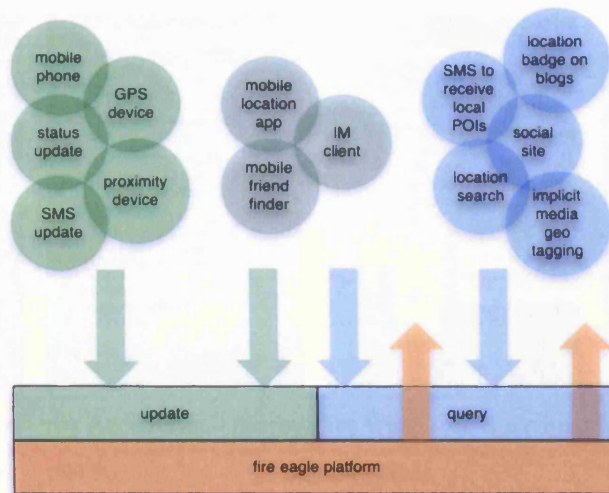
Yahoo provides Fire Eagle, a Web service which allows a user to share location data with multiple applications. A user can register their current location using different techniques (such as GPS or SMS messages), and can authorise one or more applica-

---

<sup>1</sup><http://www.facebook.com>

<sup>2</sup><https://foursquare.com>

<sup>3</sup><http://twitter.com>



**Figure 1.1: Yahoo fire eagle platform**

tions to employ this data. The platform is illustrated in Figure 1.1<sup>4</sup>. Many applications use the Fire Eagle API for various purposes, such as EagleTweet which updates Tweeter with the current location of the user, brightkite<sup>5</sup> which connects a user to nearby friends and allows to annotate photos and places, and Loki<sup>6</sup> which shares the user location with friends.

AccuTerra<sup>7</sup> is an example of a mobile application which tracks personal mobility using GPS. The application can record and visualise GPS mobility and the the places visited, and share that data with other applications such as Facebook and Google Earth<sup>8</sup>, see Figure 1.2a. Nearby places can be retrieved in relevance to the current position of the user, see Figure 1.2b. Waze<sup>9</sup> is a GPS navigation guide, see Figure 1.2c, that relies on a social network of nearby users for real-time traffic data entry. The users can feed the system with data about accidents, traffic jams, and so on which the system utilises to

<sup>4</sup>Courtesy of Yahoo Inc.

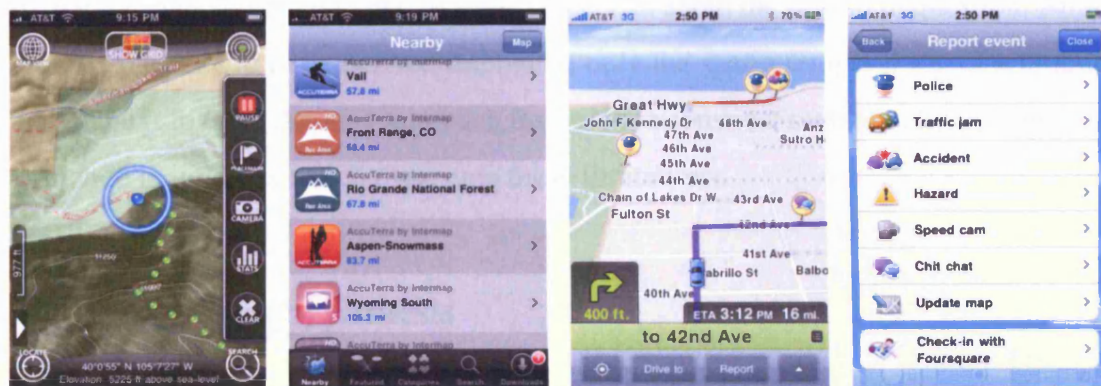
<sup>5</sup><http://brightkite.com>

<sup>6</sup><http://loki.com>

<sup>7</sup><http://mobile.accuterra.com>

<sup>8</sup><http://earth.google.com>

<sup>9</sup><http://www.waze.com>



(a) AccuTerra – track- ing GPS coordinates (b) AccuTerra – nearby places (c) Waze – GPS naviga- tion guide (d) Waze – real-time traffic data entry

**Figure 1.2: Sample personal mobility based applications**

improve its performance, see Figure 1.2d.

### 1.1.3 Current limitations

While the current state of the art in Location Based Services (LBS) and applications can automatically detect the spatio-temporal coordinates of a moving individual (the spatial coordinates as well as the time-stamp), it still generally relies on the person to manually enter other, non-spatio-temporal, data such as describing the activities taken at certain places – despite that the fact that the spatio-temporal coordinates can be automatically linked to other, often limited, data sets such as the weather information.

For instance, some of these approaches requires a Web interface, e.g. a Webpage or a Web service, to semantically describe a certain place. Although many places have their Web interfaces, many others do not. On the other hand, relying on the semantics fetched from mobile Web searches or Web services used by a person requires explicit user interactions with some system – e.g. a mobile Web search application. From a personalisation perspective, it would be more efficient to derive some semantics about that person explicitly. Hence, the former case (fetching the semantics from user interactions) require to log such interactions data (which are usually application's specific)

so that it can be analysed to derive some semantics about the person in question, while if the semantics are to be derived implicitly, only the spatio-temporal log of a mobile individual is logged. Another approach is needed to construct semantically rich profiles of mobile individuals which overcome these limitations.

#### **1.1.4 Research hypothesis**

The research hypothesis can be stated as follows. The spatio-temporal history of a person in the physical space can be analysed to extract a semantically rich profile of that individual utilising an ontology of geographical place as a sole and sufficient source of domain knowledge. The extracted profile is neither intended to be detailed, complete, nor is it a certain description of the person in question. It can be, however, employed by one or more applications to successfully personalise their results.

The following is an illustrating example. Consider a person who has a mobile phone with a GPS receiver which logs his mobility. A GPS signal is only receivable in outdoor environments, and as such its loss is a heuristic which can be employed to detect visits to buildings, such as home, work place, shops, clubs, and so on. However, a GPS signal may also be lost for other reasons, as will be shown in Chapter 3. This may cause several problems when trying to detect visits to places from GPS logs – mainly identifying false visits. As such, some procedures were needed to handle such problematic issues with personal GPS logs, to improve extracting the visited places with their times of visits, as will be shown in Chapter 3. As will be shown in Chapter 2, there are some applications which try to extract a profile from personal GPS logs which contain the visited places and their times of visits. Such profiles are very useful in many applications such as city and transportation planning, a critical review of which will be given in Chapter 2. Some applications, however, try to add more semantics into the extracted profiles by linking the visited places to associated information on the Web or by requiring the users to explicitly enter some information about the place or the visit. In our approach, we rely on the type of the visited places to provide such semantics,

without the need to log any additional semantics or any explicit data entry from the user (as mentioned in the previous section). Each of the visited places has a type, such as a school, hospital, sports club, and so on. By analysing a corpus of text describing typical human functions and services provided at certain place types, some linguistic patterns can be found to extract such place semantics (i.e. human functions and services). As will be shown in detail in Chapter 5, applying these patterns to several data sources can reveal several place semantics which are associated with place types. These semantic can be used to enrich the extracted profile (of visited places and their times of visits). For example, a person who visits a place of type university is associated with concepts such as research and higher education, a person who visits a sports club is associated with the concept of sports, and so on. However, a data model is needed to encode the resulting profile, with its spatial, temporal and semantic information, which is presented in Chapter 6 based on graph theory. The model and its associated algorithms must handle the fact the a place type maybe associated with more than one concept, as will be shown in Chapter 5. The presented profile can be used by different applications such as map personalisation applications, and a mark-up language for the profile is presented in Chapter 6 to enable such interoperability. In chapter 7, several evaluation experiments are given, which show how the resulting profile can be used to personalise a map navigation prototype. In summary, a personal GPS logs can be used to extract spatial information about a certain person (the spatial footprints of the visited places), temporal information (the times of visits), which can be linked to the semantic information (the typical human activities and services provided at the visited place types). Such a profile can be used to personalise applications such as map navigation, in an implicit manner to the end user.

## 1.2 Research challenges

To realise the research hypothesis, several challenges need to be addressed:

- **From GPS to places.** Collecting GPS mobility traces of individuals and extracting the visited places in those traces. Many of the collected logs suffered from segmentation, noisy and erroneous data, and as such appropriate pre-processing methods needed to be developed.
- **From places to visits.** Extracting and encoding the spatial and the temporal elements of visits. The users need a Web application which takes their raw GPS logs as its input, analyses it, and visualises their visits as well as providing alternative data entry methods (manual entry of places and visits).
- **Understanding place semantics.** Enhancing the concept of geographical place by inferring related concepts such as the related human activities, services, and concepts related to certain place types. To extract these terms from textual corpus, appropriate methods for information extraction needed to be developed, as well as their respective evaluation methods.
- **From visits to place semantics.** Associating each visit with its related concepts based on the the visited place type. Querying place semantics derived for the concepts associated with the visited places.
- **From visits and place semantics to a personal profile.** Modelling the resulting spatio-temporal-thematic information about a certain individual, and utilising the resulting model. Modelling the resulted profile in way which enables to utilise the profile yet avoiding hard-coding any information which can be inferred from the profile, e.g. the frequency of visits to a certain place.

## 1.3 State of the art

GPS is the main technology used to collect the mobility history of an individual using a personal mobile device, such as a mobile phone. The current technology is characterised by losing the signal inside built structures, such as buildings, and around such

structures in dense urban areas. These characteristics have their advantages and disadvantages. While losing the signal is a simple heuristic which can be utilised to detect a visit to a building, it can lead to problematic results, as it may be lost due to other reasons (and not only because of entering a building).

In the field of transportation and travel behaviour analysis, Flamm and Kaufmann introduced the notion of *personal network of usual places*; the set of geographical places which a person visits on regular bases as well as the routes taken between those places (Flamm and Kaufmann, 2007). Interviews with the participants of the data collection experiments suggested that people tend to spatially cluster their visited places around their *anchor places* (such as home and work place). This in turn affected, for example, how frequently people tend to use public transportation networks in comparison with their private transport. Nielsen and Hovgesen showed that personal mobility data collected by using GPS logging devices can reveal several interesting statistics about a certain individual (Nielsen and Hovgesen, 2004). For example, the total time of being in each of the visited places or their spatial vicinity can be found. Using other devices with GPS capabilities, such as Personal Digital Assistant (PDA), information other than the spatio-temporal can be collected, for example the purpose of a trip or a visit to a place, or the means of transport used.

However identifying the visited places and the times of their visits from GPS mobility traces is not a straight forward process. Marmasse and Schmandt developed the comMotion spatial reminder system, in which the loss of a GPS signal is the heuristic used to identify a visit to a place (Marmasse and Schmandt, 2000). It logs the spatial coordinates of a visited geographical place and relies on the user to manually label that place; simply tagging it with its name. A user then would link a list of one or more tasks to do with a certain place, so that the system alerts the user when being nearby that place again. The quality of the GPS data, namely the accuracy of the data and the availability of a GPS signal, negatively affected the process of identifying the visited places. To process GPS streams on a large scale, more reliable methods to extract the



places and their relative visits from GPS data streams needed to be developed.

Several heuristics other than losing the GPS signal can be employed to improve the place identification process. For example, several values for the spatial thresholds which define the margin of error of a place spatial boundaries can be defined. Also, several temporal thresholds which define the duration of a visit can be defined. Yet a more fundamentally different approach for extracting the visited places from GPS data streams was in need, and the answer came in the form of clustering algorithms. Successive visits to a place will form a cluster of spatial points which can be employed to define a spatial boundary for that place. However, the characteristics of GPS mobility traces required to modify the existing clustering algorithms to adapt to their dynamic nature. Zhou et al. showed that a dynamic clustering approach toward discovering Personal Gazetteer PG (a list of the visited places by a person and their relative times of visits) from mobility traces outperforms classical clustering algorithms such as the K-Means (Zhou et al., 2007). In this case, the threshold which defines a cluster is variable in response to the data, and as such the resulting clusters may vary in their number of elements inside each cluster. The set of spatio-temporal points which define a cluster would continuously change, in response to the real-world data stream.

Once the places and visits are extracted, further analysis can be performed on the results. A major direction of work is to build probabilistic models of mobility which can predict the likelihood of the next visit to a place, by a certain individual. A work by Takeuchi and Sugimoto, the City Voyager, showed how a recommender system can successfully suggest shops to visit to a certain individual based on the previously visited shops by that individual (Takeuchi and Sugimoto, 2006).

The work in the field, however, reached a limitation: a mobility trace of a person, collected with a GPS personal mobile device, can reveal the geographic places of significance to that person, along with their relative times of visits. Based on which, further analysis can be made such as the daily mobility routines or predicting the likelihood of a visit to a place. In other words, not much can be inferred about anything happening

in the visited places, unless via explicit user interaction such as filling a questionnaire or interacting with the hyperspace. What about the interests of that individual? What are the activities taken in the visited places? How can the identified places and visits be employed to improve the interaction of that person with any personalised systems, such as personalised Web search engines and mobile LBS?

As a natural evolution, other data sources were integrated into the analysis of mobility traces to obtain more semantically rich results. This is in response to the challenge that people tend to deal with places in their everyday meaning, rather than in their spatial or spatio-temporal coordinates form. This realisation has been recognised in Geographic Information Science (GIS), prior to human mobility analysis work (Rugg et al., 1997).

In general, the Web is a major source of data for that matter. Frameworks for aggregating and spatially indexing Web contents have been suggested, for example see (Himmelstein, 2005). Such systems aim to harvest the Web for all the resources which spatially refer to a certain place, and spatially index those resources. As such, a search interface would then provide the facility to search for all the data available on the Web about a certain place, turning the Web into large collection of yellow pages.

Place semantics can be provided by the public in a collaborative manner. Persson et al. developed the GeoNotes application to allow people to annotate places on the go with their ratings, descriptions, familiarity with the places, and any other comments using their mobile devices (Persson et al., 2002). A social dimension is integrated into place semantics in the form of allowing a note to be shared with friends or with the general public. The consequences include the type of place semantics to be expected in such a scenario; informal, biased, and short descriptions.

There are, however, several limitations to these approaches. Assuming that every place has its Web interface is unrealistic. While there are many commercial places which have a Web site or a Web service to their business transactions on the Web, many other places do not. Moreover, not all those Web resources include semantically rich descriptions of the functions of its places, as many of which merely includes merchand-

ise data, prices, or addresses. Another problem is that in such an approach we need to store the historical non-spatiotemporal data, such as the queries or Web services used. This adds additional data size overhead and complexities to the analysis process. Other limitations will be explored in detail in Chapter 2.

But geographical places have typical functions, which adds a human meaning to a place; an idea which had been investigated before in social and geographic sciences, for example see (Relph, 1976). The term *place affordance* was used to describe what a geographical place can provide to people in terms of functionality, services, and human activities (Jordan et al., 1998). For example, a place of type *school* is generally a place where students are taught various subjects, by teachers. Our common sense understanding of a place of the type school would associate that place with concepts like *learning, education, or teaching*. In the informatics world, however, this understanding is far from being an off-the-shelf data source. While there exist many data sources which provide some pieces of functional descriptions of geographical place types, such as Webpages or dictionary entries, there is no comprehensive system which can currently be queried for this data.

Knowledge resources that maintain information about named places have become increasingly important to support systems for geographical information retrieval (GIR). They are used to recognise place names in user queries and text documents, to disambiguate between places with the same name and to generate coordinates, i.e. geocode, for place names. They can also be used for reverse geocoding to find place names associated with coordinates and to assist in tasks of geo-data integration, particularly when matching data items that refer to the same place but with different terminology (Hill et al., 1999; Jones et al., 2001). Typical place name resources are referred to as gazetteers, geographic thesauri or place name ontologies and record at least the name and map coordinates of a place, usually in combination with the place type and its parent in a geographical or administrative hierarchy. Some of these resources are richer with regard to their conceptual, terminological and geo-spatial modelling capacity and

hence of potentially greater value in assisting with processes of GIR.

This work aims to extend the conventional content of place names resources with what is afforded at geographical places; in particular the sorts of services and the opportunities for human activities that are associated with different types of place. For instance, a *school* affords *learning* and *teaching*, while a *bank* affords the services for *storing*, *saving*, *accessing*, and *borrowing money*. The potential significance of maintaining a knowledge of geographical places affordance has long been recognised, see (Jordan et al., 1998). However, with the increasing demand for geographically based information services, there is a strong motivation to encode the geographical place affordance as an integral component of place name knowledge resources. Such knowledge resources have the potential to enhance the existing GIR services, helping to find geographical places with particular services suitable for certain personal interests or requirements.

Our approach to affordance-based enrichment of place ontologies is based on mining knowledge of services and service types from a variety of geographically focussed resources, starting from an existing typology of place types. We use a single relatively rich source of place service descriptions, the learning corpus, to identify more frequent affordance language patterns. These patterns have been applied to several other resources, the construction corpus, to extract place-type specific services and activities. The results have been integrated within an affordance-enriched place ontology that associates place types with their typical affordances. An evaluation of the resulting ontology has been performed using a web based survey of typical activities that people regard as taking place at, or being provided by, particular place types. The approach has proven to produce good results for the extracted service types.

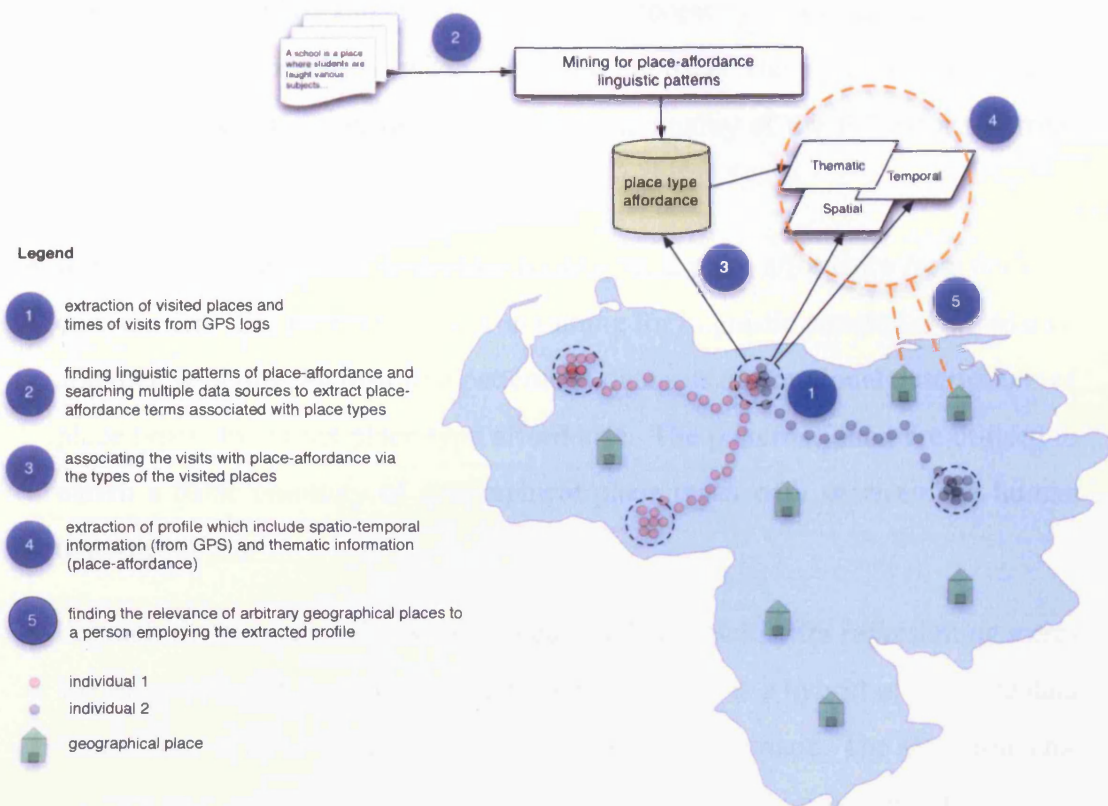
On the other hand, overlapping personal mobility with geographical place semantics raises another challenge in itself, which is how to model the resulting profile. Such a profile will contain spatio-temporal information about the visited places as well as thematic information; the concepts reflecting place affordance. As such a profile is

more likely to be deployed and utilised in a mobile environment, which is limited in terms of its computing resources, it is critical that the design of such a profile, its constructing and maintaining algorithms and its encoding format take this limitation into account. In other words, it is desirable that its design and algorithms are made as light as possible, as well as generic as possible to avoid any unnecessary interoperability problems.

## 1.4 Research objectives

Our **research methodology** can be summarised as follows. The first step is to analyse the mobility traces of an individual, collected by a personal GPS device, to extract the geographical places of significance to that person. A mapping is then made between those places to their place affordance to build a semantically rich profile of that individual, which can be employed in personalisation applications. Figure 1.3 illustrates the research framework.

- First, the GPS logs of an individual are analysed to extract a set of visited places and their relevant times of visits – hence, building a PG. This step involves handling any erroneous, noisy, and missing mobility data. See step 1 in Figure 1.3.
- The second step is to build a repository of geographical place types affordance. This step involves mining for linguistic patterns of place types affordance, and searching multiple data sets of functional descriptions of geographical place types, for those patterns. The result is a data source which associates a place type with the typical human activities and services provided in a place of such a type; see step 2 in Figure 1.3.
- A mapping is then made between the visited places and their respective place affordance, see step 3 in Figure 1.3.



**Figure 1.3: Research framework**

- A profile is built from hybrid spatial and temporal elements – driven from the GPS data, integrated with thematic information which is retrieved from the place affordance data source, see step 4 in Figure 1.3.
- This profile is employed to calculate the relevance of arbitrary geographical places to a certain person, see step 5 in Figure 1.3.

## 1.5 Research contributions

The contributions of this thesis can be summarised as follows.

- Identifying several problems with GPS data streams collected by personal mo-

mobile devices, and introducing pragmatic pre-processing procedures to reduce the effects of these problems, when analysing the data. The suggested procedures show good results in terms of improving the quality of the PG extracted from GPS logs.

- Introducing a method for extracting geographical place affordance from multiple data sources. The method is based on mining for linguistic patterns of affordance and then applying the resulting patterns to data sets of functional descriptions of place types, to extract place type affordance. The patterns found are utilised to enrich a basic ontology of geographical place types with services and human activities types.
- A graph-theoretical model is introduced as a data structure for representing a profile of mobility of an individual. This representation is a hybrid of multiple data dimensions of a PG – namely spatial, temporal and thematic. The algorithms for constructing and querying such a graph is presented and implemented.
- Showing that semantically rich profiles of an individual can be extracted solely from spatio-temporal history, representing the mobility of that individual. This technique eliminates the need to log non-spatio-temporal data elements to build such a profile.
- Suggesting a personal similarity measure of geographical places. This is an integrated measure of multiple heuristics of human mobility, which relate a geographical place to a person based on the mobility of that person – captured in PG. An algorithm which calculates this similarity is also provided, and shown how it can be employed in a prototype application.
- Finally, introducing a mark-up language for encoding PG in a unified format. The aim is avoid an anticipated problem of interoperability between PG based applications. The mark-up language is based on the graph theoretical model suggested.

## 1.6 Structure of the thesis

The rest of this thesis is structured as follows. Chapter 2 provides some background and reviews the literature on profiles for mobile and context-aware personalised systems, followed by a review of the literature on human mobility analysis, ontologies of geographical place, and extracting geographical place affordance. Chapter 3 presents the proposed procedures for pre-processing GPS data streams, along with their implementation results and application. Chapter 4 describes our natural language processing approach to mine for linguistic patterns of geographical place affordance, and presents the resulting patterns. Chapter 5 introduces our graph theoretical model of personal mobility profiles, its constructing algorithm, and the suggested personal place similarity measure, its heuristics, and utilisation algorithm. Chapter 6 describes our proposed mark-up language for encoding PG. Chapter 7 presents discussion and analysis of different evaluation experiments of the results of the mobility graph model and the personal place similarity measure and their application. Finally, Chapter 8 summarises the results and contributions, and looks at research directions for future work resulting from this thesis.



## **Background and related work**

The work presented in this thesis is built on a background in a range of research areas and technologies. These include personalised information systems, the collection, modelling, analysis, and applications of human mobility traces, knowledge representation of geographic places, and lexical analysis and textual patterns recognised from multiple data sources of natural language text.

This chapter introduces the necessary background and reviews the related work. In section 2.1, we begin with an introduction to personalised information systems, which are presented as a specialisation of context-aware information systems, given in subsection 2.1.1. In subsection 2.1.2, an emphasis is made on profiles, and in particular a special type of profiles; those which accumulate the spatio-temporal history of a mobile individual, given in subsection 2.1.3. An introduction to the related work on human mobility traces analysis is given in section 2.2, followed by a critical analysis of the major directions in the field in sections 2.3 and 2.4. Section 2.5 defines the boundaries of related work, giving examples of work which is considered out of scope for this thesis. Limitations of the state of the art is summarised in section 2.6, followed by a brief on the other techniques in human mobility analysis which are out of scope for the purpose of this work. Finally, the related work on extracting semantics of geographical places is given in section 2.7. The chapter ends with a summary in section 2.8.

## 2.1 Personalised systems

A personalised information system is an adaptive system that adapts its output based on some information about the user. In other words, while more than one user can provide the same input to the system, the system may provide a different output for each user to satisfy the need of that particular person, based on some information about each of the users. The system *personalises* its output for a certain individual in the aim of providing a more accurate answer to a user query. A few years after the emergence of the Web, vast amounts of information were available online, resulting in what became known as the problem of information overload. As users needed to access only the information relevant to their information needs, several personalisation techniques were developed (Anand and Mobasher, 2005). There is no standard definition for personalisation, the following, however, are several examples. In (Anand and Mobasher, 2005), personalisation on the Web "*is viewed as an application of data mining and machine learning techniques to build models of user behaviour that can be applied to the task of predicting user needs and adapting future interactions with the ultimate goal of improved user satisfaction*". Another view is that a personalised system aims to provide its users with their information needs, with the least explicit interactions from the users (Mulvenna et al., 2000). In (Shahabi and Chen, 2003), personalisation on the Web is divided into recommender systems (where information is customised based on its similarity with other information) and personalised Web search (customising the results of a user's query based on the user's preference, previous queries, or other users' queries). A wider definition is given in (Searby, 2003), as "*whenever something is modified in its configuration or behaviour by information about the user, this is personalisation*". This definition considers all the possible information about a user, such as location." In the remainder of this section, we survey selected examples of personalised systems.

A typical personalised system on the Web is a personalised Web search engine. A personalised Web search engine takes a user query as its input, and it employs other data

as well to customise its output to that particular user, such as the user preferences or the user Web search history. Examples of real-world personalised Web search engines include iGoogle<sup>1</sup>, and My Yahoo!<sup>2</sup>. Many configurations of the process exist, including whether the data about the user – or the user profile – are collected explicitly or implicitly, whether the profile is generated and stored on the client side or the server side, and which algorithms are applied to personalise the output. Mainly, there exist three approaches to personalise Web search results. *Re-ranking*, where the search results are reordered according to their relevance to the user, *query modification*, where the search query is modified – typically by expanding it with more search keywords – to achieve better results, and finally *filtering*, where the irrelevant results to a certain user are omitted from the output (Keenoy and Levene, 2003). Other, less common, approaches include clustering the search results to ease the browsing of the results, as used in the Vivisimo Web search engine<sup>3</sup>.

Another variation is the work in (Ferragina and Gulli, 2008), where the Web search results from multiple Web search engines are integrated, analysed, and clustered in a hierarchy on the fly, see Figure 2.1. Notably, there is no user profile constructed and maintained in this approach. A Web search query returns a set of results, each of which includes a relatively short textual description of the retrieved item – this paragraph is referred to as a *snippet*. By applying linguistic analysis techniques on the snippets, single word or multiple words labels are generated, and those labels are used as clusters, which the Web search results are grouped into. When a user selects a cluster, all the results which are not related to this cluster are filtered-out. Moreover, the clusters are arranged in a hierarchy, and re-ranked according to their relevance to the user selection (of a particular cluster), as well as to their relevance to knowledge-base of analysed text and Web pages' links (which were analysed off-line).

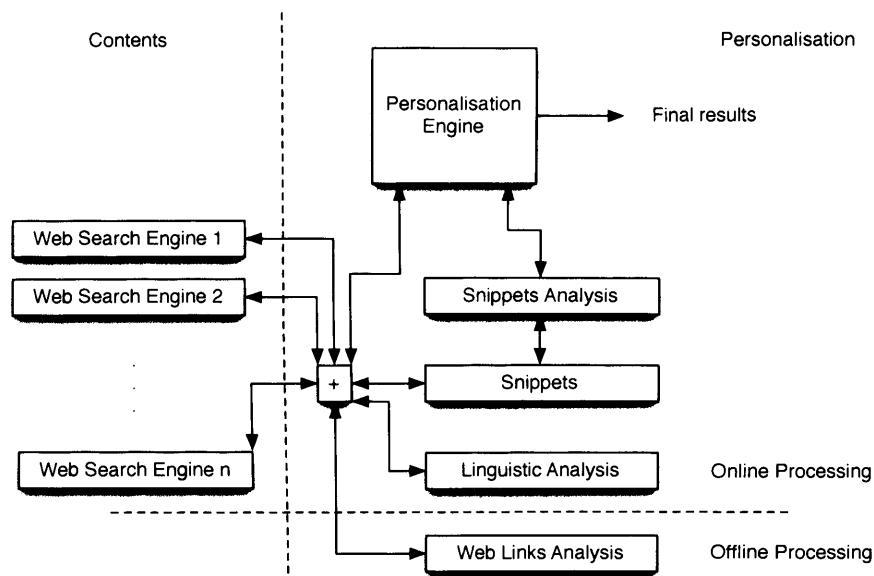
Mobasher and Cooley pioneered the field of Web usage mining for the purpose of per-

---

<sup>1</sup><http://www.google.com/ig?hl=en>

<sup>2</sup><http://my.yahoo.com>

<sup>3</sup><http://demos.vivisimo.com/projects/BioMed>



**Figure 2.1: Architecture of a sample personalised Web search engine – simplified**

sonalisation (Mobasher et al., 2000). The basic idea is to mine the log of a Web server for frequent navigation patterns within a certain Website. These usage patterns are then employed to build a predictive model of Web navigation for that particular Website, or to recommend the next Webpage to visit for a certain user based on the current Webpage being browsed. Unlike other personalisation techniques, this approach is unique in the sense that it is not based on a profile of static preferences, but it relies on the navigation behaviour of the users to find out the frequent navigation routes, which is a dynamic process. In (Eirinaki and Vazirgiannis, 2003), other examples of Web personalisation are provided, including a list of commercial tools, where the navigation behaviour, the user profile and the context of the content are analysed to improve the response to the end user.

In summary, the basic components of a personalised system are the following. First, the *contents* which the system retrieves such as the contents of a Webpage. Note that this is not limited to the text in a Webpage, but can also include other elements such as photos, videos, the link structure between different Webpages, and meta-data elements.

Second, the preferences and other usage data which the system keeps about each of its users, or the *profile*. Finally, one or more modules which apply the necessary processes to personalise the results, including the modules which are responsible for constructing and maintaining the profile, as well as other modules for monitoring the personalisation results. For further reading, see (Shahabi and Chen, 2003; Anand and Mobasher, 2005; Pretschner and Gauch, 1999; Searby, 2003) where more details are provided on different views, classifications, methods, scalability issues, privacy issues, software architectures and research challenges of personalisation.

### 2.1.1 Context-aware computing

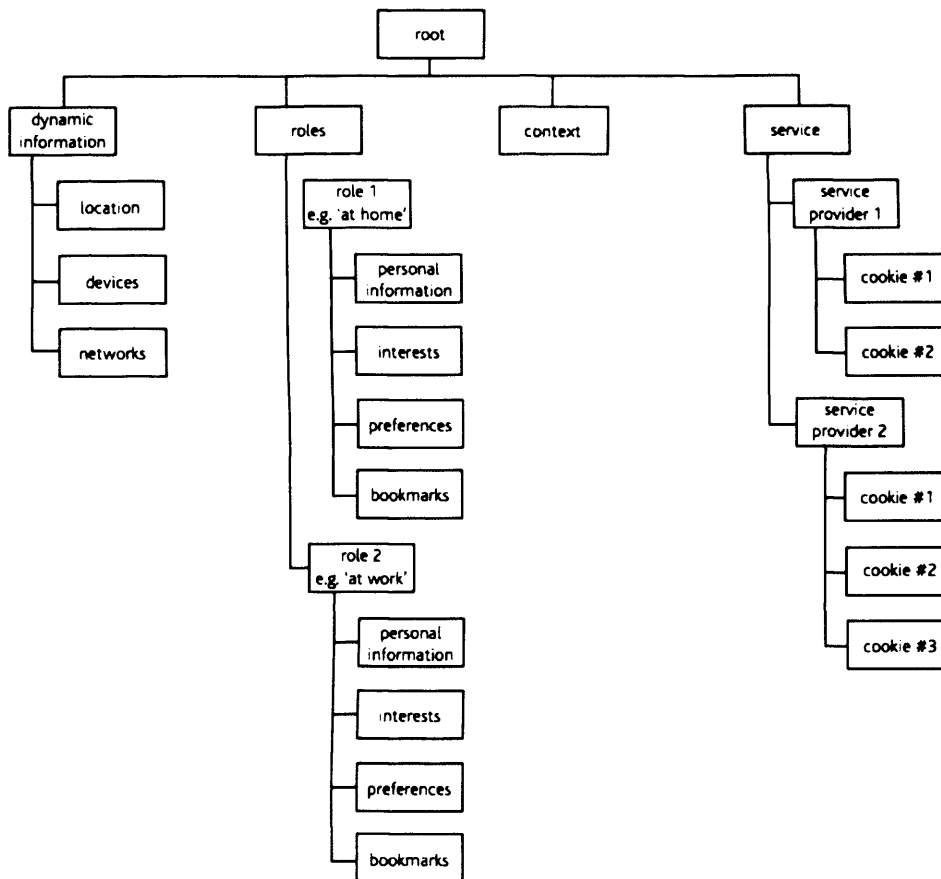
A context-aware information system is a generalisation of a personalised information system. Similar to a personalised system, a context-aware system aims to provide only the relevant data and services to a certain individual, based on some data about that individual. Unlike a personalised system, any data can be employed for that purpose – not necessarily directly related to the person or the system. For example, in a personalised Web search application, the data employed for personalisation may include the search query, search results, historical search queries and results, and the user preferences. In a context-aware search application, other data may be employed as well such as the current position of the user. Context can be defined by several dimensions, such as the location of the user, the computing resources available to the user in the current location, or even the people around the user. Accordingly, the data and the services available to the user are adapted to suit the current context, in a manner which is implicit to the user (Schilit et al., 2008).

Sleker and Burleson showed several example prototypes of embedding different types of sensors in a variety of clothes, furniture, and everyday personal tools, all of which are used to collect multiple dimensions of context data (Selker and Burleson, 2000). What characterises the notion of a context, however, is far from being a standard. As discussed in (Dey and Abowd, 2000), it can include the location, temperature, time

of the day, season, people around the user, nearby objects, their changes of status, orientation of the user, the emotional status of the user, among other dimensions. Another characterisation can include the application environment variables (processors, network capacity, cost of computing and so on), the user environment variables (location, nearby people, and any data related to the user), and the physical environment variables (such as many of the variables mentioned, or similar like lighting level). One particular element of a context definition which is of interest to the course of our work is the activity which is carried out by the user. Considering such a wide variety of data dimensions in the personalisation process has its effects on the way context-aware systems are designed and implemented, as the data of many context dimensions are continuously changing. An example is a mobile navigation guide application, which utilises the context of the user to adapt the map shown on a personal mobile device to be suitable to the situation of the user. Unlike other map-based applications on non-mobile computers, a mobile map-based application employs dynamic context elements such as the current task at hand, the weather, and the physical soundings of the user, and adapts the rendering of the map accordingly (Dransch, 2005). Interestingly, Spence et al. suggests to share the context historical data (log) between multiple users and applications (Spence et al., 2005). This work shows the need for a common data representation for profiles which are to be used in such types of applications.

Nivala and Sarjakoski defined several context dimensions for a person using a map-based application on a mobile device, in particular (Nivala and Sarjakoski, 2003a,b). These included the device computing resources, the task at hand, the user preferences, nearby people, cultural settings (language, time zone, etc.), physical surroundings, location, orientation, time, and navigation history. In general, it is fair to say that the context can be defined by any data which can characterise the system, the person who is using it, or the environment in which the user or the system runs – all of which are utilised to improve the system performance.

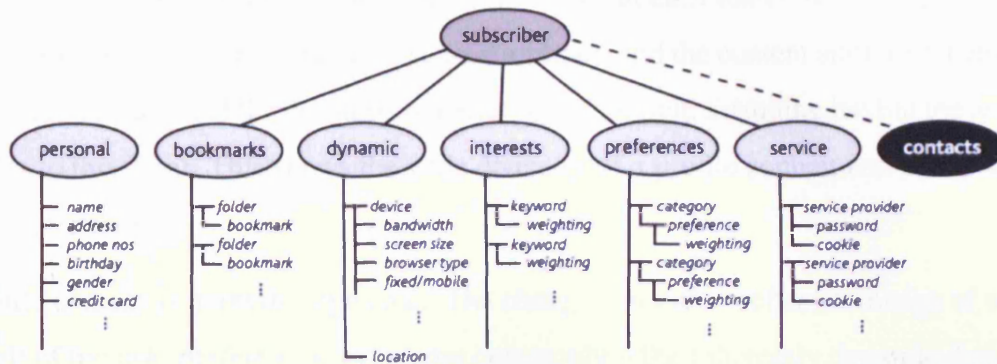
### 2.1.2 Personal profile



**Figure 2.2: A sample profile – (Newbould and Collingridge, 2003)**

A system can perform a personalisation process without the need to keep any data about the user, for a single session of the application usage. To maintain a steady improving in the results, a personalised system needs to keep some data about the user to be employed in multiple sessions of the application. This data set is referred to as the *profile*. “The profile will generally differ from system to system, storing different pieces of information in various proprietary formats depending on the exact functionality and the algorithms used” (Keenoy and Levene, 2003). In all cases, the profile is utilised to personalise the user experience of using a system. Figures 2.2 and 2.3 show sample profile from (Newbould and Collingridge, 2003) and (Searby, 2003),

respectively.



**Figure 2.3: A sample profile – (Searby, 2003)**

In general, the contents of a profile can be classified into two groups of data sets. The first data set encodes the demographic data of a certain individual, such as the name, age, occupation, preferred food, and so on. The other data set is the historical application usage data; i.e. the log of the interactions of that person with the system. For example, the previous search keywords and their relative results in a personalised Web search engine. Moreover, the latter part of the profile may also contain the analysed version of the usage data, such as a weighted vector of keywords which represent the user interests derived from the historical Web search keywords. It is a common practise that a profile is combination of the two.

In a similar manner, a profile can either be constructed explicitly, implicitly, or as a hybrid of both approaches. In the first case, the user is asked to explicitly enter the profile data to the systems, and typically this is the preferences part of a profile. In the second case, the system implicitly monitors the user interactions data with the system, analyses it, and logs the analysed results. In general, achieving the personalisation process implicitly is more efficient than the achieving it explicitly – including the construction of the profile. A well-known measure of the performance of a personalised Web search engine is the number of mouse clicks a user needs to take in order to reach a satisfactory result. In (Teevan et al., 2010), a number of data elements are impli-



citly and explicitly monitored and analysed in a collaborative manner to improve the Web search results. These are the user feedback about each search result (explicit), the clicks the user takes on the search results (implicit), and the content similarity between the search results and the desktop computer data (implicit, assuming having the access rights to this data). This shows the wide diversity of a profile content and personalisation algorithms.

Profile content is normally dynamic. The change can either reflect a change of some or all of the user preferences, or – more commonly – the inherently dynamic data elements of a profile, such as the user position or the interactions data with the system. Algorithms for maintaining the content of a profile are a necessary part of a personalised system. Moreover, a profile can be constructed in a collaborative manner. In other words, the preferences and the behaviour data of the other users are considered as well when deriving the profile of a certain user (Crossley et al., 2003). This is related to whether or not a profile is designed to be shared between multiple personalisation systems, as will be explained in the next paragraph.

Finally, personalisation as a technology is currently heavily employed by a large number of service providers on the Web. Examples include some of the major service providers such as the *Amazon* e-commerce company<sup>4</sup>, where costumers are offered suggestions of items to buy based on theirs as well as other costumers previous purchases. Another example is the *iGoogle* personalised interface to the *Google* Web search engine<sup>5</sup>. As such, it is currently common that a single individual has more than one profile, which are utilised by different systems. Designing profiles to be as interoperable as possible is a business need which raises many research challenges (Schuurmans and Zijlstra, 2004). Some of those challenges include allowing a user full control of all the contents of his profiles (in different domains such as commercial, healthcare, education, etc...), the synchronisation of contents of such profiles and privacy issues. In that regard, it is worthwhile to mention that personalisation is not

---

<sup>4</sup><http://www.amazon.com>

<sup>5</sup><http://www.google.com>

necessarily restricted to Internet-based systems (Riecken, 2000).

### 2.1.3 Personal Gazetteers (PG)

One particular form of profile which is currently an active research and applications area is the Personal Gazetteer (PG). Similar to other forms of profiles, no specific definition of a PG content or structure exists. The term was originally coined by Zhou et al. (Zhou et al., 2004).

As mentioned, one of the major context dimensions which are logged in a profile for a mobile personalised system is the user position. Additionally, it is common that the position data is augmented with its time-stamp. A spatio-temporal history of an individual can be analysed to extract a list of the visited places by that person and their relative times of visits – a PG.

The word *personal* is used to indicate that the device used to collect the spatio-temporal history is a personal mobile device, such as a mobile phone, and the word *gazetteer* is used to indicate that a PG is a dictionary of the visited geographic places, and not merely a set of raw spatial co-ordinates.

The next section introduces the related work of extracting and employing PG.

## 2.2 Analysing human mobility

The instantaneous spatio-temporal co-ordinates of a mobile person have been employed in a variety of classical LBS (such as navigation, advertising, games, emergency response, etc...) to characterize the context of a mobile person (Steiniger et al., 2006; D’Roza and Bilchev, 2003; Adamczyk et al., 2007; D’Roza and Bilchev, 2003). Example applications include real-time monitoring and management of vehicles (Lei and 0002, 2006).

Collecting such data for a period of time is currently feasible due to the availability of spatially-aware personal mobile devices, such as mobile phones<sup>6</sup>. Typically, the GPS is the technology employed to capture the positional data of a mobile person. Current technology is characterised by its inaccurate readings (positional readings are within a diameter of about 10 meters error tolerance), and that a GPS signal can not be received without a direct line of sight to a minimum number of satellites. In other words, it is not receivable inside built structures or may be not receivable in the vicinity of such structures, for example, as in the case of a downtown area of a typical modern city where a person is usually surrounded by a number of multiple-story buildings, what is referred to in the literature as *urban canyons* (Marmasse and Schmandt, 2000; Moreira and Santos, 2005). These characteristics affect the way personal mobility data are analysed, as well as the results of such analysis, in particular for constructing PG, as will be given.

Analysing this data can reveal several statistics about a user, such as the geographical places which that person visits on regular basis, the times of such visits, their routes taken, and the spatial area in which that person typically moves. In (Mountain and Raper, 2001), the spatio-temporal history of a user is visualised to help determining spatio-temporal episodes, such as repetitive visits to certain places, in an interactive manner. Several configurations of such episodes can be found, by setting different spatial or temporal limits which define an episode. Such results are utilised by a wide variety of applications, such as city and transportation planning, for example. Schonfelder et al. conducted several studies for traveller behaviour analysis which aimed to identify pattern of human mobility traces (Schonfelder et al., 2002; Schonfelder and Samaga, 2003). Such research helped to improve applications such as traffic safety, understudying routes and means of transport choices made by the users, for example. Flamm and Kaufmann emphasised the personal view in such studies (Flamm and Kaufmann,

---

<sup>6</sup>In addition to a GPS receiver to determine the instantaneous location, a spatially-aware mobile phone may also have a digital compass or an accelerometer sensors, allowing for a range of novel applications, for example see (Simon et al., 2006).

2007). Their research showed that, for each individual, the routes taken and the visited places were correlated with the locations of home and work place of a person who provided the mobility data. Other research aimed to analyse spatio-temporal history of a group of people in parks and site-seeing and tourism places, in order to determine optimal walking paths and buildings' locations, for example (O'Connor et al., 2005). In (Nielsen and Hovgesen, 2004), it was shown that using devices with more than only GPS capability, such as voice recording or taking digital notes, can reveal deeper understanding into human mobility behaviour. Devices such as modern mobile phones or PDA can, for example, log the purpose of a visit to a place.

From a personal applications perspective, a PG is suggested as a way to improve way-finding applications – the assumption is that a person's familiarity with a geographic area can be employed to provide a more efficient navigation guide (Schmid and Richter, 2006). In general, a data set representing the mobility trace of a certain individual can be analysed either *i*) to reveal some understanding about that person, or *ii*) to predict the next spatio-temporal point for that person (in other words, the likelihood of the next place to visit, and the time of such a visit) , where the latter direction of analysis typically employs the results of the former.

It has been shown that a person's preferences of geographic places are correlated to their frequencies and times of visits by that person (Froehlich et al., 2006), in other words their PG. A PG can be constructed by different approaches, and are built for different purposes. In general, either the spatio-temporal history is solely employed to extract the PG, or some additional data sources are integrated into the construction process, to help building a more informative PG. The following sections critically review a selection of works on analysing personal mobility trails, within the context of constructing PG.

## 2.3 Statistical analysis

The simplest technique to construct a PG from a set of spatio-temporal points, provided via a GPS equipped personal mobile device, is to explicitly and manually or semi-automatically identify the places and their relative visits. The *comMotion* spatially-aware reminder prototype learns the frequently visited spatial co-ordinates and explicitly asks to *label* an identified place by the user (simply tag it with a name) (Marmasse and Schmandt, 2000). Next, textual and auditory notes are to be manually associated with places.

Toward an automated extraction of PG, simple spatio-temporal thresholds can be employed to identify the places and their relative visits. In Project Lachesis (Hariharan and Toyama, 2004), the notions of a place, a route, and a visit are defined by repetitive occurrences of their spatio-temporal co-ordinates, and within the limits of a minimum and a maximum spatial and temporal threshold. For example, a place is defined by multiple occurrences of spatial points within a certain spatial distance, at different times (indicating different visits to that place). Depending on the application, different values can be used for those parameters.

Clustering a personal mobility log can automatically extract a PG to a good extent, with some clustering algorithms aiming to classify the derived places into certain types. Simple heuristics can be employed to define the notions of *home* and *work place* considering typical human behaviour, for example see (Moreira and Santos, 2005). Home, for example, is typically a place that is visited over night, almost everyday.

The K-means algorithm is used to cluster a set of spatio-temporal points to extract a PG in (Ashbrook and Starner, 2003). The idea is to randomly assign each point in the set to a cluster, find the centre of each a cluster, and to minimise the error in a cluster in an iterative manner, where the error is the sum of the squared distance of each point in the cluster to its centre. The algorithm is non-deterministic, as its output depends on the initial assignment, and is sensitive to noise, as erroneous points may affect the

centres of clusters. Another limitation is that it requires the number of clusters – places in this case – to be set prior to any processing, which is unrealistic. Regardless of those computational limitations, no distinction is made between the extracted places in terms of their place-type.

Density-based clustering is used as an alternative in (Zhou et al., 2004, 2007, 2005b). In this approach, a cluster is defined with two constraints. First, each two distinct data items in that cluster are within a certain maximum distance of each other, and secondly a set of points forms a cluster if it has a minimum number of data items. While this approach provided improved results compared to the K-means algorithm in terms of the discovered places, it is – similar to the K-means – limited in terms of differentiating those places.

The K-means and DBSCAN (Birant and Kut, 2007) clustering algorithms were used to extract PG in (Lim and Bhatnagar, 2008), while Neural Networks, Naive Bayes and Decision Trees were used to build predictive models to determine the importance of a visited place to a certain individual, based on the frequency of visits and the times of visits. The study found that while some places were frequently visited, others were labelled important and some places were both important and frequently visited. In (Nurmi and Bhattacharya, 2008), the Bayesian analysis showed an improvement in comparison with the K-means analysis for automatically extracting PG, as the former did not require any parameters initialisation.

Some progress has been made applying other machine learning algorithms, as in (Liao et al., 2006, 2005) where a variation of the Markov Network (MN) is used. Following the extraction of a PG, machine learning techniques are applied in order to semantically label the extracted places – specifically by the activities taken in those places. The process required manual data entry from the people who provided their mobility logs, to train the MN. Moreover, the results were limited to simple descriptions to a few types of the visited places, namely places in which a person has labelled as either being *at home*, *at work*, *shopping*, *dining out*, or *visiting*. Other places not falling into

these categories were not identified.

A PG construction algorithm can be designed to detect certain types of places, and as such the resulting PG only includes the visits to places of such types. The *CityVoyager* prototype analyses the mobility history to find the frequently visited shops by a certain individual, and recommends other shops to visit accordingly (Takeuchi and Sugimoto, 2006). A candidate visit is detected when a GPS signal is lost for a certain time limit and within a certain spatial limit nearby a shop, where the actual locations of the shops are known in advance. On the opposite, when a GPS signal is detected again for a certain time limit, this indicates being outdoors. Moreover, the routes taken are also taken into account (along with the frequency of visits) to determine a rate for each shop, which in turn is taken into account when recommending a shop to visit. The work identified the need to define multiple properties for each place – shop in that case – which can be employed to provide finer recommendations. In other words, *something more than the place type needed to be integrated into the analysis*. Other work on recommendation systems associated the position of a mobile person with their Web navigation experience, over a period of time, in order to provide more useful recommendations (Brunato and Battiti, 2003).

Relying solely on the spatio-temporal history of a person and quantitative analysis techniques, a PG of that person can be extracted, to a good extent. Moreover, a predictive model for that person can also be built, which can predict the likelihood of the next visit by that person. However, the resulting PG is limited in terms of the semantics of its contents. The majority of the literature focus on correctly identifying the visited places, while it falls short in revealing any understanding about the person who visited those places. For example, the previous methods may well learn to identify the work place of a person, as well as predict the next visit to that place, but it can not say much about what type of work this person undertakes. A PG built employing only the mobility history and quantitative analysis techniques is limited in terms of its qualitative elements – which is identified as one of the major challenges of interpreting positional

data (Hightower, 2003).

## 2.4 Semantic analysis

One way to build more informative PG, specifically in terms of the semantics of their contents, is to integrate additional data sources into the PG construction process. The idea is to associate a visit to a place with other data either related to the person, the place, or both. As such, a resulting PG not only includes the visited places with their relative times of visits, but also it logs, for example, the services used by a certain individual. This type of personal mobility analysis can be classified as *hybrid analysis*, as it includes additional data sources other than the spatio-temporal mobility data, as well as utilising different data analysis techniques. A wide variety of data sources can be employed via different techniques, as follows.

In (Stopher et al., 2007), the participants in a mobility log data collection experiment were asked to explicitly paper-log their spatio-temporal history, along with the type of the visited places (to be selected from a fixed set of either being a *home*, a *work place*, a *school*, or a *grocery shop* – if applicable). The aim was to identify the purpose of a trip taken by some individual, to support transportation planning.

In general, the Web represents a rich data source to extract geo-referenced data. Web contents are either explicitly or implicitly geo-referenced. Frameworks for aggregating and geo-referencing Web contents, such as the Internet Derived Yellow Pages (IDYP) (Himmelstein, 2005), collect the data referring to the same geographical place from multiple Web sources, and provide a spatial search interface for that data. For examples, see (Morimoto et al., 2003; Amitay et al., 2004; Silva et al., 2006). Another way is to directly expose the information about geographical places to the Web. Examples include any classical Webpage of a certain place such as a restaurant, a coffee shop, an hotel, and so on. Moreover, in project *CoolTown* (Pradhan, 2000; Pradhan et al., 2001) a place is explicitly assigned a unique URI, and as such it can be ac-



cessed from the Web along with its associated data and Web services. In those cases, the assumption is that each place either already has its Web interface (typically one or more Webpages which contains some textual descriptions of that place), or it has to be registered for one (as the case in project *CoolTown*).

Semantics associated with geographical places can also be provided by the general public, and not necessarily by governmental or other formal organisations. As such, that data can be any descriptions about a certain place, such as some experiences some people had at that place. This requires to establish the necessary communication channels which facilitate an easy to use place tagging interfaces on the social Web. For examples, see (Yu and Shaw, 2008; Ito et al., 2005). In (Scharl et al., 2008) a framework for aggregating geo-referenced data on the Web and then presenting it for manual annotation is introduced. On the other hand, visual analytics tools allow to visualise a trace of personal mobility, identifying the PG it contains, and annotate it for various analytical purposes, as in (Andrienko et al., 2007; Kapler et al., 2005). Moreover, recent advances in personal mobile devices allowed to have places annotated on the go. In such a scenario, a personal mobile device (such as a mobile phone with Internet connectivity and spatial-awareness) is used to associate an activity with either the arriving to or the leaving from a place (Sohn et al., 2005), for example. Tagging places via mobile devices can also include ratings, comments, informal descriptions, recommendations, familiarity with a place, and whether or not to share the tag with other people, or any other note (Persson et al., 2002; Zhou et al., 2005a). As those place descriptions are typically provided by the general public in natural language, general purpose knowledge-bases can be utilised to linguistically analyse such data, and extract structured place semantics, for example see (Hockenberry and Selker, 2006). Other approaches of tagging places via mobile devices include auditory (associating a voice note with a place), visual (associating a photo with a place), and map-based annotations (employing a map to choose a place from, and annotate it) (Wang and Canny, 2006). In (Jones et al., 2007), historical mobile search queries are correlated to places, to improve future searches. One of the interesting findings of the research is that “*queries give a*

*sense of place*" – for the context of this work, mobile search queries are a possible data source to mine for place semantics. A similar work is (Pramudiono et al., 2002). Moreover, some research work assumed the accessibility to various data sources on a mobile device, such as the calendar, address book, phone calls log, navigation routes, to mine for place semantics, for examples see (Liu et al., 2006; Hinze and Voisard, 2003).

Another direction of work is to correlate a spatio-temporal human mobility trace with other data which implicitly share the same spatio-temporal history. For example, in *personal life log* applications, the data collected from multiple sensors, such as audio-visual, GPS, weather detection sensors, and accelerometers are fused to compose a multidimensional view of someone's life experiences, where the places and visits discovered can be associated with the other data collected. For examples, see (Aizawa et al., 2004; Kim et al., 2006a; Saponas et al., 2008). Interestingly, in (Kim et al., 2006b) an attempt to semi-automatically *classify* the activities taken by a certain individual is made, via analysing their life-log content. By analysing the data collected from multiple sensors (for example, an Artificial Neural Network (ANN) is trained by pictures of objects found in typical home and work indoor environments, to detect a person's presence in such places), the activities related to *being at home* and *being at the work place* could automatically be inferred. The other activities, however, needed some manual data entry of the activities actually taken, to be used as a training data set to the machine learning module.

Having qualitative data elements in a PG can facilitate the logical reasoning and semantic matching of its content. Ontologies were proposed as a means for facilitating logical reasoning in context-aware environments, modelling various elements of such environments such as space, time, events, activities and so on. For examples see (Chen et al., 2004; Wang et al., 2004). Similar to some of the work mentioned earlier, in (WeiBenberg et al., 2006), the spatio-temporal history along with the services used are utilised to derive *situations*, where the context of an individual is not changed for

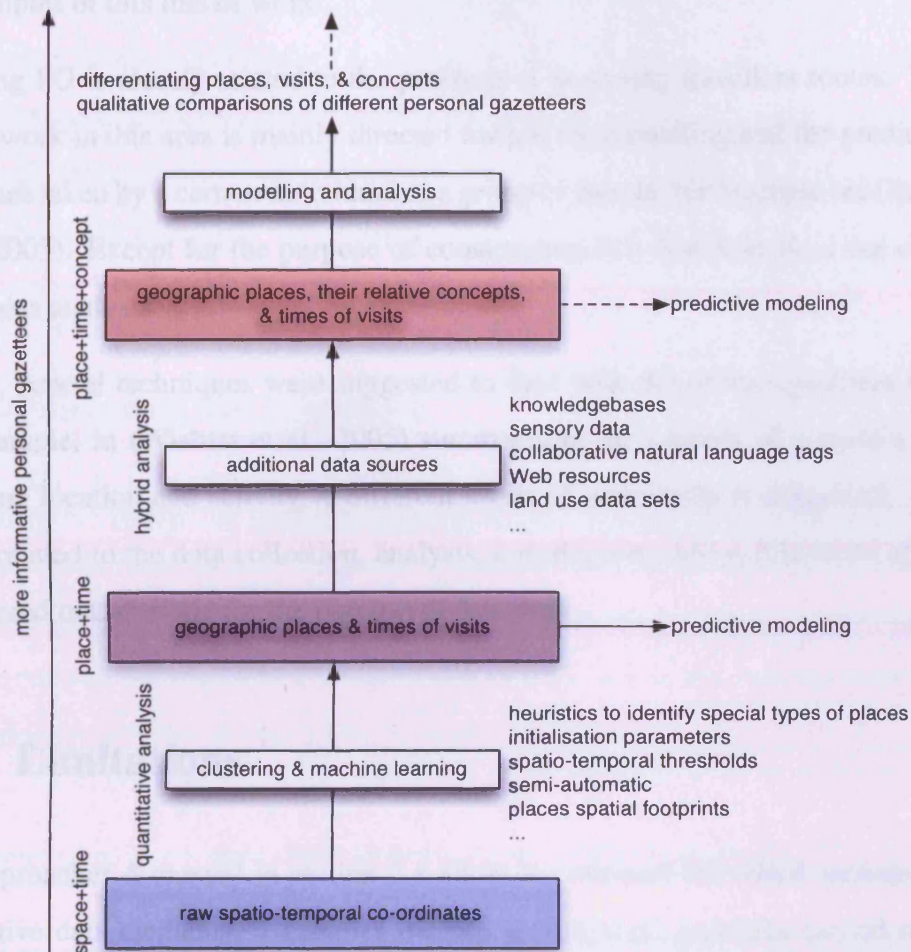
a period of time. Examples include driving a car or watching a football match. The deduced status is in turn utilised to semantically match the available information and services, to the current situation. Similar to the previous work, the assumption is that all the data elements are accessible, a condition which is usually satisfied for a specific application. Examples of similar work include (Zipf and Jöst, 2006; Hong et al., 2009; Schmidt-Belz, Barbara et al., 2002; Kikiras et al., 2006).

The Figure 2.4 summarises the state of the art in personal mobility analysis for the purpose of constructing PG. In general, PG can be classified into three generations. The first generation of PG were simply an accumulation of spatio-temporal points. The second generation of PG are a list of the visited places of significance, as well as their relative times of visits. Finally, the third generation of PG included the visited places with their relative times of visits, as well as other – typically qualitative – data elements related to those places and visits. Predictive modelling of the data elements of the third generation PG is yet an open research area.

## 2.5 Other techniques

The work discussed in subsections 2.3 and 2.4 illustrated several examples of how to construct a PG from a spatio-temporal history of some individual, by different approaches. In those works, the emphasis is to employ the *personal* mobility history as the primary data source to extract the PG. In other words, the mobility traces of other people, whom may be spatio-temporally or socially related to the individual in question, are not considered in the PG construction process.

Similar mobility patterns of a group of different individuals are suggested as a measure of common interests between those people. In other words, people who share the same interests probably spend similar amounts of time at the same places (Terry et al., 2002). Another direction of work is to employ such patterns to personalise the information for a number of individuals, based on their spatio-temporal histories. For example,



**Figure 2.4: State of the art personal mobility analysis for the purpose of constructing personal gazetteers.**

see (de Spindler et al., 2006). Analysing the mobility history for a group of people, considering their social or spatio-temporal relatedness to each other, is considered out of scope for the purpose of this work.

The spatio-temporal history of a person has also been employed to learn and detect the activities taken by that person, in indoor environments. This direction of work is out of scope for the purpose of this work, and the reader is referred to (Nguyen et al., 2005; Niitsuma and Hashimoto, 2007; Mozos et al., 2007; Eagle and Pentland, 2006),

as examples of this line of work.

Studying PG is closely related to the problem of analysing travellers routes. The research work in this area is mainly directed toward the modelling and the prediction of the routes taken by a certain individual or a group of people, for example see (Patterson et al., 2003). Except for the purpose of constructing PG, this work does not consider any routes analysis.

Finally, several techniques were suggested to deal with the privacy problem in LBS. For example, in (Wishart et al., 2005) summarising the context of a mobile person, including location and activity, at different levels of granularity is suggested . Privacy issues related to the data collection, analysis, construction, and applications of PG are considered out of scope for the purpose of this work.

## 2.6 Limitations

The approaches discussed in section 2.4 allow to construct PG which included many qualitative data elements. Examples include the physical activities carried out by a certain individual, the comments people associate a certain place with, or the data typically requested by people at a certain place. However, the approaches given in the current literature toward the construction of such PG have their limitations, as given in the following.

### **The need for an explicit Web interface for *every* place**

If the semantics of a place are to come from the Web, as being either directly stated or indirectly mined from some Web resources, a Web interface needs to exist for that place. In other words, the place in question should have some Web interface, typically a Webpage, which includes enough data to describe that place. This is not always a realistic assumption. Many businesses have their Webpages already, but on the other

hand, others do not. Moreover, consider for example a Website of a typical business chain which consists of multiple branches, such as a Website of a restaurant chain. The Webpage which provides the data for a certain branch would typically include the address and the opening times of that branch, only. No specific data for that place is available, which describes, for example, the quality of the services provided at that certain place.

### **The need for explicit user interactions**

The work which relied on the general public to provide place semantics requires explicit users interactions for data entry. As mentioned in subsection 2.1.2, a personalised system is more efficient when the personalisation is more implicit to the user. In a similar manner, constructing a semantically rich PG implicitly, without explicit user interactions to provide its semantics, is more efficient.

### **The need to log more data other than the spatio-temporal mobility**

More than one of the works given earlier had to log additional data other than the spatio-temporal co-ordinates, such as the Web services or the data requested, to construct a semantically-rich PG. In a mobile deployment scenario, the size of the PG file and the computational power required to process it are concerns, due to the limited hardware and communication resources of personal mobile devices. As such, a light-weight solution is more desirable.

### **PG which depend on certain data elements are typically application-specific**

A PG which associates the spatio-temporal history of some individual with Web search experience, for example, utilises the resulting PG to improve any future searches. Similarly, a system which associates the mobility history with the data or Web services

requested is likely to utilise such a PG to improve its results. Typically, such systems do not share their profiles.

In a situation where there are multiple PG for the same individual, the non-spatio-temporal parts of such PG may differ, their mobility data are likely to be the same – for the same individual<sup>7</sup>. Such a data redundancy is not desirable, in particular in a mobile deployment scenario, for the same reasons mentioned earlier.

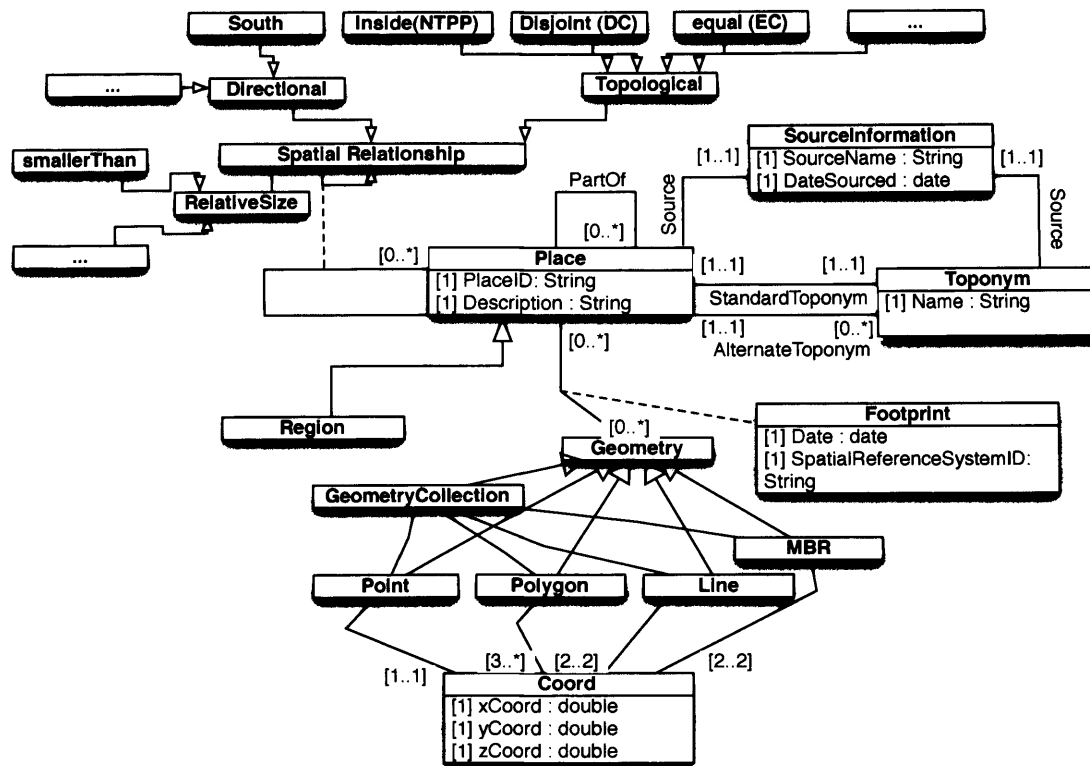
Another approach is needed to construct semantically-rich PG which overcome these limitations.

## 2.7 Extracting place affordance

The recent trend of geo-referencing resources on the web has driven the need for designing and building geographical place ontologies. A place ontology is a model of terminology and structure of geographic space and named place entities (Egenhofer, 2002). It extends the traditional notion of a gazetteer to encode semantically rich spatial and non-spatial entities, such as the historical and vernacular place names and the events associated with a geographic place (Perry et al., 2006), as well as the qualitative spatial relationships between instances of geographical places. An ontology of a geographic *place* is seen to play a key role in facilitating GIR on the web (Gregory et al., 2002; Jones et al., 2002; Smart et al., 2010). An example geographic place ontology is shown in Figure 2.5. In this example, the ontology is employed to support the representation and manipulation of geo-referenced resources on the Semantic Web (Abdelmoty et al., 2009). In this context, a geographical place is primarily distinguished by belonging to a category of place types, e.g. roads and streets, rivers and streams, etc., as well as by the identification of its name and physical structure; its location and shape.

---

<sup>7</sup>Some variations may result depending on the algorithm used to construct the PG in each application.



**Figure 2.5:** An example of a geographical place ontology employed in GIR on the Semantic Web (Abdelmoty et al., 2009).

Functional differentiation of geographical places, in terms of the possible human activities that may be performed in a place or *place affordance*, has been identified as a fundamental dimension for the characterisation of geographical places (Relph, 1976). For Relph, the unique quality of a geographical place is its ability to order and to focus human intentions, experiences, and actions spatially. It has been argued that place affordance is a core constituent of a geographical place definition, and thus ontologies for the geographical domain should be designed with a focus on the human activities that take place in the geographic space (Kuhn, 2001; Frank, 2003). The term “action-driven ontologies” was first coined by (Câmara et al., 2000) in categorising objects in geospatial ontologies. Affordance of geospatial entities refers to those properties of an entity that determine certain human activities. In the context of spatial information



theory, several works have attempted to study and formalise the notion of affordance (Rugg et al., 1997; Kuhn, 2007; Sen, 2008, 2007; Raubal and Kuhn, 2004; Scheider and Kuhn, 2010). The assumption is that affordance-oriented place ontologies are needed to support the increasingly more complex applications requiring semantically richer conceptualisation of the environment. Realising the value of the notion of affordance for building richer models of geographic information, the Ordnance Survey (the national mapping agency for the UK) proposed its utilisation as one of the ontological relations for representing their geographic information (Hart et al., 2004) and made an explicit use of a “has-purpose” relationship in building their ontology of buildings and places<sup>1</sup>.

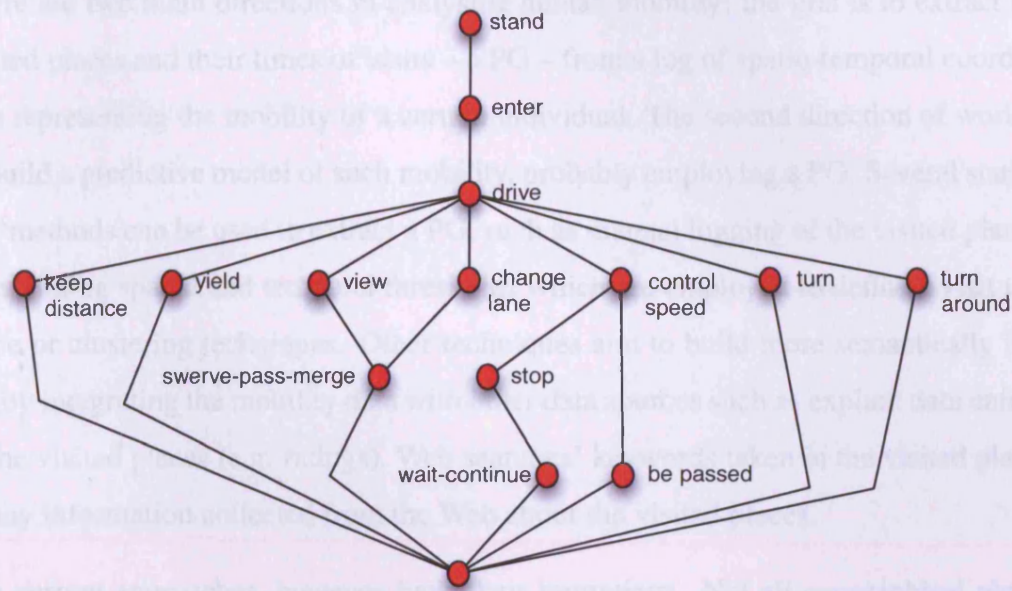
There exist several approaches in the literature which address the issue of extracting place semantics, ranging from understanding natural language descriptions, experimentation with user-based population of place semantics and domain experts evaluations, to a hybrid of those techniques. Some of those methods address ontological issues, e.g. the identification and classification of geographical place categories (Tversky and Hemenway, 1983; Smith and Mark, 2001; Edwards and Templeton, 2005), while others focus on populating the ontologies, such as the web mining techniques employed in the domain of GIR (Popescu et al., 2008).

Hierarchies of human actions in a geographical domain were extracted by analysing formal textual resources, such as the German traffic code documents in (Kuhn, 2001, 2000; Sen, 2007). (Kuhn, 2001) suggests a stepwise methodology to develop ontologies in support of human activities. The aim is to extract human action hierarchies, such as shown in Figure 2.6. (Sen, 2008) builds on this work and employs a word frequency co-occurrence model on two traffic code texts to extract nested and sequential affordances.

In the above approaches, hierarchies of action entities have been considered independently of the geographic entities they are associated with, and a link from the derived

---

<sup>1</sup><http://www.ordnancesurvey.co.uk/oswebsite/ontology>.



**Figure 2.6: Action hierarchy derived from traffic code texts in (Kuhn, 2001) (reproduced).**

action hierarchies needs to be made to the relevant entity classes that afford those actions. Specific text documents relevant to a particular domain were used in (Kuhn, 2001), where the automation of text analysis is noted to be problematic and manual intervention is needed. (Sen, 2007) investigates a simple case study of linguistic analysis of noun and verb phrases in formal texts of traffic codes and proposes a probabilistic approach to linking two parallel taxonomic hierarchies of entities and actions.

## 2.8 Summary

Personalisation is a widely applied technology to improve information retrieval on the Web, and in context-aware and mobile systems. While there are many configurations, architectures, techniques and methods to apply achieve personalisation, a common element which is employed in mobile systems personalisation is the user's location.

There are two main directions in analysing human mobility; the first is to extract the visited places and their times of visits – a PG – from a log of spatio-temporal coordinates representing the mobility of a certain individual. The second direction of work is to build a predictive model of such mobility, probably employing a PG. Several statistical methods can be used to extract a PG, such as manual logging of the visited places, determining spatial and temporal thresholds which are employed to define a visit to a place or clustering techniques. Other techniques aim to build more semantically rich PG by integrating the mobility data with other data sources such as explicit data entries of the visited places (e.g. ratings), Web searches' keywords taken in the visited places or any information collected from the Web about the visited places.

The current approaches, however, have their limitations. Not all geographical places have associated information on the Web, or if place semantics are to be provided by the users then this requires additional user interactions. Moreover, this comes with an overhead as it requires to log historical interactions data. A more generic semantic descriptions of the visited places is desirable, as it can be associated with mobility logs to extract semantically rich PG, which are generic enough to be employed by different applications.

The next chapter gives the work done in this thesis to extract PG from GPS logs, while the next chapter introduces the work done to extract place semantics based on the notion of place affordance.

# **Improving the quality of PG extraction by pre-processing**

## **3.1 Introduction**

The spatio-temporal data streams provided by GPS-empowered personal mobile devices often suffer from noisy, erroneous, or missing data. These problems can have impact on the quality of applications which employ such data, such as building a PG. In this chapter, several pragmatic heuristics are introduced and employed to design and build simple procedures for pre-processing GPS data streams, to improve their quality. The proposed heuristics and procedures are shown to provide good improvements within the context of automatically constructing PG.

## **3.2 Personal Gazetteers from GPS data streams**

A distinctive characteristic of an increasing number of personal mobile devices is their spatial awareness. At present, one of the main technologies used to identify locations for outdoor environments is GPS. In association with such devices, a wide variety of location-based applications are emerging. For example, many services on the Web today offer to store, tag, visualize, or optionally share people mobility data collected by such devices. These data sets may be points of spatial significance, such as locations

of homes, work places and restaurants, as well as spatial routes such as the trips of the people who provided their data. Examples of such Web services are described (FireEagle, 2010).

As mentioned, GPS is the main technology currently employed to construct PG. Manually constructing a PG has its obvious limitations; a person has to manually enter all the data of the places and their relative times of visits, as well as to remember all that data between the data entries sessions, including the less memorable ones (such as being at bus stops or stopping at a news agent). In contrast, mining personal mobility data streams for the purpose of automatically constructing a PG neither needs data to be entered manually, nor to remember any of it. In automatically constructed PG, the loss of a GPS signal is commonly employed as a heuristic to identify places in urban areas, as a GPS signal is lost inside buildings. This heuristic, however, leads to poor quality identifications of places, as a GPS signal may also be lost for various other reasons; such as having insufficient visible GPS satellites in dense urban areas (Marmasse and Schmandt, 2000; Moreira and Santos, 2005). This problem, along with other issues such as faulty and noisy GPS data readings, has led to a significant drop of quality of automatically built PG – as will be shown in the next section. Several approaches addressing the problem were suggested in the literature, as follows.

Additional hardware may be used to improve the quality of GPS data streams, such as the case of Assisted Global Positioning System (A-GPS) (Djuknic and Richton, 2001). The basic principle of how A-GPS works is to utilise the wireless network available to a device, e.g. a mobile phone, to obtain additional information from a location server, to improve the accuracy of the GPS receiver on the device and to reduce the time required to obtain a satellite fix (Feng and Law, 2002). In (Lea et al., 2006), A-GPS is employed to improve the quality of automatically built PG. A-GPS, however, may not be available to all mobile devices or communication networks. On the other hand, additional data sets, such as a transportation network data set, can also be employed to reconstruct a poor quality GPS log. The idea is simple: as the GPS signal of a moving

object (e.g. a mobile phone, and obviously in the case of PG excluding a device on an aircraft, for example) can only be received in open spaces (i.e. on the road network), and as the road network coordinates are known, any spatio-temporal point within a GPS track that is outside the scope of the road network coordinates is considered as noise. Such data sets, however, are typically of a commercial or governmental nature and are not available for public use<sup>1</sup>. An alternative is to develop heuristics and algorithms for the purpose of improving the quality of GPS data streams, which is the direction taken in this work. A number of techniques have been proposed to automatically construct PG with minimal errors, using only raw GPS data (see Chapter 2), where neither additional hardware nor additional data sets are required. The majority of these solutions, however, are relatively complex, and some require additional parameters to be set in advance prior to any processing (e.g. setting the cluster size in clustering algorithms). Moreover, there are some other types of GPS log-based applications which do not require any analysis of their raw GPS data streams. For example, some applications aim to simply visualise the GPS logs for the purposes of tagging or sharing on the Web. In other words, there is no need to cluster a GPS log to simply, and quickly, visualise it without errors.

The work in this chapter is proposed to improve the quality of automatically constructed PG from GPS data streams by employing a set of relatively simple heuristics, as a preprocessing step. The proposed procedures showed effective improvements of the quality of the collected GPS logs, for the purpose of automatically constructing PG. Hence, the proposed procedures could improve the quality of any applications which could be built employing PG. Common problems found in GPS data streams are first explained, followed by the proposed pre-processing procedures, and then their application on sample real-world data sets is carried out and the results are provided.

---

<sup>1</sup>In recent years, there have been some attempts to collect such data sets in a collaborative manner by the general public, for example (openstreetmap, 2010). The quality of the data sets collected via this approach, however, is not guaranteed as it is collected by non technicians.

**Table 3.1: Possible problems of GPS data streams and their causes, data, and algorithmic effects**

Problem	Possible causes	Possible data effects	Possible algorithmic effects
loss of GPS signal	insufficient number of satellites to a GPS receiver	segmented logs	false identifications of places and trips
erroneous GPS readings	hardware	spatio-temporal points which are out of harmony with the rest of their track	false identifications of places and trips and false calculations
GPS signal availability in open-spaces	normal GPS operation	spatio-temporal points of visits to open-space places exist in a log	false identifications of visits to open-space places
inaccurate GPS readings	accuracy of GPS receivers	different GPS readings of the exact same space and time co-ordinates	inaccurate identifications of places and trips
other problematic GPS readings	hardware	a single spatio-temporal point in a track or only spatial co-ordinates in a spatio-temporal point	false identifications of places and trips

### 3.3 Problems of personal GPS data streams for building PG

Typically, a GPS data file consists of a set of one or more *tracks*. Each track, in turn, consists of one or more spatio-temporal *points*; a pair of latitude and longitude coordinates along with their relative time-stamp. Other information may also be included (GPX, 2009) (such as satellite number). During the course of work in this chapter, 2 people contributed their GPS logs for analysis. Those logs were collected using the GARMIN GPS60<sup>2</sup> handsets, over a four weeks period. Close examination of these data sets revealed several problems. These are discussed below. Table 3.1 summarizes the possible problems of GPS data streams and their causes, data, and algorithmic effects, found in the course of this work for the purpose of automatically constructing PG.

#### 3.3.1 The loss of GPS signal

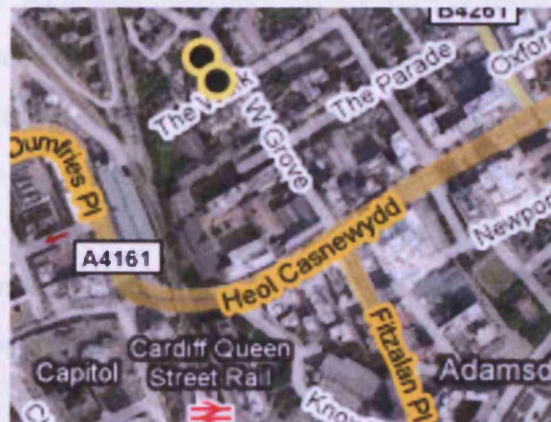


Figure 3.1: A segmented track due to the loss of a GPS signal

GPS signal is lost in heavily built-up urban areas, known as urban canyons. Typical examples of urban canyons can be found in a down-town area of a modern city, where

<sup>2</sup><http://www.garmin.com>.



buildings of five stories or higher are found on both sides of a road. In such environments, a GPS signal is lost due to an insufficient number of satellites which are on a direct line of sight with a GPS receiver. Another scenario is the loss of a GPS signal while in a tunnel or an indoor pathway, for example inside a covered shopping mall. As the loss of a GPS signal is commonly used as the heuristic employed to infer a visit to a place, this problem can significantly reduce the quality of personal gazetteers as it leads to false identifications of places in urban areas. Because of a GPS signal being lost, the normally continuous spatio-temporal points in a single track are often segmented. In other words, a single trip is misleadingly represented as several trips, and possibly causing misinterpretation of visits to places. See Figure 3.1.

### 3.3.2 Erroneous GPS readings

Some GPS readings were obviously out of harmony with the rest of the other spatio-temporal points in the same track, which are referred to as *overshoot points* in the context of this work. Although overshoot points did not directly affect place identifications, such data lead to many other problems such as the incorrect visualization of trips. An exception is when an overshoot point is at either end of a track, or both, where it leads to false place identifications. Moreover, any calculations which are to be performed on a track which contains overshoot points will probably be inaccurate, as the spatial and temporal coordinates of an overshoot point are false. Figure 3.2 shows a track with overshoot points.

### 3.3.3 Open-spaces detection

As a GPS signal is normally receivable only in outdoor environments (parks, open-space areas of restaurants, and so on), an additional effort is required to identify visits to open spaces. This is based on the assumption that the loss of a GPS signal is the heuristic employed to identify a visit to a place.

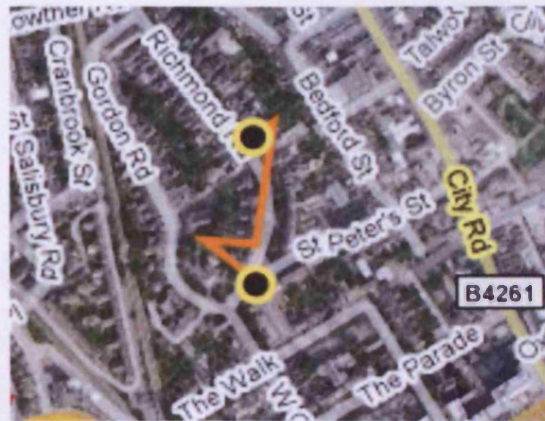


Figure 3.2: A track with overshoot points

### 3.3.4 Inaccurate GPS readings

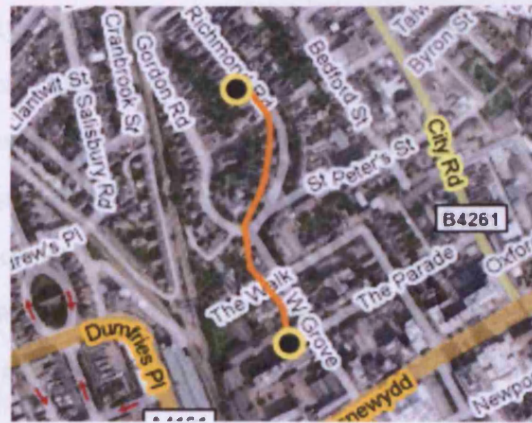
In the course of this work, it was found that different commercial brands of GPS logging devices provided different GPS readings, for the exact same location and time. The devices tested were GARMIN GPS60, GPSMAP 60CSx, and Sony GPS-CS<sup>3</sup>. Moreover, three GPS devices of the same commercial brand provided slightly different GPS readings for the exact same location and time. Furthermore, the same device provided different GPS readings for the exact same location, at different times.

Figure 3.3 shows the effects of accuracy of GPS reading. While the track is generated from a handset carried by a walking person, there is a noticeable shift from the road network.

### 3.3.5 Other issues with GPS readings

Although of a rare occurrence, other problems appeared with GPS readings that needed to be addressed. For example, some tracks had only a single spatio-temporal point. Another example is a track which had only spatial coordinates for its points, with no time-stamp.

<sup>3</sup><http://www.sony.com>.



**Figure 3.3: A track with inaccurate GPS readings**

Some of the problems which were introduced previously may be rooted in various hardware causes. For the purpose of this work, a pragmatic view is taken for identifying problems with GPS data streams, as well as for their proposed solutions. The objective is to provide fairly practical and easy to implement solutions for some of these problems – or more precisely, their effects.

### 3.4 Improvements procedures

The procedures proposed are based on several heuristics reflecting common sense knowledge of human mobility, namely:

1. *Merging of segmented tracks.* A pair of consecutive tracks are merged based on two constraints. Firstly, the temporal difference between the last point of the first track and the first point of the second track has to be less than a certain time threshold. Secondly, the velocity of the virtual track which is constructed temporarily beginning with the last point of the first track, and ending with the first point of the second track, has to be less than or equal to the mean velocities of first and the second tracks. The tracks are merged if the two conditions are

satisfied. Simply put, a mobile person is unlikely to accelerate, on a certain part of his track, more than the mean velocity of that track. As such, if some parts of that track are lost and the track is segmented, then the segments are merged back – if the constraints mentioned previously are satisfied. The only exception is the case when a mobile person decides to accelerate to a velocity which is higher than the track mean velocity, while the GPS signal is lost. For the purpose of this work, this exception is considered of a rare occurrence, and as such is not studied further.

2. *Elimination of erroneous GPS readings.* For the purpose of this work, an overshoot point is any spatio-temporal point in a track which is at more than a certain distance from its previous and next points, in that track. That distance is calculated as the mean of distances between each pair of consecutive points in that track, multiplied by a certain factor. Such points are eliminated.
3. *Elimination of other problematic GPS readings.* A track with a very low number of spatio-temporal points was found to be uninformative, as it could neither be used for visualization purposes, nor to identify the visited places correctly. Such tracks are eliminated.

### 3.4.1 Overshoot points filter

Equation 3.1 defines the set of overshoot points as explained below:

$$O := \{p \mid d(p, p') \geq D \text{ and } d(p, p'') \geq D\} \quad (3.1)$$

Where  $O$  is the set of overshoot points,  $p$  is an overshoot point,  $p'$  and  $p''$  are the previous and the next spatio-temporal point to  $p$  respectively,  $d$  is the distance between any two given spatio-temporal points, and finally  $D$  is a certain distance. The purpose

**Algorithm 3.1: Overshoot points filter****Input:** A set of tracks  $T$  and an overshoot threshold  $oT$ **Output:** Overshoot points are eliminated from  $T$ .

- 1: **while** there exists a  $track \in T$  **do**
- 2:    $D = \text{mean distance}(track) \times oT$
- 3:   **while** there exists a  $point \in track$  **do**
- 4:     **if**  $\text{distance}(point, \text{previous point}) \geq D$  **and**  $\text{distance}(point, \text{next point}) \geq D$  **then**
- 5:       **eliminate}(point)**

of this procedure is to eliminate such points. The procedure is illustrated in Algorithm 3.1. In practice, an *overshoot threshold* ( $oT$ ) of 3 provided satisfactory results.

For each track, Algorithm 3.1 (step 1) defines a spatial distance for that track (step 2) as indicated above. For each point in the track (step 3), the constraint between that point and its previous and next point is applied (both spatial differences should be less than the spatial distance defined in step 2), and the point is eliminated if it satisfies the constraint of being an overshoot point (step 5).

**3.4.2 Segmented tracks merger**

Equation 3.2 defines the set of segmented tracks as explained below:

$$S := \{r \mid v(r, r'), \bar{V}(v) \leq \bar{V}(r), \bar{V}(v) \leq \bar{V}(r'), \text{and } t(r, r') \leq tT\} \quad (3.2)$$

Where  $S$  is the set of segmented tracks,  $r$  is a segmented track,  $r'$  is the track next to a segmented track  $r$ ,  $v$  is a virtual track which is temporarily constructed between any two given tracks,  $\bar{V}$  is the mean velocity of a given track,  $t$  is the time difference between any two given tracks, and finally  $tT$  is a temporal threshold. This procedure is proposed to merge segmented tracks based on the first heuristic explained

**Algorithm 3.2: Segmented tracks merger****Input:** A set of tracks  $T$  and a time threshold  $tT$ **Output:** Segmented tracks in  $T$  are merged.

- 1: **while** there exists a  $track \in T$  **do**
- 2:    $v = \text{construct virtual track}(track, track')$
- 3:    $t = \text{time difference}(track, track')$
- 4:   **if**  $\text{mean velocity}(v) \leq \text{mean velocity}(track)$  **and**  $\text{mean velocity}(v) \leq \text{mean velocity}(track')$  **and**  $t \leq tT$  **then**
- 5:     **concatenate}(track, track')**

at the beginning of this section. See Algorithm 3.2. In the current implementation,  $\text{timethreshold}(tT)$  is set to 30 minutes (more on the time threshold is given in the next section).

For each track, Algorithm 3.2 (step 1) defines its virtual track (step 2) and time difference between this track and its next track (step 3) as explained above. Each track is then checked if it is a segmented track (step 4), and merged with the next track accordingly (step 5). The constraint for merging is to check the mean velocity of the virtual track is less than the mean velocity of the track as well as the next track, and the time difference is less than the time threshold given (see above).

An earlier version of this procedure was tested, in which only a spatial and a temporal threshold were employed to merge segmented tracks. A spatial threshold of 100 meters and a temporal threshold of 60 seconds were used. Any pair of consecutive tracks in which the spatio-temporal distance between the last point of the first track and the first point of the second track fell within the specified thresholds, were merged. While that provided positive results for tracks with relatively short periods of GPS signal loss, it failed for the longer ones. As such, the current version was developed.

**Algorithm 3.3: Noise filter****Input:** A set of tracks  $T$  and a noise threshold  $nT$ **Output:** Noisy tracks are eliminated from  $T$ .

- 1: **while** there exists a  $track \in T$  **do**
- 2:   **if**  $points\ count(track) \leq nT$  **then**
- 3:     **eliminate**( $track$ )

**3.4.3 Noise filter**

In practice, some GPS logs suffered from having uninformative tracks; tracks with only one or two spatiotemporal points. This is in particular true for the very first track in such logs – even after applying Algorithms 3.1 and 3.2 to those logs. In this work, such tracks are referred to as *noisy tracks*, and are defined as (see Equation 3.3):

$$N := \{t \mid points\ count(t) \leq nT\} \quad (3.3)$$

Where  $N$  is the set of noisy tracks,  $t$  is a noisy track,  $pointscount(t)$  is the number of spatio-temporal points in that track, and  $nT$  is the noise threshold. As being uninformative, such tracks are eliminated, see Algorithm 3.3. In the current implementation,  $nT$  is set to 1. In other words, any track with less than 2 spatio-temporal points is eliminated from the log.

**3.5 Results**

The basic step in constructing a PG is the detection of candidate places in a GPS log. As mentioned earlier, the common technique is to consider the loss of a GPS signal, or the last point in a track, as the visited place. This often leads into falsely identified places, as the GPS logs typically contain erroneous, noisy, or corrupted data. Moreover,

while this is a significant indicator of typical places, it misses out other places; open-space places where the GPS signal is still detectable, such as road-side cafes, parks, and similar places<sup>4</sup>.

To apply the standard approach, simply the spatial coordinates of the last spatio-temporal point in a track are considered as the spatial coordinates of the visited place. As mentioned earlier, as GPS logs typically suffer from segmented tracks, this approach is very susceptible to incorrectly identifying places – due to the intermittent loss of the GPS signal. As such this approach stands to benefit from the proposed pre-processing procedures.

For detecting open-space places, different constraints were employed. Namely, the minimum time ( $tMin$ ) and the maximum time ( $tMax$ ) for which a person is moving within an area, which is in turn defined by a spatial buffer ( $sMax$ ). Different values of these heuristics can be employed to detect different types of open-space places, within the same track<sup>5</sup>. For each track, all its points are checked for point sets which obey the spatial constraint, and the resulting sets are then checked for the temporal constraints. The points which do not satisfy all the constraints are eliminated, and if some sets overlap, then the set with earliest time-stamp is considered. By employing different thresholds, different place types are detected; namely short-stay and long-stay candidate places. Examples of the former type are places where a person stays for relatively short times and is moving within a short spatial distance, such as short lunch-brakes in parks, stopping at petrol stations, or having a morning coffee from a street vendor. Examples of the latter type are typically places of leisure times, such as trips to a zoo or outdoor sports areas. For short-stay candidate places the thresholds are set as the following:  $sMax$  is 25 metres,  $tMin$  is 5 minutes, and  $tMax$  is 30 minutes, while

---

<sup>4</sup>Although this is not the main focus of this work, different spatio-temporal constraints were employed to detect different types of visits, as will be explained.

<sup>5</sup>The main focus of this work is to extract the visits to closed places, as those places are of interest in terms of extracting semantically rich profiles, and as such visits to open-spaces were not presented as procedures.



the long-stay thresholds are:  $sMax$  is 250 metres,  $tMin$  is 30 minutes, and  $tMax$  is unlimited.

Typically, a person visits many places, often more than once. As such, a number of candidate places which represent the same place, may be detected within a single GPS log, and more likely in multiple GPS logs of the same individual. For example, a typical GPS log of one week period would include several candidate *home* and *work* places, and perhaps few other candidate places, while a typical GPS log of a one month period would also include several candidate *shop* and *restaurant* places, and perhaps *relatives* or *friends houses*. Moreover, multiple GPS logs of the same individual would include several candidate *home* places – all representing the same place in reality, i.e. *home*, among other possibly duplicated candidate places. This would clutter the PG, and would add complexities to the analysis of its spatio-temporal patterns. This problem may be solved by clustering the candidate places into final places for inclusion in the constructed PG. In this work, a modified version of the DBSCAN algorithm is used for clustering candidate places. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm defines a cluster by a minimum number of points inside it, with each pair of points are mutually density-connected<sup>6</sup> (Birant and Kut, 2007). Unlike the K-means algorithm, the DBSCAN algorithm does not need any initialisation, i.e. the number of places to be determined in advance, and it can define clusters with different sizes<sup>7</sup>. Compared to the original DBSCAN algorithm, this work employs a variable distance spatial buffer. The spatial buffer is determined based on the type of the candidate place, being a short-stay or a long-stay candidate place. The modified DBSCAN retains the minimum number of points for a cluster criterion, which can be tuned to define how permissive the clustering is. Allowing places with only one point in a cluster increases the number of incorrect places, while filtering out such places creates a personal gazetteer with fewer incorrect places but misses out

---

<sup>6</sup>Either the two points are within a certain spatial distance or there exists one or more points connecting the two points, where each pair of points in the sequence are within the spatial distance.

<sup>7</sup>For more details on DBSCAN for PG, see pages 29–30.

some infrequent places. It is a trade-off that needs to be decided upon based on the target application of the resulting personal gazetteer.

The next subsections give the results of two experiments of evaluating the proposed procedures introduced previously. In each experiment, a different data set is used. While the two data sets were collected over the same period, the number of visited places in each differs, as well as the settings of the mobility area, as explained. The types of the visited places in the experiments are given in Table 3.2.

**Table 3.2: Visited places types of collected GPS logs in Experiments 1 and 2**

Person	Visited place types
Experiment 1	university, theatre, supermarket, station, sport centre, spa, ski resort, shop, school, restaurant, parking place, library, house, home, community centre, club, church, cafe, bar, bank, airport.
Experiment 2	university, shop, restaurant, library, home, bank.

### 3.5.1 Experiment 1

In this experiment, a semi-automated evaluation approach of the results is followed. A four weeks data collection period was logged using a personal mobile GPS logger, along with a paper log. The PG resulting from applying the pre-processing procedures and PG construction approach mentioned earlier, are evaluated based on the paper log.

#### Data

The data used in this experiment is collected in a typical urban area, with one excursion into a rural setting. The mobile device used to acquire the data was a Garmin GPS 60,

with a high collection frequency set. The person was instructed to acquire a satellite fix after leaving a place, which, while unrealistic, does improve the consistency of the data.

Since the actual locations of the visited places are known (from the paper log), a polygon of a place can be generated using any of the free applications on the Web such as Google Earth<sup>8</sup>, for example. For each place visited, a polygon was manually created representing the shape of the place, with an additional spatial buffer of between 10 and 20 meters (depending on how near the next possible place is). A point-in-polygon test is employed to determine whether a detected place in the resulting PG – constructed by applying the procedures described earlier – actually identifies a visited place or not.

## Results

The person identified 36 places in the collected data. Table 3.3 shows the results of places detection, with and without applying the preprocessing procedures. *True positives* is used to refer to the detected places in the constructed PG which lie inside their respective generated polygons, while the *false positives* are those places in the PG which lie outside their respective polygons. *Precision* and *recall* have the standard information retrieval interpretations (precision is the number of the correctly identified places out of all the identified places, while recall is the number of the correctly identified places out of the total number of places in the collected data), and the *F-score* is the harmonic mean of the precision and the recall.

A first look at the results shown in Table 3.3 reveals that the results show an acceptable recall, but a low precision. Clearly, the preprocessing procedures do not change the number of correctly identified places, while at the same time nearly halving the number of false positives. A further analysis of the results is given in section 3.6, in order to illustrate the nature of the improvements introduced by the preprocessing techniques.

---

<sup>8</sup>[http://www.google.co.uk/intl/en\\_uk/earth/index.html](http://www.google.co.uk/intl/en_uk/earth/index.html)

**Table 3.3: Results of constructing a PG without and with preprocessing**

Result	Detection strategy	
	No preprocessing	Preprocessed
True positive	25	25
False positive	41	23
False negative	11	11
Precision	0.38	0.52
Recall	0.69	0.69
F-score	0.49	0.6

### 3.5.2 Experiment 2

In this experiment a different data set is evaluated for a further testing of the proposed procedures. The same semi-automated evaluation approach is followed.

#### Data

The data set in this experiment is collected roughly within the same urban area of the previous experiment, with no excursion into any rural areas. The same mobile device and data collection procedures were used.

The data was collected over a period of four weeks. Compared to the data set used in the previous experiment, only five distinct places were visited in this data set. Moreover, all the places are in a relatively small area; which lie within a circle of 1.2 kilometers diameter, where no two visited places were spatially adjacent or close, compared to the first experiment, see Fig 3.4.



**Figure 3.4:** Places which were visited in the second experiment – all of which are within a relatively small urban area; which lie within a circle of 1.2 kilometers diameter.

### Results

Table 3.4 shows the results of this experiment with and without applying the proposed procedures. In each case, the results are shown when the candidate places were filtered, as well as when they were not. See Table 3.4.

**Table 3.4: Results with and without preprocessing and with and without filtering the candidate places.**

Result	Detection strategy			
	No preprocessing		Preprocessed	
	No filter	Filtered	No filter	Filtered
True positive	2	2	2	2
False positive	18	4	6	1
False negative	3	3	3	3
Precision	0.1	0.33	0.25	0.66
Recall	0.4	0.4	0.4	0.4
F-score	0.16	0.36	0.30	0.5

## 3.6 Evaluation

### 3.6.1 Analysis of experiment 1

#### Analysis of the false negatives

The false negatives are places visited by the person, while logging his mobility as instructed, which were not part of the resulting PG – i.e. *undetected* after preprocessing and clustering his log. Examining the false negatives reveals that 10 of these places lie in the city centre, part of which is shown in Figure 3.5. This is to be expected, because although the person was instructed to acquire a satellite lock with the GPS receiver immediately after leaving any building, the signal is easily lost before the actual destination is reached (in particular, in relatively narrow roads, and depending on the satellite constellation).

The places which do not appear correctly in the constructed PG (the false negatives) can roughly be classified into two categories. There are those places that do appear in the PG, but where the place location is incorrect (i.e. inexact, Places 7, 45 and 54 on the

map in Figure 3.5) due to errors in the original GPS location calculation. The second set of undetected places are those where the original track ends before the destination place is reached (and as such they result in totally different places compared to the actually visited places, Places 19, 23, 26 on the map in Figure 3.5). The problem is basically caused, in both cases, by being mobile in heavily built-up areas. In such urban canyons, only a few satellites are visible to the mobile device, which reduces the accuracy of the calculated locations, which in turn leads to errors of the first category. Moreover, walking around buildings' corners often leads to losing the direct line of sight to one of the satellites visible to the device, causing the premature end of GPS tracks, which leads to errors of the second category. Another possibility is that when visiting two places which are spatially close to each other, with a distance less than the clustering threshold.



**Figure 3.5: Detection of errors in urban canyons. The GPS signal was lost before reaching the destination places: Place 19, Place 23, Place 26. The places which were not correctly identified due to errors in the GPS readings are: Place 7, Place 45, Place 54.**

To summarize, false negatives are caused either by erroneous GPS readings, losing the

GPS signal before reaching the destination place, or visiting two places one of which is frequently visited, while the other is less visited and spatially adjacent or close to the first place – in which case the latter place will not appear in the personal gazetteer (as both places will be detected as a one place).

### **Analysis of the false positives**

False positives are the places in the resulting PG, after applying the preprocessing procedures and the clustering algorithm, which were incorrectly identified as they do not match the actually visited places (compared to the paper log). For the false positive places, around half the false positives have corresponding false negatives – for the reasons which were analyzed in the previous subsection. These are places that are basically identified as candidate places in the resulting PG, but due to the errors in the accuracy or the completeness of the original GPS data, the locations of the candidate places were incorrect.

### **Analysis of the clustering threshold**

Analysis of the false positives also revealed that most of them had only one candidate location (spatial point) associated with, so an additional filtering step was performed. After clustering places, all the clusters with only one candidate place were filtered from the set of places, the results of which can be seen in Table 3.5.

One problem that this extra filtering step exposes is that there are some cases where the preprocessing heuristics tend to merge tracks which should not be merged. This leads to the slightly lower number of false positives when applying the preprocessing procedures, as it removes one of the two candidate places that the algorithm detected in the original data, and then the cluster is filtered. Nevertheless, when compared with Table 3.3, the results clearly show that filtering place clusters which have only one candidate place drastically lowers the false positives rate, with precision values in the



**Table 3.5: Evaluation results with clusters of one candidate place filtered**

Result	Detection strategy	
	No preprocessing	Preprocessed
True positive	17	14
False positive	3	2
False negative	19	22
Precision	0.85	0.88
Recall	0.47	0.39
F-score	0.61	0.54

high eighties, while the number of correctly identified places drops by roughly a third. The non-preprocessed detection strategy with the additional filtering actually has the highest F-score of all combinations, although this does come at the price of detecting less than half the places the user visited.

### 3.6.2 Analysis of experiment 2

For this data set, the results of this experiment show that the recall value was neither affected when applying the proposed procedures, nor when filtering the candidate places. The same applies to the missed places. The precision value, on the other hand, was significantly improved after applying the preprocessing procedures, as well as when filtering the candidate places. As in the previous experiment, the threshold of filtering the candidate places is set to one.

As mentioned, the data collected in this experiment represents a walking person within a heavily built-up, and a relatively small, urban area. For the same reasons explained in the previous subsection, a large number of false positives was found, compared to the

paper log. Similarly, by simply preprocessing the log, the number of incorrectly identified places is significantly reduced. As for the clustering, the filtering of candidate places with only a single point in a cluster reduced the false positives to nearly one fifth. The results of this experiment further supported the argument of pre-processing. Applying a set of simple preprocessing procedures has significantly improved the results of processing the GPS log of a person, for the purpose of automatically constructing a PG.

### 3.6.3 Discussion

While the number of visited places is different in the two experiments, the results showed some similarity. The proposed procedures showed effective reduction of false positives, without reducing the true positives. On the other hand, the analysis showed that many of the false positives are actually related to some of the false negatives. In other words, these are places which were detected in the PG, but were identified as false places compared to the actually visited places – which, in turn, did not appear in the PG, for the reasons explained in the previous subsection. The following is an elaboration of some suggestions to reduce the false negatives, and hence the false positives, i.e. what are the possible ways to increase the recall.

As mentioned previously, the false negatives are either caused by erroneous GPS readings, losing the GPS signal before reaching the destination places due to insufficient number of visible satellites, or visiting spatially adjacent or close places. For frequently visited places, the GPS track end point, which is employed to identify candidate places, will usually be correct, and as such could be used to correct the false tracks to that place. Thus it might be possible to see if these tracks match with previous routes taken by the person. If they do, then assigning them to the end-point of the route they are most spatially similar to, would correct the errors in the original GPS data. The question is how to ensure that this does not lead to less frequent places that lie on very frequent routes, as they are amalgamated into the more frequent route. An

adequate similarity measure is needed, which may also employ the times of visits in a way similar to (Chang et al., 2007).

An alternative is to employ a combination of intelligent heuristics, additional data sets, and additional hardware to improve the quality of GPS personal mobility logs, if applicable. In such a scenario, additional data sets in particular may be useful to correct visits to spatially adjacent or close places, on the assumption of having data sets representing the exact polygons of the visited places. For example, a recent work of (Schmid et al., 2009) reduced a mobility track into its basic elements, typically stops and changes of course of mobility (e.g. at streets intersections). In this way a mobility track is significantly compressed in terms of number of spatio-temporal points, allowing efficient indexing and processing. This, however, requires that at least all the basic elements of mobility identified earlier – in particular destination places, are logged correctly. The work presented in this thesis is a contribution toward that. Other solutions could possibly utilize the heuristics introduced in this work to produce a better quality GPS-based mobility logs.

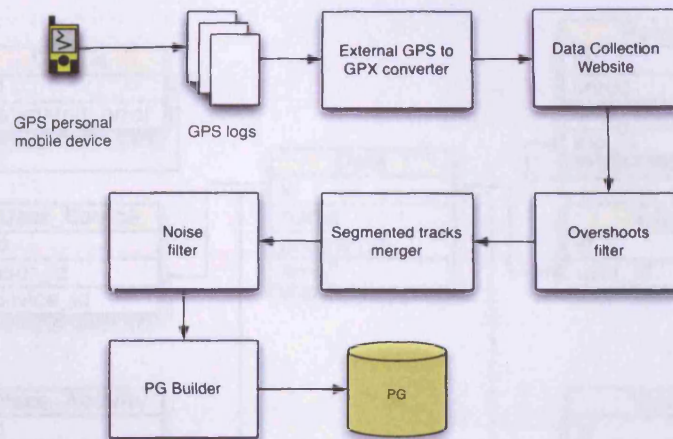
## 3.7 Application

The proposed procedures were implemented in a data collection application, with a Web interface for users to upload, visualise, and annotate their GPS personal mobility logs.

The users could use any personal GPS logging device to collect their logs. The users were asked though to use an external GPS to GPX file format converter, to unify their file types<sup>9</sup>. Once the users upload their GPX files, the application applies the Overshoots, Segmented tacks, and the Noise filters to the uploaded data. Once processed, the users are then able to visualise the resulting PG and tag it with related visit information. The Web interface is written in PHP, and the back-end database is implemented in

---

<sup>9</sup><http://www.gpsbabel.org>



**Figure 3.6: Place Activity Time data collection application architecture**

Oracle. The application architecture is given in Figure 3.6. OpenLayers<sup>10</sup> open-source maps visualization library, and the Google Maps API<sup>11</sup>, were also used for map-based visualisation of the GPS logs.

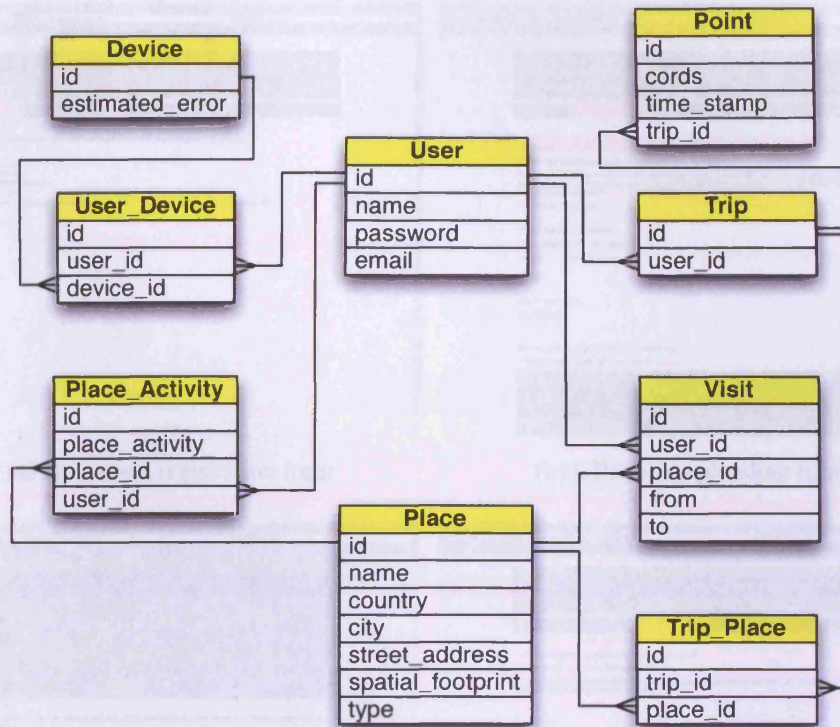
The database keeps the basic user data, e.g. name, email, password, and the used GPS device type. This is to allow for a flexible processing of the data based on the device specific error rate. Each user may upload one or more trips, where each is represented by its spatio-temporal points. A trip is analysed to extract ideally its two distinctive visits to its source and destination places. The visits extracted from each trip – and hence their relative places – are visualised on a map for the user to annotate with one or more activities. The database schema is shown in Figure 3.7.

New users can register, or log in into the system as shown in Figure 3.8a if they are already registered. The application allows two ways for the users to provide their personal mobility data. As mentioned, the users can upload their GPX files to annotate, or alternatively they can manually enter their data.

As shown in Figure 3.8b, the users are given detailed instructions on how to get valid

<sup>10</sup><http://www.openlayers.org>

<sup>11</sup><http://code.google.com/apis/maps/index.html>



**Figure 3.7: Place Activity Time application database schema**

logs: to check for a valid satellite fix before moving, to check the device clock is set correctly, and that only GPX files are accepted. Then a user can select any trip from the uploaded log to annotate. Once a trip is selected, it is visualised on a map with the source and destination places shown differentiated from the rest of the trip. The user can click on a place to enter its name, type, times of visit, and the activities taken at the place, see Figure 3.8d. The system attempts to automatically infer the visits times, simply by retrieving the times-stamps of the last spatio-temporal point in the current trip and the first spatio-temporal point in the next trip (assuming that the user logs his mobility all the time). Pre-processing the logs not only improved the place identification for the purpose of constructing the PG, but it also allowed to infer the times of visits more accurately, as it corrected many segmented tracks.

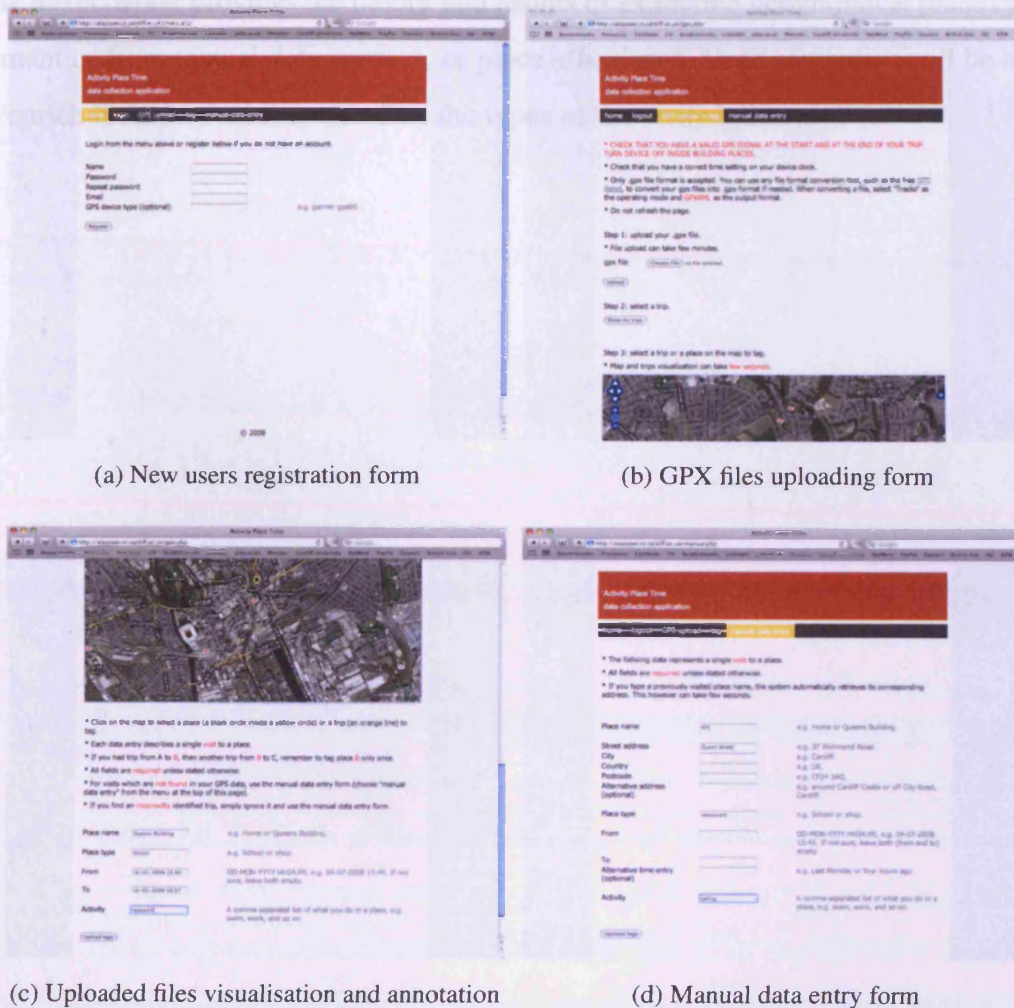


Figure 3.8: Place Activity Time – Web interface snapshots

### 3.8 Conclusion

This chapter elaborated on list of problematic issues which were found in real-world GPS logs. The chapter introduced several pragmatic solutions to those problems, as a pre-processing step prior to any further analysis to the logs. The proposed procedures showed good results. Merging segmented GPS tracks, eliminating overshoots points and uninformative points from GPS tracks allowed to extract better quality PG.

Pre-processing GPS logs in this way can help to improve other applications which employ such logs, and not necessarily PG-type of applications, only.

The next chapter provides the theory and results of extracting geographical place types semantics from textual data sources, or place affordance. Such semantics will be used to enrich the extracted PG, based on the types of the visited places.

# Extracting place-related services and activities

## 4.1 Place Ontology

Geographical places are normally associated with specific functions, economic activities, or services that they provide to individuals. These dimensions of a geographical place definition are typically evident in catalogues of place type specifications, produced by national mapping and other geographic data collections agencies, and are used for the purpose of classification of place entities. For example, the following descriptions are parts of the definitions associated with place types in the Ordnance Survey Mastermap specification <sup>1</sup>.

**Amusement park:** A permanent site providing entertainment for the public in the form of amusement arcades, water rides and other facilities.

**Comprehensive school:** A state school for teenagers, which provides free education.

Annotations assigned by people to a particular geographical place or place type are another valuable resource from which place affordance concepts can be harvested. For

---

<sup>1</sup><http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/>



example, the DBpedia data set, created by extracting structured data from the Wikipedia, provides detailed geographic information for over 400,000 places<sup>2</sup>. On the other hand, Wordnet and OpenCyc knowledge-bases contain several descriptions of place types and human activities. Moreover, in a mobile context, people annotations of geographical places are seen as a source of place semantics (Wang and Canny, 2006; Espinoza et al., 2001). Classification of economic activities of business establishments is often used for place type categorisation. For example, national bodies such as the Office of National Statistics of the UK (ONSUK)<sup>3</sup> and Eurostat (the statistical office of the European commission), produce classifications and definitions of economic activities, for classifying business establishments by the type of economic activity in which they are engaged<sup>4</sup>. Notably, a business place can be associated with a number of services, where some of these are principal activities that determine its primary classification while others are ancillary activities (such as accounting, transportation, purchasing, and repair and maintenance, etc...) that exist solely to support the principal ones.

The place ontology in Figure 2.5 is extended here with the concepts of Place Type and Service/Activity Type. Affordance of place is used here to mean the specific notion of a service offered at that place or an activity that can be practised there.

The proposed model is shown in Figure 4.1<sup>5</sup>. Here, a place instance is associated with a particular place type which in turn identifies the set of possible services linked to the place. The class of Service/Activity types in this model is used to encompass both the notions of economic as well as other human activities commonly associated with a place type. A place type can afford one or more service types.

Relevance of particular services to a place may depend on the application or context in which this ontology is used, and which can therefore determine the strength of the

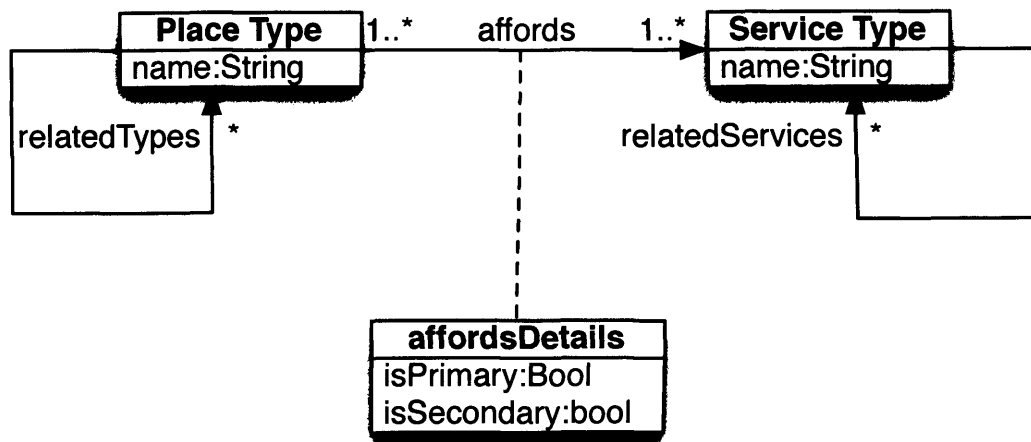
---

<sup>2</sup><http://wiki.dbpedia.org/datasets>

<sup>3</sup><http://www.statistics.gov.uk>

<sup>4</sup>See The Standard Industrial Classification of all Economic Activities (SIC), [http://www.statistics.gov.uk/methods\\_quality/sic/downloads/sic2007explanatorynotes.pdf](http://www.statistics.gov.uk/methods_quality/sic/downloads/sic2007explanatorynotes.pdf)

<sup>5</sup>RWPO stands for Real World Place Ontology, which will be fully explained in the next section.



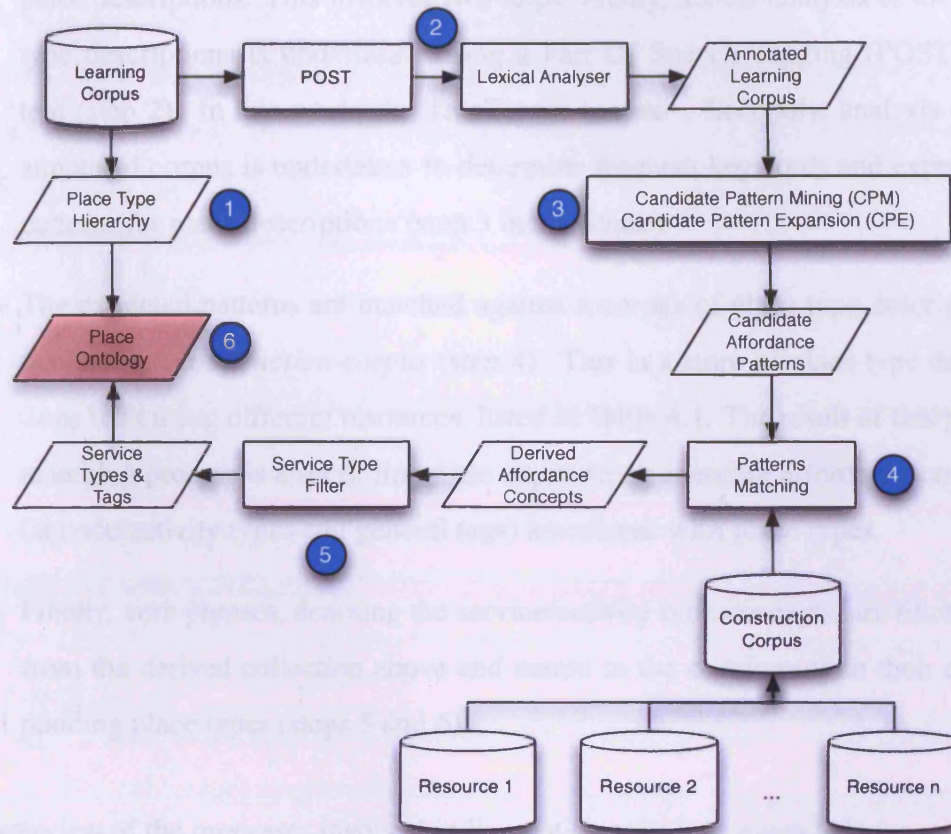
**Figure 4.1: Place ontology (RWPO).**

association between the two classes. For example, while the service of *catering* may be available at places of type *university* (assuming having a canteen inside the university) or *hotel*, it is more related to places of the latter type than the former. This is referred to as "*isPrimary*" and "*isSecondary*" in the Figure 4.1.

## 4.2 Ontology building approach

The methodology introduced in this work is based on frequency-based lexical analysis of a corpus of place type descriptions, denoted the *learning corpus*, to discover candidate patterns for service and activity type identification. The set of patterns identified is then applied on a corpus of multiple resources of place type definitions, denoted the *construction corpus*, to extract possible service and activity types, which are then evaluated. The methodology is illustrated in Figure 4.2, and summarised in the following steps:

- A catalogue of place type descriptions from the OS (referred to as the Real World



**Figure 4.2: System architecture for service types extraction and place ontology population.**

Object Catalogue (OSRWOC)<sup>6</sup> is used to populate place type concepts in the ontology. A total of 550 place types were identified, linked together in a hierarchy and stored in the ontology (step 1 in Figure 4.2). The associated place type descriptions from the catalogue form the *learning corpus*. The resulting place ontology will henceforth be referred to as the Real World Place Ontology (RWPO).

- Linguistic analysis of the *learning corpus* is applied to determine patterns of

<sup>6</sup>OSRWOC: <http://www.ordnancesurvey.co.uk/oswebsite/docs/legends/os-mastermap-real-world-object-catalogue.pdf>

place descriptions. This involves two steps. Firstly, lexical analysis of the place type descriptions is undertaken using a Part Of Speech Tagging (POST) system (step 2). In this work, the TreeTagger is used<sup>7</sup>. Secondly, analysis of the annotated corpus is undertaken to determine frequent keywords and expression patterns for place descriptions (step 3 in the figure).

- The extracted patterns are matched against a corpus of place type descriptions, denoted the *construction corpus* (step 4). This is a store of place type descriptions built using different resources, listed in Table 4.1. The result of this pattern matching process is a set of linguistic expressions denoting affordance concepts (service/activity types and general tags) associated with place types.
- Finally, verb phrases, denoting the service/activity type concepts, are filtered out from the derived collection above and stored in the ontology with their corresponding place types (steps 5 and 6).

An overview of the processes involved in linguistic analysis is given below.

**Table 4.1: Data sources for place type definitions used as a construction corpus**

Data source	No. of place types
Ordnance Survey RWOC	555
Wikipedia	345
Oxford English Dictionary OED	273
OpenCyc	191
WordNet	192

### 4.2.1 Candidate Pattern Mining (CPM)

The algorithm starts by iterating over each set of tokens, for each place type, constructing forward combinations of lemmas. For example, if the token set is  $sT := \{l_1, l_2, l_3\}$ ,

<sup>7</sup><http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/DecisionTreeTagger.html>.

**Algorithm 4.1: Candidate Patterns Miner (CPM)**

**Input:** Corpus of place types ( $pT$ ), where each  $pT$  has a set of string tokens ( $sT$ ), and each  $sT$  is annotated with its *POST* and *lemma*, no. of tokens threshold  $nT$ , and frequency threshold  $fT$

**Output:** A set of candidate patterns ( $cP$ ).

- 1: for each  $sT$  in each  $pT$ , construct  $cP$  of forward combinations of lemmas
- 2: for each  $cP$ , count its frequency in corpus
- 3: sort  $cP$  set by frequency
- 4: for each  $cP$ , delete  $cP$  if its no. of lemmas  $< nT$
- 5: for each  $cP$ , delete  $cP$  if its frequency  $< fT$
- 6: sort  $cP$  set by no. of lemmas
- 7: for each  $cP$ , if  $cP$  is part of next  $cP$  and its frequency  $<$  the frequency of next  $cP$  then delete  $cP$

then the forward lemma combinations are  $cP := \{l_1, l_2, l_3, l_1l_2, l_1l_3, l_1l_2l_3, l_2l_3\}$ . The frequency of these lemmas in the learning corpus is calculated. Thresholds for the length and the frequency of occurrence of the resulting lemmas is set to determine a set of plausible patterns (a frequency threshold of 5 is used in this experiment). Further filtering of the patterns is done to exclude redundancy (patterns subsumed by others). See Algorithm 4.1.

In step 1, the algorithm constructs a set of forward combinations of lemmas as explained above, for reach string and for each place type. In step 2, the frequency of each of the constructed sets is counted, for the whole corpus. In step 3, the lemma sets are sorted by their counted frequencies from the previous step. In steps 4 and 5, each lemma set is checked for two constraints. Namely, if the number of lemmas inside the set is less than a given threshold, and if the set frequency is less than a given number, respectively. In step 6, the surviving lemma sets from the previous two steps are sorted by their frequencies. A final check step is performed in step 7, where each lemma set is checked if it is part of its next lemma set (in the sorted set) and if its frequency is

**Algorithm 4.2: Candidate Patterns Expansion (CPE)****Input:** A set of annotated place types  $pT$  and a set of candidate patterns ( $cP$ )**Output:** Each candidate pattern is expanded with its preceding POST in each token set in  $pT$ .

- 1: for each  $cP$  do steps 2-3
- 2: for each  $pT$  do step 3
- 3: for each token set in  $pT$ , if  $cP$  is part of token set lemmas, then concatenate  $cP$  with preceding POST until the end of token set

less than the frequency of its next lemma set, in which case it is deleted.

**4.2.2 Candidate Pattern Expansion (CPE)**

The output of the CPM is a set of candidate patterns, where each is a set of lemmas corresponding to a set of tokens from which the lemmas were derived, for example “*a place where*” or “*a place for*”. Typically, a set of lemma tokens is followed by different tokens in different sentences of place type descriptions. An expansion process is then applied to retrieve all those possible tokens in the place type definitions. The algorithm concatenates a set of lemma tokens with the POST of its preceding tokens. For example, a candidate pattern  $\{l_1l_2\}$  from CPM may be expanded as follows:  $\{l_1l_2DT VBN, l_1l_2VB NNS, l_1l_2\dots\}$  where *DT*, *VBN*, *VB*, *NNS* are examples of standard POST. A full list is given in Table 4.2.

In step 1, each of the given candidate patterns is checked for steps 2 and 3. For each place type (step 2), for each token set for the given place type, it is checked if the given candidate pattern is contained inside that token set. If the condition is satisfied, then the POST of the rest of the sentence (from the token set) is concatenated with the candidate pattern (as explained above).

A pattern takes the form of:  $\{place\text{-keyword linker} \dots POST^* \}$ . The list of *place-*

*keywords* compiled from the place descriptions in the learning corpus is: *place, building, establishment, facility, institution and organisation*. A *linker* is a word which may immediately follow a place-keyword, such as *where* or *of*. POST\* is list of one or more part of speech tags and “...” indicates a non-consecutive occurrence of tokens in the pattern. The list of POSTs identified in the patterns is then used to infer relevant place concepts. These take the form of concatenated sets of nouns and verb phrases. It is important to note that action verbs on their own are not sufficient to identify a service or activity. Instead, it is more meaningful to formulate a verb phrase, composed of a combination of a verb and one or more nouns. For example, a pattern such as *place where ... NN\* ... VBN*, e.g. “*place where tennis is taught or played*”, will produce the concepts “*teachTennis*” and “*playTennis*” instead of the verbs *teach* and *play*. The set of patterns identified in this experiment is listed in the following section. The patterns identified are used to match against place type descriptions in the construction corpus and to identify and automatically select relevant service concepts from the extracted nouns and verb phrases. It is possible that place description in different resources in the construction corpus will yield similar verb phrases. These are filtered out manually to avoid redundancy.

**Table 4.2: List of part of speech tags (POST) used for lexical analysis ((Santorini, 1990)).**

POST	Definition	POST	Definition
CC	coordinating conjunction	DT	determiner
IN	preposition or subordinating conjunction	JJ	adjective
JJR	adjective, comparative	JJS	adjective, superlative
NN	noun, singular or mass	NNS	noun, plural
NP	proper noun, singular	NPS	proper noun, plural
TO	to	VB	verb, base form
VBD	verb, past tense	VBG	verb, gerund or present participle
VBN	verb, past participle	VBP	verb, non-3rd person singular present
VBZ	verb, 3rd person singular present	WDT	determiner
WP	pronoun	WRB	adverb

## 4.3 Extracted patterns

The result of mining the learning corpus is a set of candidate place definition patterns. The most frequent patterns identified are extracted as listed below.

**Pattern 1** is defined as,

*place-keyword where ... NN\* ... NNS\* ... VBN ... NN\* ... .*

Concepts extracted are in the form of:  $VB^*+NN^*$  and  $VB^*+NNS^*$ . An example is: *Tennis centre is an establishment where tennis is taught and played.* The lemma form is: *an establishment where tennis be teach and play*, the POST form is: *DT NN WRB NN VBZ VBN CC VBN SENT*, and the service types extracted are: *teachTennis* and *playTennis*.

**Pattern 2** is defined as,

*place-keyword of ... NN\* ... NNS\* ... .*

Concepts extracted are in the form of:  $NN^*$  and  $NNS^*$ . An example is: *Winter Garden is an indoor garden containing exotic plants, or a place of entertainment, usually associated with coastal resorts.* The lemma form is: *an indoor garden contain exotic plant, or a place of entertainment, usually associate with coastal resort*, the POST form is: *DT JJ NN VBG JJ NNS, CC DT NN IN NN, RB VBN IN JJ NNS SENT*, and the concepts extracted are: *entertainment* and *coastalResorts*. Note the adjective concatenation in *coastalResorts*.

**Pattern 3** is defined as,

*place-keyword of ... VBG ... NN\* ... NNS\* ... .*

Concepts extracted are in the form of:  $VBG^*+NN^*$  and  $VBG^*+NNS^*$ . An example is: *Polytechnic: an institution of higher education offering courses at*



*degree level or below.* The lemma form is: *an institution of high education offer course at degree level or below*, the POST is: *DT NN IN JJR NN VBG NNS IN NN NN CC RB SENT*, and the concepts extracted are: *OfferingHigherEducation*, *OfferingDegree*, *OfferingDegreeLevel*, and *OfferingCourses*.

**Pattern 4** is defined as,

*place-keyword for ... NN\* ... NNS\* ... .*

Concepts extracted are in the form of: *NN\** and *NNS\**. An example is: *Hall is a large building for meetings or entertainment.* The lemma form is: *a large building for meeting or entertainment*, the POST form is: *DT JJ NN IN NNS CC NN SENT*, and the service types extracted are: *entertainment* and *meetings*.

**Pattern 5** is defined as,

*place-keyword in which ... NN\* ... VBN ... NNS\* ... NN\* ... .*

Concepts extracted are in the form of: *VB+NN\** and *VB+NNS\**. An example is: *Corn mill. A gristmill or grist mill is a building in which grain is ground into flour, or the grinding mechanism itself.* The lemma form is: *a gristmill or grist mill be a building in which grain be grind into flour, or the grind mechanism itself*, the POST form is: *DT NN CC NN NN VBZ DT NN IN WDT NN VBZ VBN IN NN, CC DT VBG NN PP SENT*, and the concepts extracted are: *grindGrain*, *grindFlour* and *grindMechanism*.

**Pattern 6** is defined as,

*place-keyword used to ... NN\* ... VB ... NN\* ... .*

Concepts extracted are in the form of: *VB+NN\**. An example is: *a smelter is an establishment used to extract metal from ore by melting. see building.* The

lemma form is: *an establishment use to extract metal from ore by melting. see building*, the POST form is: *DT NN VBN TO VB NN IN NN IN NN SENT VB NN SENT*, and the concepts extracted are: *extractMetal*, *extractOre*, and *extractMelting*.

**Pattern 7** is defined as,

*place-keyword used for ... NN\* ... VBG ... NN\* ... .*

Concepts extracted are in the form of: *VBG+NN*. An example is: *A meteorological Centre is a facility used for studying and recording facts about the weather*. The lemma form is: *a facility use for study and record fact about the weather*, the POST form is: *DT NN VBN IN VBG CC VBG NNS IN DT NN SENT*, and the concepts extracted are: *studyingWeather* and *recordingWeather*.

**Pattern 8** is defined as,

*place-keyword VBG ... NN\* ... NNS\* ... .*

Concepts extracted are in the form of: *VBG+NN\** and *VBG+NNS\**. An example is: *a civic centre is a building containing municipal offices*. The lemma form is: *a building contain municipal office*, the POST form is: *DT NN VBG JJ NNS SENT*, and the concept extracted is: *ContainingMunicipalOffices*. Note that there is no *linker* in this pattern.

**Pattern 9** is defined as,

*place-keyword that/which ... VB/VBG to/for ... NN\* ... NNS\* .*

Concepts extracted are in the form of: *VB/VBG+NN\** and *VB/VBG+NNS\**. An example is: *a Leisure Pool is a facility that may include swimming pools, fun*

*pools, flumes and other recreational activities.* The lemma form is: *a facility that may include swimming pool, fun pool, flume and other recreational activity*, the POST form is: DT NN WDT MD VB NN NNS, NN NNS, NNS CC JJ JJ NNS SENT VB NN SENT, and the concept extracted is: IncludeSwimming, IncludeFun, IncludeSwimmingPools, IncludeFunPools, IncludeFlumes, IncludeRecreationalActivities.

## 4.4 Results

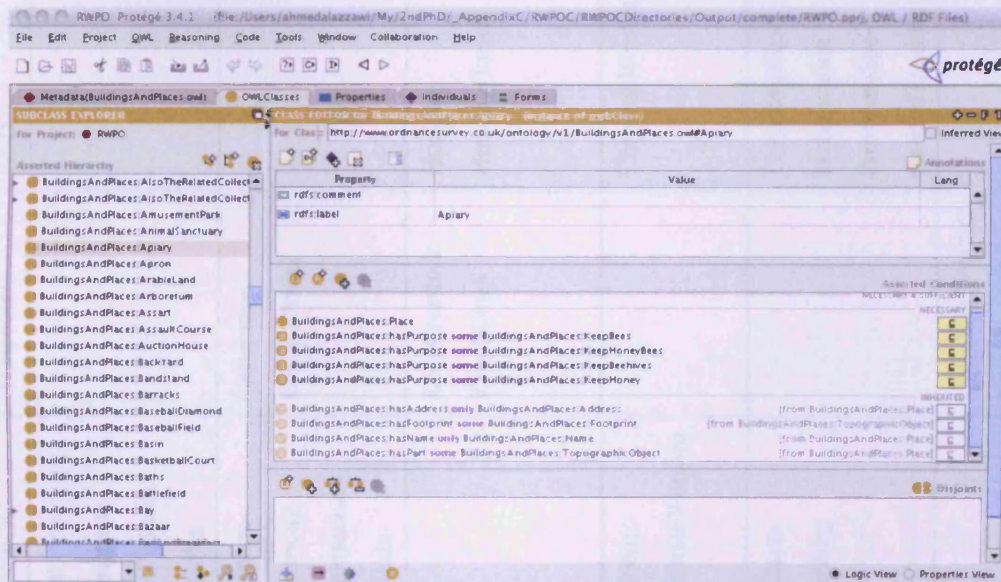


Figure 4.3: A snapshot of RWPO in Protégé

The patterns identified were applied to place descriptions compiled in the construction corpus. A set of just over 1550 concepts was extracted containing a mixture of verb phrases and nouns. Verb phrases were filtered out and used as service/activity concepts and nouns were stored as general tags of their corresponding place types. In many cases, the latter set of tags included references to the typical consumers and providers

**Table 4.3: Sample of concepts extracted from individual data sources in the construction corpus**

Place type	Service Concepts and General Tags				
	OSRWOC	Wikipedia	OED	OpenCyc	WordNet
Fire Station	Fire Equipment FireFighters Vehicles			KeepFireFig- htingEquipment KeepFirefighters	
Power Station	Generate- Electricity GenerateLar- geScale	Generation ElectricPower		Transformers Generators	
Stable		KeepLivestock KeepHorses	Housing HousingHorses	Farm Ranch Animals Horses	
Church	Worship ChristianWorship		ChristianWorship	HoldChristian Services HoldActivities	Public Worship
College	Students Learning		ProvidingHig- herEducation ProvidingSpe- cializedTraining		
Weather station		ObservingWeather ObservingClimate Observing- Instruments			

Place type	Service Concepts and General Tags				
	OSRWOC	Wikipedia	OED	OpenCyc	WordNet
		ObservingAtmo- sphericConditions ObservingWe- atherForecasts			
Foundry	MeltMetal MeltGlass MeltShapes FormMetal FormGlass FormShapes			DoMetalCasting	
Glasshouse		CultivatePlants			Glass Cultivation Exhibition Plants Controlled- Conditions
Shipyard	BuildShips RepairShips	BuildShips RepairShips	ConstructShips ConstructBoats		
Arsenal	MakeAmmunition MakeMilitatryEquipment MakeWeapons StoreAmmunition StoreMilitatryEquipment StoreWeapons	Construction Repair Storage Ammunition Weapons			Manufa- ctureArms

of the services for a particular place type. No attempt to refine this set was made in this work, but it should be possible to identify pattern expressions to decode some of these concepts in the future.

The system was developed using the Jena Semantic Web Toolkit<sup>8</sup> and the place ontology was represented in OWL. Figure 4.3 shows a snapshot of the ontology in Protégé<sup>9</sup>, showing a sample of place types and associated service concepts.

Employing multiple data sources in the construction corpus was valuable, as in many cases the sources complemented one another with regards to missing place types and service/activity concepts, as shown in Table 4.3. For example, for the place type *Arsenal*, the tag extracted from Wikipedia was “*Weapons*”, while the OSRWOC provided the concepts “*StoreWeapons*” and “*MakeWeapons*”.

## 4.5 Evaluation

Two evaluation tests were undertaken to measure the quality of the developed ontology. The first is a qualitative user-based evaluation and the second is an evaluation against an ontology created by experts. A set of 20 place types was used in both experiments for comparison.

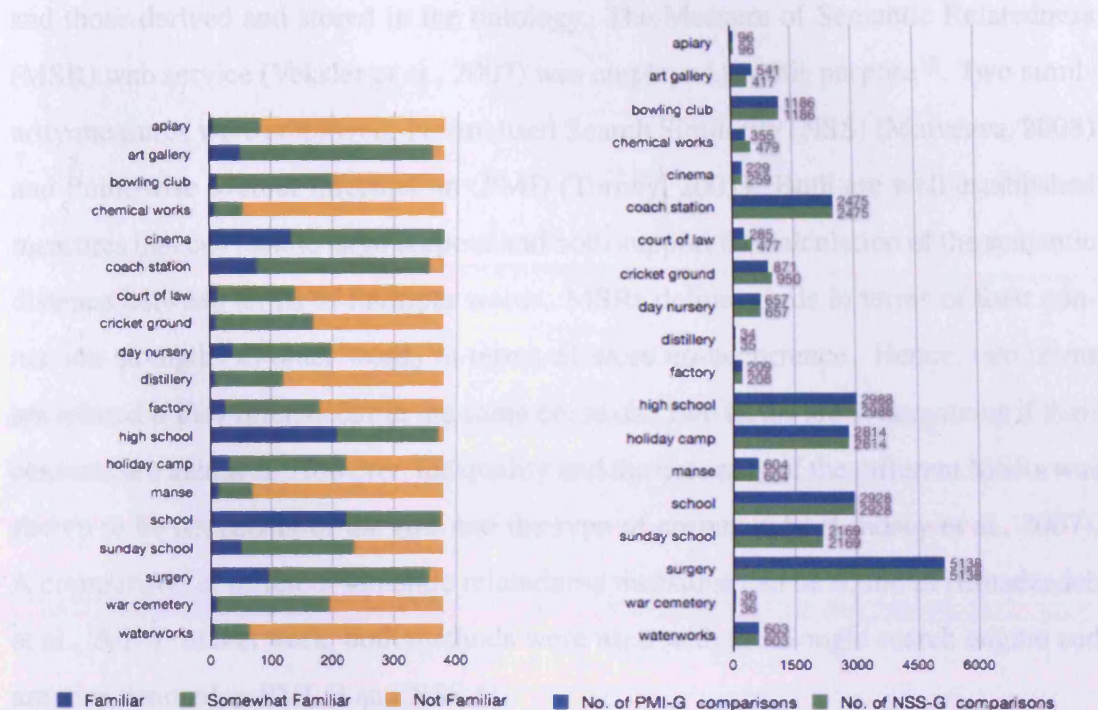
### 4.5.1 User-based evaluation

A qualitative user-based evaluation in the form of an online questionnaire was carried out to assess the quality of the resulting ontology. Users were presented with a list of place types and asked to identify for each place type one or more typical service concepts that can be provided or activities that can be undertaken by people in a place

---

<sup>8</sup><http://jena.sourceforge.net>.

<sup>9</sup>Protégé is a free, open source ontology editor and knowledge-base framework, <http://protege.stanford.edu>



(a) The participants familiarity with the place types used the survey (b) The number of terms compared for each place type using two measures of semantic relatedness

**Figure 4.4: Survey statistics.**

of such a type. The semantic similarity between the service/activity concepts derived in this work and the concepts identified by the users was then measured. The questionnaire was hosted for a period of two weeks, during which students and staff from the university were invited to take part. A total of 872 people attempted the questionnaire, of which 388 (44.5%) completed the questionnaire. Around half of the participants were under 25 years old, 90.6% were native English speakers, and 69.1% were females. The participants were asked to identify their familiarity with each of the given place types as either: “*familiar: you visit or have visited places of this type on a daily basis or a few times a week*”, “*somewhat familiar: you visited places of this type a number of times before*”, or “*not familiar: you have not visited places of this type before*”. The distribution of user familiarity with place types is plotted in Figure 4.4a.

Semantic similarity was measured between the concepts identified by the participants

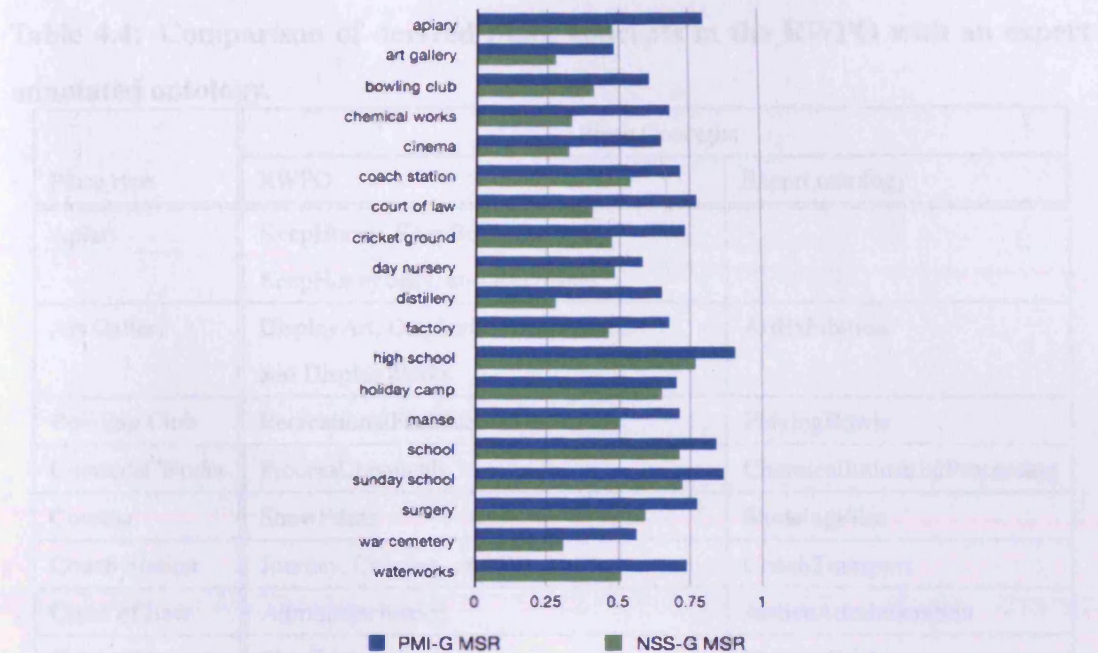
and those derived and stored in the ontology. The Measure of Semantic Relatedness (MSR) web service (Veksler et al., 2007) was employed for this purpose<sup>10</sup>. Two similarity measures were employed: Normalised Search Similarity (NSS) (Matveeva, 2008) and Point-wise Mutual Information (PMI) (Turney, 2001). Both are well established measures that can handle large corpora and both support the calculation of the semantic distance between terms of multiple words. MSRs define words in terms of their connection strengths to other words in terms of word co-occurrence. Hence, two terms are related if they often occur in the same contexts. Two terms are synonymous if their contexts are identical. However, the quality and the accuracy of the different MSRs was shown to be dependent of the size and the type of corpus used (Lindsey et al., 2007). A comparison of different semantic relatedness measures can be found in (Emadzadeh et al., 2010). In this work, both methods were used with the Google search engine and are thus denoted as PMI-G and NSS-G.

For each place type, a combined measure of semantic relatedness is calculated as the average of relatedness of all the extracted terms in the RWPO ontology with the terms submitted by users. To compare two terms, MSR sends search queries to the Google search engine (consisting of the original individual terms as well as of their concatenation, where a term may consist of multiple words). Based on the number of results returned, a semantic distance is calculated. Figure 4.4b shows the number of term comparisons for every place type using both MSRs. The number of term comparisons generally correlates with the participants familiarity with the place types considered as shown in the figure. The result of the average MSR is plotted in Figure 4.5 for each of the place types considered using both NSS and PMI. The strength of the similarity using PMI, for most the place types, approached or exceeded 50%. It is noted that the overall performance of the two MSRs was highly correlated (with a correlation coefficient of 0.795). A similar observation was also found in (Lindsey et al., 2007). Figure 4.6 shows sample term comparisons using both NSS and PMI. For each pair of the terms compared, the first term is a service/activity type from the survey and the second

---

<sup>10</sup><http://cwl-projects.cogsci.rpi.edu/msr>.





**Figure 4.5: The average semantic relatedness between RWPO derived concepts and those identified in the questionnaire.**

is a concept from the RWPO. The experiment demonstrates how the automatically extracted service/activity types can be semantically close to the user perception of place affordance.

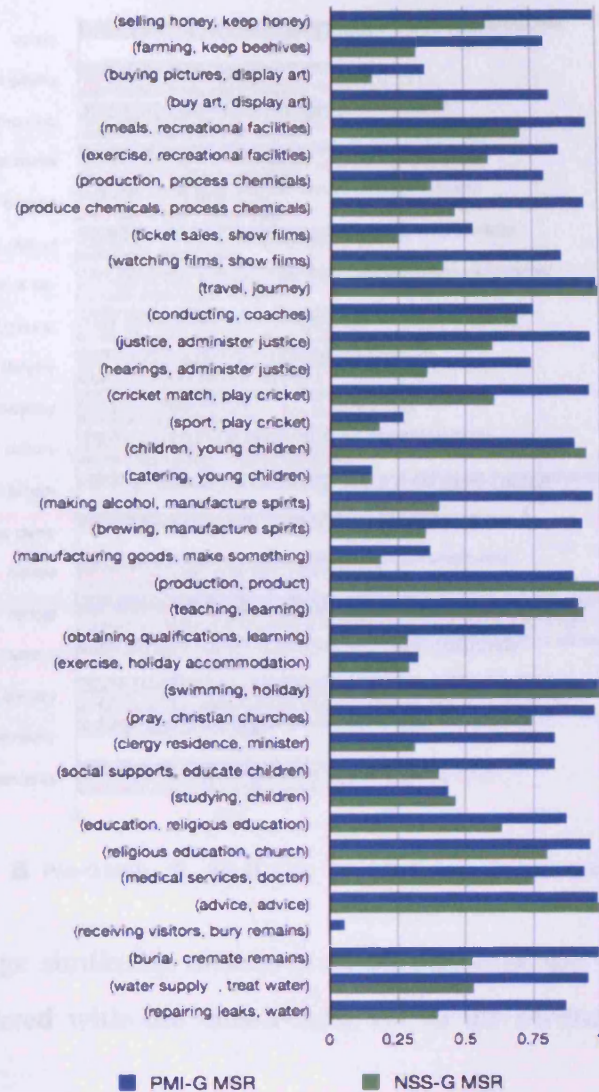
#### 4.5.2 Comparison with an expert ontology

An ontology was developed by the Ordnance Survey to describe building features and the place classes surveyed, “with the intention of improving the use of the surveyed data and enabling semi-automatic processing of these data<sup>11</sup>”. The ontology is expressed in a controlled natural language, called *Rabbit*, that supports the authoring of OWL ontologies by domain experts (Denaux et al., 2010).

<sup>11</sup>Building and Place ontology: <http://www.ordnancesurvey.co.uk/oswebsite/ontology>.

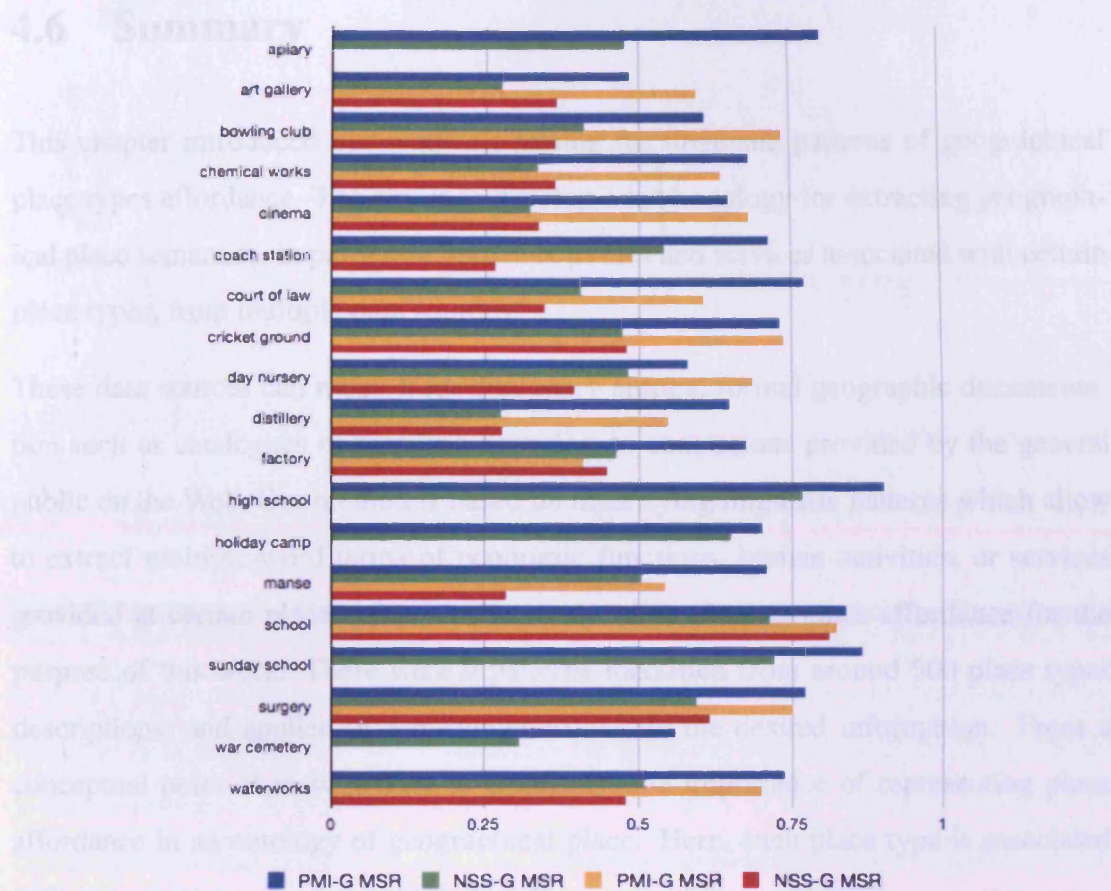
**Table 4.4: Comparison of derived place concepts in the RWPO with an expert annotated ontology.**

Place type	Place Concepts	
	RWPO	Expert ontology
Apiary	KeepHoney, KeepBeehives, KeepHoneyBees, and KeepBees	–
Art Gallery	DisplayArt, DisplayPublicViewing, and DisplayWorks	ArtExhibition
Bowling Club	RecreationalFacilities	PlayingBowls
Chemical Works	ProcessChemicals	ChemicalIndustrialProcessing
Cinema	ShowFilms	ShowingFilm
Coach Station	Journey, Coaches, and Passengers	CoachTransport
Court of Law	AdministerJustice	JusticeAdministration
Cricket Ground	PlayCricket	PlayingCricket
Day Nursery	YoungChildren	InfantCare
Distillery	DistilAlcoholicSpirits, and ManufactureSpirits	DistillingSpirit
Factory	MakeSomething, ConstructSomething, Product, Workers, Goods, Machines, ManufactureMachine, ManufactureGoods, AssembleMachine, and AssembleGoods	Manufacturing
High School	Learning, Teenagers, and Girls	–
Holiday Camp	Holiday, HolidayAccommodation, and Entertainment	–
Manse	Minister and ChristianChurches	HousingMinister_Religion
School	EducateChildren, Learning, and Children	Education
Sunday School	Church, ReligiousEducation, and Children	–
Surgery	Doctor, Dentist, Vet, Advice, Treatment, MedicalPractitioner, and Patients	HealthCareProvision
War Cemetery	CremateDeadBodies, CremateRemains, BuryDeadBodies, and BuryRemains	–
Waterworks	TreatWater, SupplyWater, Water, and WaterSupply	WaterTreatment and SupplyWater



**Figure 4.6: Sample terms comparisons using the semantic relatedness measures. The first is a term collected in the questionnaire and the second is a term in the RWPO ontology.**

A place concept in the ontology is defined as a kind of topographic object with the properties *purpose*, *address*, and *name*. The property “purpose” is used to hold notions of activity or processes associated with a place and is defined by experts and explicitly stored in the ontology. Table 4.4 shows the high correlation between the values of this property in the expert ontology with those derived and stored in the RWPO for the set



**Figure 4.7: Average similarity measures of service concepts extracted for some place types compared with the benchmark set in the Ordnance Survey expert ontology.**

of place types in this experiment. Figure 4.7 shows a semantic comparison between the concepts of place types from RWPO and the collected concepts in the user study, for the same place types, employing the two semantic measures mentioned earlier (the green and blue bars). For the same place types, the same comparisons are performed between the concepts from the expert ontology and the collected concepts in the user study (yellow and red bars).

## 4.6 Summary

This chapter introduced our work on mining for linguistic patterns of geographical place types affordance. The aim is to develop a methodology for extracting geographical place semantics, in particular human activities and services associated with certain place types, from multiple data sources.

These data sources can range from dictionary entries, formal geographic documentation such as catalogues of mapping agencies, to annotations provided by the general public on the Web. Our method is based on identifying linguistic patterns which allow to extract multiple-word terms of economic functions, human activities, or services provided at certain place types – collectively referred to as place affordance for the purpose of this work. There were 9 patterns identified from around 500 place types descriptions, and applied to 5 resources to extract the desired information. From a conceptual point of view, we try to emphasise the importance of representing place affordance in an ontology of geographical place. Here, each place type is associated with one or more service or activity types.

Applying the patterns found provided good results. Using multiple data sources for the information extraction was beneficial as many resources complemented the others in terms of the extracted terms. There were, however, some negative results and limitations, more on which will be given in Chapter 8.

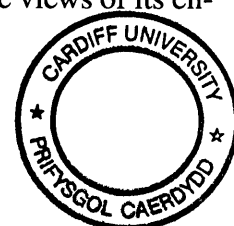
# Graph model of human mobility

## 5.1 Introduction

This chapter introduces the proposed model of human mobility, specifically seen as a semantically rich PG. We begin by discussing different views of modelling human mobility data as described in section 5.2, followed by introducing the graph based model in section 5.3. Section 5.4 introduces the details of the algorithm designed to construct the model from GPS data. To show the usefulness of the suggested model, section 5.5 introduces the details of an algorithm which employs the model to find the relatedness of an arbitrary geographical place to a certain person, as a possible personalisation application. Finding the most related places to a person is useful in many scenarios, such as filtering unrelated places on a map displayed on a mobile device. Finally, the chapter ends with a conclusion in section 7.3.

## 5.2 Modelling human mobility

Space can be conceptualised at different levels of hierarchy. A space can be seen from a geometrical perspective where its entities are defined by their spatial co-ordinates. A space can also be seen from a topographical perspective where it is defined by its places, regions, and paths. For example, in (Kuipers, 2000) different levels of spatial hierarchy were defined in the context of a mobile robot and its possible views of its en-



vironment (such as sensory, motion, trajectory-following, turning and moving actions perspectives). For the context of this work, *Personal Mobility (PM)* is the trace of a human who is moving in the physical real-world<sup>1</sup>. A PM data set consists of spatio-temporal points representing:

- i*) the places a person visited; and
- ii*) the spatial routes taken to travel between those places – along with their relative time-stamps.

PM is:

- i*) multi-directional in space (a person can move in any direction in space) and uni-directional in time (time goes forward only for any person);
- ii*) dynamic (as it reflects the mobility of an individual); and
- iii*) unique: no more than one person can acquire the exact same PM data set.

To help modelling PM, it is conceptualised at different levels of abstractions – including different aspects of place representation, see Figure 5.1.

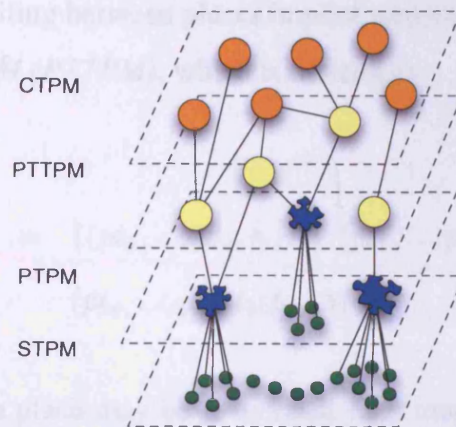
On a primitive level, it can directly capture the GPS trace as a set of unique spatio-temporal points: *Spatio-temporal PM (STPM)* is defined as:

$$STPM := \{(s_1, t_1), (s_2, t_2), (s_3, t_3), \dots, (s_n, t_m)\} \quad (5.1)$$

where each pair  $(s_n, t_m)$  consists of a spatial point  $(s_n)$  and a temporal point  $(t_m)$ . For simplicity, it is assumed that the space is two dimensional. Namely, a spatial point is defined by a pair of longitude and latitude coordinates, specified by a coordinate

---

<sup>1</sup>Note the terminology difference with PG, as each refers to a different view, as will be explained later.



**Figure 5.1: Different levels of abstracting personal mobility**

system. Note that a spatial point may be repeated in a STPM data set, only with different time-stamps. A person can physically be only in a single place at any specific time. No distinction is made at this level between any geographical features in STPM, such as a geographic place or a road network.

A subset of a STPM data set is the set of places which a person visits such as home and work place, and so on. These places are obtained by further abstraction of the *STPM* by grouping spatio-temporal points (of the GPS trace) that represent a visit to a specific place. Hence the *place-temporal PM (PTPM)* is defined as:

$$PTPM := \{(p_1, t_1, t_2), (p_2, t_3, t_4), (p_1, t_5, t_6), \dots, (p_a, t_b, t_{b+1})\} \quad (5.2)$$

where each triplet  $(p_a, t_b, t_{b+1})$  consists of a spatial footprint of a place ( $p_a$ ) and a pair of time-stamps ( $t_b, t_{b+1}$ ) (representing the beginning and the end of a time interval of a visit to a place). Similar to the spatial points in a STPM, a place in PTPM may be repeated with different time-stamps.

Typically, a geographical place has a place type such as a hospital, a university, a



school, and so on. Travelling between places implies moving in a space of place types; a *place-type-temporal PM (PTTPM)*, which is defined as:

$$PTTPM := \{(pt_1, \dots, pt_c, t_1, t_2), (pt_1, \dots, pt_d, t_3, t_4), \dots, (pt_e, \dots, pt_f, t_b, t_{b+1})\} \quad (5.3)$$

Note that in a PTTPM, a place may be associated with more than one place type and vice versa. As such, multiple place types may be paired with the same time interval.

As presented in the place model in Chapter 4, place type may be associated with one or more concepts, such as the services or the human activities typically provided at a place of such a type. Moving between place types implies moving between the concepts reflected by those place types, or the *concept-temporal PM (CTPM)*, which is defined as:

$$CTPM := \{(c_1, \dots, c_g, t_1, t_2), (c_1, \dots, c_h, t_3, t_4), \dots, (c_1, \dots, c_i, t_b, t_{b+1})\} \quad (5.4)$$

Similar to the PTTPM, multiple concepts may be associated with the same time interval.

To support the analysis at the different levels of abstraction, a model of *PM* needs to capture the different abstractions presented earlier along with their inter-relationships. Based on the above, a profile of *PM* needs to:

*i)* represent the quantitative elements of *PM*, namely the spatial and the temporal footprints, to facilitate statistical analysis (such as counting the number of visits to a certain place, predicting the likelihood of a visit to a place at a certain time, and so on);

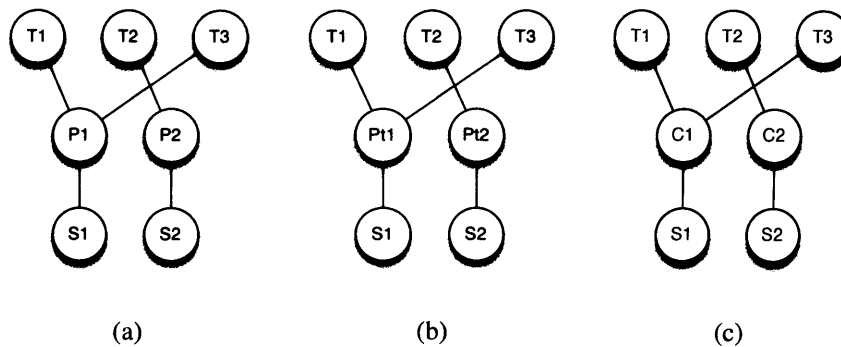
- ii) represent the qualitative elements of PM (as the objective is to construct a semantically-rich PG), namely the concepts about visited places, to facilitate information retrieval; and
- iii) for a certain place type with multiple concepts, the model should allow to differentiate between its different concepts for different people. However, as the assumption is that only the STPM is collected, this differentiation should be based on their relative STPM only.

A profile of PM is modeled as a coloured, weighted, and undirected graph as follows.

### 5.3 Personal Mobility Graph

One possible approach of modelling a *PM* is a tripartite graph between the spatial, temporal, and thematic dimensions of a human movement. Thus, a person visits a geographical place (represented by its name, place type, and its associated concepts) that is located in a certain geographical location, for a specific time interval.

An instance of this graph can represent individual place instances, as shown in Figure 5.2a. Another version can represent a place type as shown in Figure 5.2b, or a place concept as shown in Figure 5.2c.



**Figure 5.2: Possible models of *PM***

The nodes in the graph represent various elements captured from *PM* abstractions, and the edges represent their inter-relationships. Namely, the *S* nodes represent the spatial footprints of places, the *P* nodes represent the places, the *Pt* nodes represent the place types, and the *C* nodes represent concepts associated with certain place types. The edges in the graph associate each spatial location with its place (or a place type, or its relative concepts), and its time interval (of a visit).

A measure of strength between a person and a place (or a place type, or its relative concepts) can be inferred by analysing the spatial and temporal edges on the graph. The strength of this association can be represented as a weight on the relevant node in the graph.

For our purposes, the PMG is defined as follows:

**Personal Mobility Graph (PMG).** *A PMG is a coloured, weighted, and undirected graph  $G$  such that  $G = (C, S, T, SE, TE)$ .  $C$  is the set of weighted-concepts nodes – or conceptual nodes,  $S$  is the set of spatial footprints nodes – or spatial nodes,  $T$  is the set of temporal footprints nodes – or temporal nodes,  $SE$  is the set of spatial edges, where each spatial edge connects a spatial node to a conceptual node, and finally  $TE$  is the set of temporal edges, where each temporal edge connects a temporal node to a conceptual node. Each node and edge is unique in a PMG.*

**Conceptual Node.** *A conceptual node is a node  $c \in C$  where  $c$  is a pair of a concept and a weight. The concept is inferred or retrieved for a spatial node(s), and its weight is calculated based on the semantic relatedness of this concept to the other concepts in a PMG, as well as the spatio-temporal elements of a PMG – as will be explained in detail.*

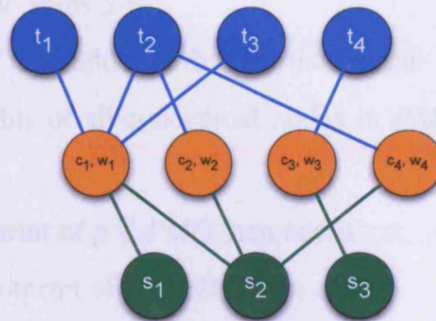
**Spatial Node.** *A spatial node is a node  $s \in S$  representing the spatial footprint a geographical place.*

**Temporal Node.** *A temporal node is a node  $t \in T$  representing the temporal footprint of a visit by a person to a certain place, i.e. the beginning and ending time-stamps of the period of a visit.*

**Spatial Edge.** A spatial edge is an edge  $se \in SE$  which connects a conceptual node to a spatial node.

**Temporal Edge.** A temporal edge is an edge  $te \in TE$  which connects a conceptual node to a temporal edge.

The following is an illustrating example.



**Figure 5.3:** A sample PMG. Conceptual nodes are coloured in orange, spatial nodes and edges are coloured in green, and temporal nodes and edges are coloured in blue.

**Example** Figure 5.3 shows a sample PMG. The figure is read as follows. A person visited places  $p_1$ ,  $p_2$ , and  $p_3$  (represented by  $s_1$ ,  $s_2$ , and  $s_3$  respectively). Place  $p_1$  gives conceptual node  $(c_1, w_1)$  and is visited at time  $t_1$  (as there is a spatial edge which connects  $(c_1, w_1)$  to  $s_1$  and a temporal edge which connects  $(c_1, w_1)$  to  $t_1$ ). The next visit is to place  $p_2$  at time  $t_2$  (as there are temporal edges which connects  $t_2$  to  $(c_1, w_1)$ ,  $(c_2, w_2)$  and  $(c_4, w_4)$  which are all connected via spatial edges to  $s_2$ ). Note that the conceptual node  $(c_1, w_1)$  is shared between places  $p_1$  and  $p_2$ . If the second visit was also made to  $p_1$ , then  $t_2$  would have been connected to  $(c_1, w_1)$  only – which is the case of the third visit. A final visit is made to  $p_3$  which is associated with a single conceptual node  $(c_3, w_3)$ . The process of assigning the weights will be explained below. Note that there is no direct connection between the spatial and the temporal nodes, as the main

**Algorithm 5.1: PMG Constructor****Input: STPM and data sources of place semantics****Output: PMG.**

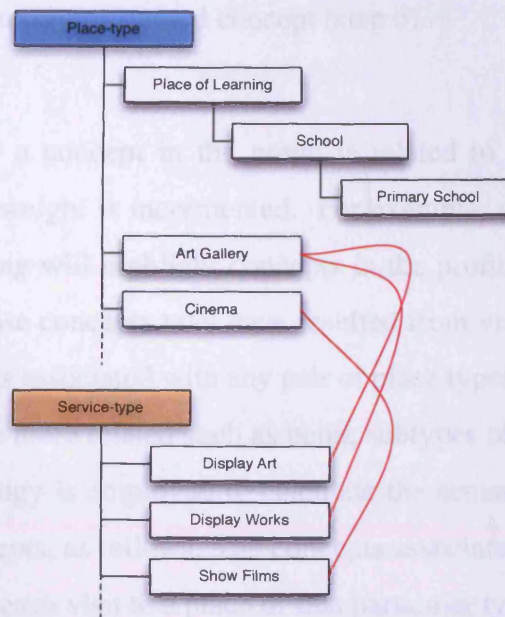
- 1: Cluster STPM into a set of places  $P$  and their relative times of visits
- 2: Retrieve place semantics  $pS$  for each place  $p$  in  $P$
- 3: For each visit to  $p$  do steps 4-9
- 4:     For each  $pS$  of  $p$  do steps 5-6
- 5:         If  $pS \notin PMG$  then add it with its initial weight
- 6:         Update weights of all conceptual nodes in  $PMG$  based on their semantic similarity with  $pS$
- 7:     If the spatial footprint of  $p \notin PMG$  then add it
- 8:     If the temporal footprint of  $p \notin PMG$  then add it
- 9:     Connect the spatial and temporal footprints of this visit to all its  $pS$  by adding the necessary spatial and temporal edges

interest is to associate the concepts with spatial and temporal tags. As mentioned, in the case of  $t_1$  for example, it can be inferred from the embedded information in the graph that it is associated with  $s_1$ .

## 5.4 PMG Constructor Algorithm

Algorithm 5.1 is designed to construct and update a PMG. It accepts a STPM of a certain individual and any data source of place semantics as its input, and produces a PMG as an output. In this work, place semantics are retrieved from an ontology of geographical place affordance, where each place type is associated with one or more service or human action types – as will be explained in more detail.

The algorithm begins by extracting the places visited with their relative times of visits from the inputted mobility log (step 1). For each place found, the algorithm retrieves



**Figure 5.4: Place-affordance ontology – a snapshot.**

its related concepts from the ontology (step 2). The necessary data sets which relate the spatial footprint of a place to its type are assumed to be available.

In this work, an ontology of geographical place affordance is used as a data source to fetch place semantics for certain place types. Each class in a place-type hierarchy is associated with one or more service or human activity types. These are the typical services and activities which are provided/occurring at a certain place type. For example, a place of type *art gallery* provides functions such as *display art* and *display works*, while a *cinema* is associated with *show films*, and a *distillery* with *distil alcoholic spirits* and *manufacture spirits*, and so on, see Figure 5.4.

For each visit to a place (step 3), each of its retrieved concepts is (step 4):

- i*) inserted to the graph with an initial weight if the concept does not exist in the graph (step 5); and
- ii*) the weights of all the concepts in the graph are updated based on their semantic

similarity with the currently processed concept (step 6).

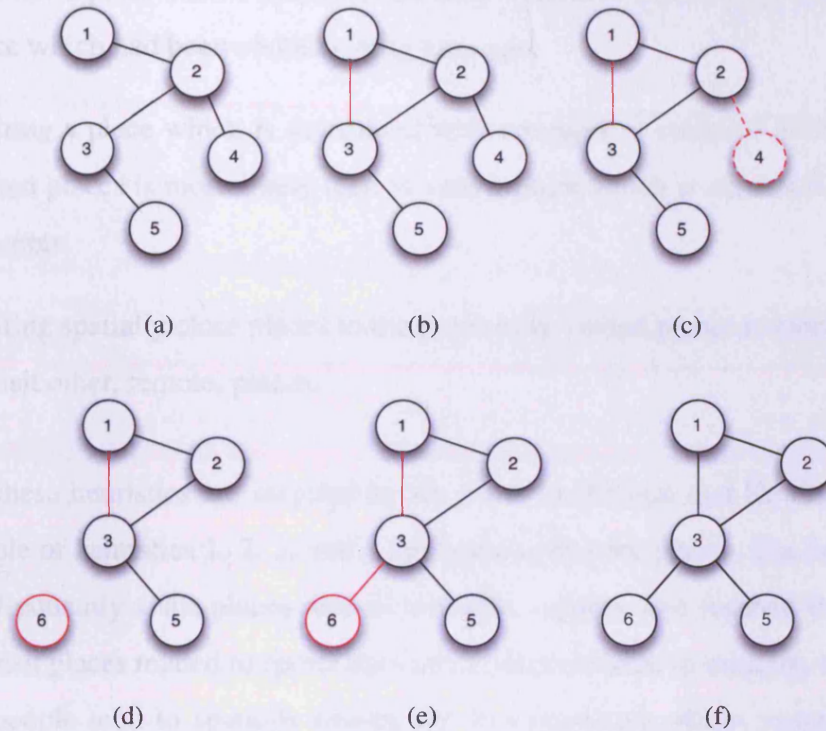
Simply put, the more a concept in the graph is related to the currently processed concept, the more its weight is incremented. For example, visiting a place which is associated with *learning* will highlight concepts in the profile such as *education* and *teaching* (although those concepts may have resulted from visiting other places). The idea is that the concepts associated with any pair of place types are more related if their relative place types are more related such as being subtypes of a super place type). see Figure 5.4. The ontology is employed to calculate the semantic relatedness between any pair of given concepts, as follows. The concepts associated to the given place type is incremented by 1 at each visit to a place of that particular type. The weights of other concepts are incremented according to their relevance to the given place type, and up to 1, see Equation 5.5:

$$weight := \frac{1}{SP(pt_1, pt_2) + LCA(pt_1, pt_2)} \quad (5.5)$$

The weight increment is a hybrid measure of the shortest path (SP) and the Lowest Common Ancestor (LCA) of the currently processed place type (of the place of the current visit), and another given place type (i.e. a place type which is associated with one or more concepts, and the weights of those concepts are being incremented in the weight update step). The suggested measure takes into account the siblings as well as the inheritance relations between the place types (Schwering, 2008). The weight increment is normalised to 1. In Figure 5.4, if a visit is made to a place of type Primary School, then the weight of the concepts associated with Primary School are incremented by 1, the weights of the concepts associated with School are incremented by  $1/1+2 = 0.333$ , while the weights of the concepts associated with place type Cinema are incremented by  $1/4+3 = 0.142$ , for example.

Finally, for each visit, its spatial and temporal footprints are inserted to the graph (con-

sidering if the spatial footprint exists in the graph or not) (steps 7-8), and then connected to the concepts of this visit as required (step 9).



**Figure 5.5: GED Illustrated.**

## 5.5 Personal Place Similarity Algorithm

To develop the appropriate algorithm to find the relatedness of any geographical place to a PMG, a number of heuristics are employed. Namely:

1. To visit a frequently visited place is more likely than to visit a new place.
2. To visit a place which previously had been visited for a long time (summing all the time periods of all the previous visits to that place) is more likely than to visit a new place or a place which had been visited for a short time.



3. To visit a place at a certain time period at which it had been frequently visited before is more likely than to visit that place at another time.
4. To visit a place which had been recently visited is more likely than to visit a place which had been visited a long time ago.
5. Visiting a place which is associated with concepts in common with previously visited places is more likely than to visit a place which is associated with other concepts.
6. Visiting spatially close places to the previously visited places is more likely than to visit other, remote, places.

Some of these heuristics are inspired by the work in (Flamm and Kaufmann, 2007). An example of heuristics 1, 2, 3, and 4 are home and work places. For heuristic 5, if a person frequently visits places related to tennis, squash, and football then its more likely to visit places related to sports than other places related to military, for example. Usually, people tend to spatially cluster the less important places around the more frequently visited places, such as tending to shop in nearby places to home or work place (Flamm and Kaufmann, 2007), as an example of heuristic 6. These heuristics are employed in Algorithm 5.2, as will be explained.

Graph Edit Distance (GED) is a distance measure between two graphs. It is employed in inexact graph matching problems, where the objective is not only to determine if the given graphs are equal or not, but also how similar is one graph to another. Namely, GED is the *minimal cost* of modifying one graph into the other. A cost is assigned to each possible graph modification operation – namely the addition or deletion of a vertex or an edge. The GED is the total cost of all the modification operations required to transform one graph into the other. Although there can be several costs to transform one graph into another (resulting from different sequences of modification operations which lead into the same result), the GED is the minimum total cost to achieve the

**Algorithm 5.2: Personal Place Similarity****Input:** PMG and a set of places  $P'$ **Output:** places in  $P'$  are sorted based on their GED with PMG.

- 1: For each place  $p'$  in  $P'$ , do steps 2-4
- 2: Add  $p'$  to  $PMG$
- 3: Calculate  $GED(p', PMG)$
- 4: Remove  $p'$  from  $PMG$
- 5: Sort places in  $P'$  by their relative  $GED$

transform. Note that the specific cost of each alteration operation (the addition or deletion of a vertex or an edge) is application-dependent (Gao et al., 2010).

GED can be applied to labelled or non-labelled edges or vertices graphs. An example is shown in Figure 5.5. Assume that each graph modification operation has one unit cost, then the GED of transforming the graph shown in Figure 5.5a into the graph in Figure 5.5f equals 4. In Figure 5.5b an edge is inserted between vertices 1 and 3, in Figure 5.5c vertex 4 is deleted along with its connecting edges, in Figure 5.5d vertex 6 is inserted, and finally in Figure 5.5e an edge is inserted which connects the vertices 3 and 6.

GED can be used to calculate the cost of adding a new place to a PG. This cost represents the relevance of a place to a person, based on the person's PG.

This algorithm ranks a given set of arbitrary places based on their similarity to a PMG of a person, in a mobile context. The heuristics introduced in the section are employed to design a similarity measure – a *cost* of a visit – to compare any given place to the PMG of a certain individual. For any given PMG and a place, Algorithm 5.1 is used to process a hypothetical visit to that place<sup>2</sup>. The *cost* of this visit is defined based on the notion of GED and the heuristics introduced earlier, as will be given. For a given set of places, the cost of each place is calculated, and the set is sorted accordingly. See

---

<sup>2</sup>Algorithm 5.1 is overloaded as necessary.

Algorithm 5.2.

The heuristics are utilised to define the cost of GED for PMG, as follows (each cost definition corresponds to a heuristic, in order):

1. The *temporal degree of a conceptual node*. The more temporal edges connecting a conceptual node to temporal nodes the less it costs to insert another temporal edge from that conceptual node to a new temporal node.
2. *Anchor concepts*. The more total time a concept accumulates (summing the time intervals of all the temporal nodes connected to that conceptual node) the less it costs to insert another temporal edge from that conceptual node to a new temporal node.
3. *Temporal patterns*. For all the temporal nodes connected to a certain conceptual node, the common time interval between those temporal nodes is probably the time interval at which this concept is going to be revisited, i.e. the less it costs to connect that conceptual node to a temporal node of such a time interval.
4. *Temporal closeness*. For a newly inserted temporal node, the more temporally close its conceptual node is (comparing to the last visit time it was visited) the less it costs to insert a temporal edge to connect that conceptual node to the new temporal node.
5. The *spatial degree of a conceptual node*. The more spatial edges connecting a conceptual node to spatial nodes, the less it costs to connect that conceptual node to a new spatial node. If a concept is derived from more than one place, it is a heuristic of a more interest in that concept.
6. *Spatial closeness*. For a newly inserted spatial node which its associated concept is already in the PMG, the more spatially close its conceptual node is (comparing to its spatial footprints) the less it costs to insert a spatial edge to connect that conceptual node to the new spatial node.

The GED is calculated over multiple dimensions, see Equation 5.6:

$$GED := \alpha + \beta + \chi + \delta + \epsilon + \eta \quad (5.6)$$

- $\alpha$  is the cost of the temporal edges for the given place, i.e. it is the ratio of the temporal degree of all the conceptual nodes of that place (the count of their relative temporal edges) to the temporal degree of the given PMG (the total count of its temporal edges).
- $\beta$  is the ratio of the total time of all the conceptual nodes of that place (the total time of their relative temporal nodes) to the total time of all the temporal nodes in the PMG.
- $\chi$  is the cost of the conceptual nodes for the given place, i.e. it is the ratio of the total weight of all the conceptual nodes of that place to the total weight of all the nodes in the PMG.
- $\delta$  is the ratio of the temporal difference between the current time and the last visit in the PMG (the temporal node which is most temporally-close to the current time) to the temporal difference between the current time and the last visit to the concepts related to the given place.
- $\epsilon$  is the cost of the spatial edges for the given place, i.e. it is ratio of the spatial degree of all the conceptual nodes of that place (the count of their relative spatial edges) to the spatial degree of the given PMG (the total count of its spatial edges).
- $\eta$  is the difference in spatial distance between the spatial footprint of the given place and the current position (of a person, in a mobile context).

## 5.6 Conclusion

This chapter presented different views of human mobility, such as moving between geographical places, their types or their relative concepts. Accordingly, the chapter introduced a model for PG based on graph theory which encodes spatial, temporal and thematic information. The graph building algorithm explained how each of its concepts is associated with a weight, and this weight is updated according to the relevance of a concept associated with the concepts of the currently visited places. The model utilises the notions of graph theory and several heuristics based on human mobility trends to calculate the relatedness of a geographical place to a certain individual, based on PG. The next chapter introduces a mark-up language for the model.

---

## *Chapter 6*

# **Personal gazetteers mark-up language**

## **6.1 Introduction**

In general, collecting data on a large scale, and processing collections of such data, had often led to interoperability problems, basically resulting from adapting heterogeneous data representations. Bio-medical, geo-spatial, and Web Services contents and usage data, for example, all had their integration problems, where several specifications had been developed as solutions throughout the years.

A PG can be utilized for various purposes, such as predicting a next visit for a certain individual, and as such inferring the context of the visit and personalizing any data or services available to that individual accordingly (possibly by integrating other data sets). For multiple individuals, a set of PG can help to identify statistically significant places in a neighborhood, and hence plan transportation routes, optimal places of public services, and so on.

In a mobile setting, many context dimensions other than the space and time may as well be logged for further analysis, for example physical surroundings, orientation or nearby friends (Steiniger et al., 2006; Nivala and Sarjakoski, 2003b). Moreover, a PG may be mapped to various data sets with the aim of transforming the raw spatial coordinates – of the visited places – into place semantics (Hightower, 2003), to infer further information about a certain individual. For example, mobile people may tag places with various descriptions, ratings, or activities taken in those places (Espinoza

et al., 2001; Zhou et al., 2005a; Sohn et al., 2005). The services and data used by a mobile person may also be associated with their spatio-temporal history (WeiBenberg et al., 2006). In fact, the spatio-temporal history itself may provide some clues about the semantics of the visited places in a PG, such as the home or work places (Liao et al., 2006). All that adds a thematic dimension to the spatio-temporal elements of a PG.

Based on the above, the same individual is likely to end up with more than one PG, where each PG is employed by a different application. Moreover, these multiple PG will probably share much of their contents – a scenario which happened on the Web, for stationary applications. On a planet-scale, this will lead to large amounts of duplicated, yet non-interoperable, data. Although ontologies have been proposed to solve data semantics interoperability problems in mobile environments in general (Chen et al., 2004; Heckmann and Krüger, 2003; Heckmann et al., 2005), no common model or an encoding format of particularly PG exists at present. This work proposes a lingua-franca mark-up language for PG which is aimed to solve this issue. A PMG is implemented in Personal Gazetteers Mark-up Language PGML, a markup language proposed as a common encoding format for personal gazetteers, and which is based on the PMG.

## 6.2 PGML

The PGML is an XML language for encoding the PG, based on the PMG introduced in the previous chapter. Designed to be scalable, a PGML avoids explicitly encoding any elements which may be embedded in its data, as well as any elements which may be inferred based on its data. For example, the various statistics which could be found based on the contents of a PGML file, such as the most visited place or the name and address of that place (which could be found by mapping its spatial footprint to the required data sets), are not part of a PGML file. Such a design allows the PGML to

be flexible enough to be utilized by different applications. For example, assume that a place is considered familiar to a certain person if it is visited a certain number of times. Different applications may define that threshold differently. The same argument applies to place names. A single spatial footprint can refer to a different place names, after different organizations or languages. Separating embedded and inferred elements of a PG from the PGML allows a more flexible and scalable deployment, without losing the necessary data to infer those elements on demand. The next section introduces a motivating scenario to implement the PGML.

A PGML file consists of thematic, spatial footprints, and temporal footprints nodes elements, as well as spatial and temporal edges elements. Each thematic node element has an id, a concept, and a weight elements. Each spatial footprint element has an id, latitude, and longitude elements. A temporal footprint element has an id, and “*from*” and “*to*” time-stamps elements. A spatial edge element has an id, a thematic node id, and a spatial footprint node id elements. Finally, a temporal edge element has an id, a thematic node id, and a temporal node id elements. Partial PMG for the users encoded by PGML is shown in Figures 6.1, 6.2., 6.3 and 6.4, and the PGML XML Schema is given in Figure 6.5.

### 6.3 Motivating scenario

Consider a person navigating in a typical urban setting, with two PG-based applications which run on his mobile device. A PG constructed by a Point of Interest (POI) recommendation application may log the spatial footprints of the visited places, their names, types, and perhaps their relative times of visits, with the objective of recommending similar places around, while navigating. On the other hand, that person also uses a traffic analysis application which builds a PG of the most visited places and their times of visits, and regularly synchronises the PG contents with a remote server via a wireless Internet connection. The contents of both PG are nearly the same. Encoding

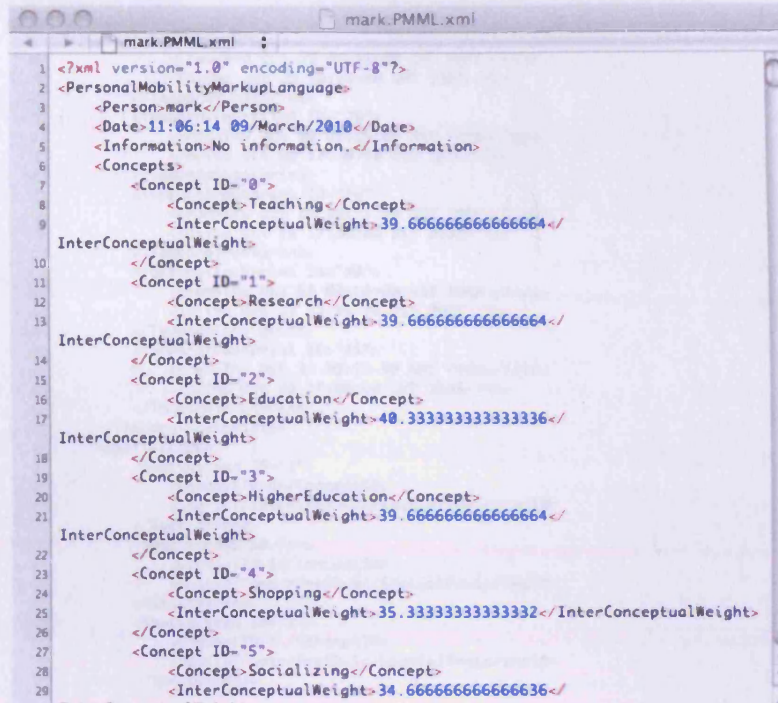


Figure 6.1: Partial XML encoding of the PMG for user 1

both PG in one PGML file will save data storage space, computational power, and the use of allocation hardware – all are necessary to optimize the performance of mobile devices, which typically have limited resources. The varying elements which are not shared between the two PG may be kept on remote applications servers or in separate application files on the mobile device.

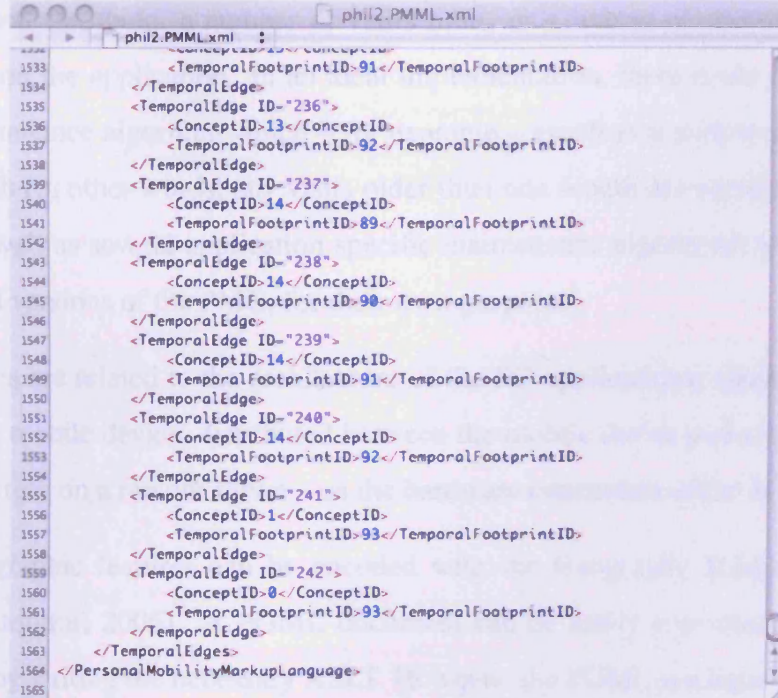
## 6.4 Discussion

The size of a PMG, and hence its relative PGML file, is important to consider, specifically for a mobile device deployment scenario. A PMG maintenance algorithm needs to be developed, to keep the relative PGML file size as small as possible. Different maintenance algorithms may employ different criteria to maintain a PMG, such



```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <PersonalMobilityMarkupLanguage>
3   <Person-mark</Person>
4   <Date>11:06:14 09/March/2010</Date>
5   <Information>No information.</Information>
6   <Concepts>
7     <Concept ID="0">
8       <Concept>Teaching</Concept>
9       <InterConceptualWeight>39.666666666666664</
10 InterConceptualWeight>
11     </Concept>
12     <Concept ID="1">
13       <Concept>Research</Concept>
14       <InterConceptualWeight>39.666666666666664</
15 InterConceptualWeight>
16     </Concept>
17     <Concept ID="2">
18       <Concept>Education</Concept>
19       <InterConceptualWeight>40.333333333333336</
20 InterConceptualWeight>
21     </Concept>
22     <Concept ID="3">
23       <Concept>HigherEducation</Concept>
24       <InterConceptualWeight>39.666666666666664</
25 InterConceptualWeight>
26     </Concept>
27     <Concept ID="4">
28       <Concept>Shopping</Concept>
29       <InterConceptualWeight>35.333333333333332</ InterConceptualWeight>
30     </Concept>
31     <Concept ID="5">
32       <Concept>Socializing</Concept>
33       <InterConceptualWeight>34.666666666666636</
```

Figure 6.2: Partial XML encoding of the PMG for user 2



```
1533 </ConceptID>
1534 </TemporalFootprintID>
1535 </TemporalEdge>
1536 <TemporalEdge ID="236">
1537   <ConceptID>13</ConceptID>
1538   <TemporalFootprintID>92</TemporalFootprintID>
1539 </TemporalEdge>
1540 <TemporalEdge ID="237">
1541   <ConceptID>14</ConceptID>
1542   <TemporalFootprintID>89</TemporalFootprintID>
1543 </TemporalEdge>
1544 <TemporalEdge ID="238">
1545   <ConceptID>14</ConceptID>
1546   <TemporalFootprintID>90</TemporalFootprintID>
1547 </TemporalEdge>
1548 <TemporalEdge ID="239">
1549   <ConceptID>14</ConceptID>
1550   <TemporalFootprintID>91</TemporalFootprintID>
1551 </TemporalEdge>
1552 <TemporalEdge ID="240">
1553   <ConceptID>14</ConceptID>
1554   <TemporalFootprintID>92</TemporalFootprintID>
1555 </TemporalEdge>
1556 <TemporalEdge ID="241">
1557   <ConceptID>1</ConceptID>
1558   <TemporalFootprintID>93</TemporalFootprintID>
1559 </TemporalEdge>
1560 <TemporalEdge ID="242">
1561   <ConceptID>0</ConceptID>
1562   <TemporalFootprintID>93</TemporalFootprintID>
1563 </TemporalEdge>
1564 </TemporalEdges>
1565 </PersonalMobilityMarkupLanguage>
```

Figure 6.3: Partial XML encoding of the PMG for user 3

```

florian2.PMML.xml
florian2.PMML.xml
448 <From>Wed Oct 29 09:15:00 GMT 2008</From>
449 <To>Wed Oct 29 15:15:00 GMT 2008</To>
450 </TemporalFootprint>
451 <TemporalFootprint ID="78">
452 <From>Thu Oct 09 09:15:00 BST 2008</From>
453 <To>Thu Oct 09 17:30:00 BST 2008</To>
454 </TemporalFootprint>
455 <TemporalFootprint ID="79">
456 <From>Thu Oct 16 09:15:00 BST 2008</From>
457 <To>Thu Oct 16 17:30:00 BST 2008</To>
458 </TemporalFootprint>
459 <TemporalFootprint ID="80">
460 <From>Thu Oct 23 09:15:00 BST 2008</From>
461 <To>Thu Oct 23 17:30:00 BST 2008</To>
462 </TemporalFootprint>
463 <TemporalFootprint ID="81">
464 <From>Thu Oct 30 09:15:00 GMT 2008</From>
465 <To>Thu Oct 30 17:30:00 GMT 2008</To>
466 </TemporalFootprint>
467 </TemporalFootprints>
468 <SpatialEdges>
469 <SpatialEdge ID="0">
470 <ConceptID>0</ConceptID>
471 <SpatialFootprintID>0</SpatialFootprintID>
472 </SpatialEdge>
473 <SpatialEdge ID="1">
474 <ConceptID>1</ConceptID>
475 <SpatialFootprintID>0</SpatialFootprintID>
476 </SpatialEdge>
477 <SpatialEdge ID="2">
478 <ConceptID>2</ConceptID>
479 <SpatialFootprintID>1</SpatialFootprintID>
480 </SpatialEdge>

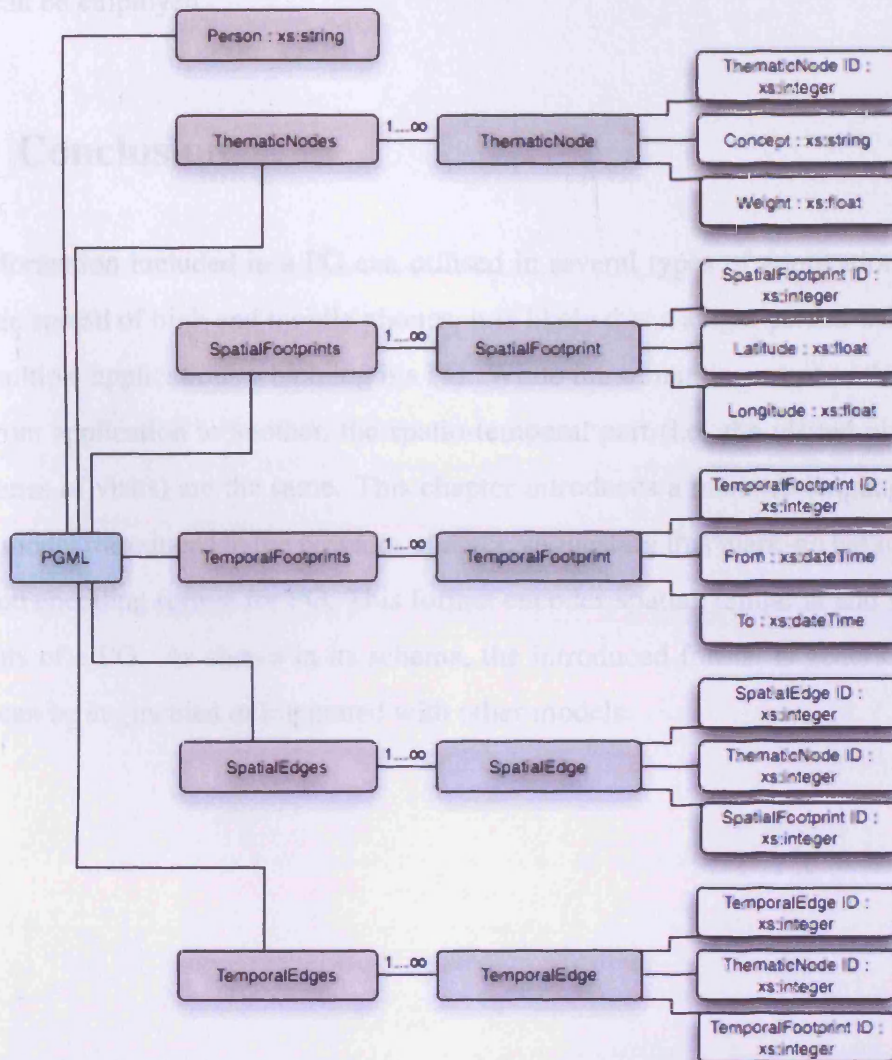
```

**Figure 6.4: Partial XML encoding of the PMG for user 4**

as a temporal threshold, a number of visits limit, or a hybrid of several measures – depending on the application. In an ideal implementation, there could be one global PMG maintenance algorithm which – for example – employs a temporal threshold of a one month (in other words, any visits older than one month are eliminated from the PMG), as well as several application specific maintenance algorithms which produce customized versions of the PMG, for their own purposes.

These issues are related to the architecture of the PG applications; whether the PG is kept on the mobile device, distributed between the mobile device and a remote server, or kept entirely on a remote server – as the hardware constraints differ in each case.

Many geographic features can be encoded with the Geography Markup Language (GML) (Burggraf, 2006). A PGML document can be easily converted into a GML document by writing the necessary XSLT. However, the PGML is a light-weight XML encoding for PG which preserves the PMG structure.



**Figure 6.5: PGML XML schema diagram**

As mentioned, as PGML is based on a graph-theoretical model – the PMG, the mathematics of graph theory can be utilized to perform several analytical processes, for a set of PGML. Namely, two PMG can be compared or a set of PMG can be sorted based on the similarities of their relative graphs. Such flexibility facilitates the employment of PGML not only for any individual, but for a group of people as well. As with the maintenance of a PGML file, it is possible that several measures to compare a set of

PMG can be employed.

## **6.5 Conclusion**

The information included in a PG can be utilised in several types of applications. With the wide spread of high end mobile phones, it is likely that a single person will end up with multiple applications which log his PG. While the semantic part of such PG may vary from application to another, the spatio-temporal part (i.e. the visited places and their times of visits) are the same. This chapter introduces a mark-up language based on the model introduced in the previous chapter, suggesting this mark-up language as a common encoding format for PG. This format encodes spatial, temporal and semantic elements of a PG. As shown in its schema, the introduced format is generic enough that it can be augmented or integrated with other models.

# Chapter 7

## Profiles evaluation

This chapter give the results of a pilot and a group profiles modelled as PMG, in subsections 7.1.1 and 7.1.2. In section 7.2, an evaluation is given for the PG constructing approach in subsection 7.2.1, the PMG model design in subsection 7.2.2, and the resulting profiles in subsections 7.2.3 and 7.2.4 respectively. The chapter ends with a conclusion in section 7.3.

### 7.1 Results

#### 7.1.1 Pilot profile

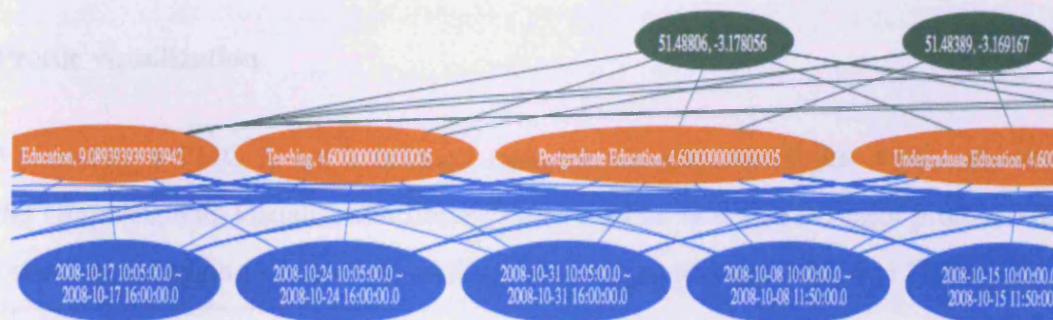


Figure 7.1: Partial visualization of the pilot profile.

## Data

A person who lives in a typical modern city was asked to log her spatio-temporal history for a period of one month, filling out a paper form. The data set had two parts; namely the weekly-routine visits and the irregular visits (visits to places outside the weekly-routine). For each visit, the place name, its type, and the beginning and the ending time-stamps of each visit to that place was recorded.

Her log contained 148 visits to 13 distinct places, which are of 10 place types, and all of which are within the city. The data set was fed into an Oracle database.

## Profile

A Java program was written to implement Algorithms 5.1 and 5.2, to construct the resulting profile. The Jena Semantic Web Framework<sup>1</sup> was used to retrieve the ontology.

As the log did not contain the actual spatial footprints of the places visited, a manual geocoding was conducted to augment the database with the respective spatial footprints (based on the place names as well as the approximate positions of the visited places, as given by the data provider).

## Profile visualization

Visualizing a PMG is not a primary task for the purpose of this work. However, the GraphViz data visualisation library<sup>2</sup> was used to visualize the pilot profile, which helped in debugging some of the coding flaws of generating PMG. See Figure 7.1.

---

<sup>1</sup><http://jena.sourceforge.net>.

<sup>2</sup><http://www.graphviz.org>.

**Table 7.1: Visited places types of collected GPS logs**

Person	Visited place types
user 1	university, shop, restaurant, library, home, bank.
user 2	university, theatre, supermarket, station, sport centre, spa, ski resort, shop, school, restaurant, parking place, library, house, home, community centre, club, church, cafe, bar, bank, airport.
user 3	university, supermarket, service station, rugby club, restaurant, home, bar.
user 4	university, supermarket, primary school, leisure centre, home, community centre, church, beach, allotment.

### 7.1.2 Group profiles

The following subsection given the results of an experiment for extracting a set of PMG from real-world GPS data logs, employing the methods introduced in this thesis. Namely, the raw GPS logs were pre-processed (as in Chapter 3), enriched with place affordance concepts (as in Chapter 4) and encoded as the graph model and mark-up language introduced in Chapters 5 and 6, respectively.

#### Data

Four individuals were asked to collect their GPS mobility traces for a minimum period of one month, and up to 2 months, in a typical modern city. The group used Garmin GPS-60 devices<sup>3</sup> to collect their data, along with a manual paper log of the visited places' names, types, and relative times of visits.

<sup>3</sup><http://www.garmin.com>.



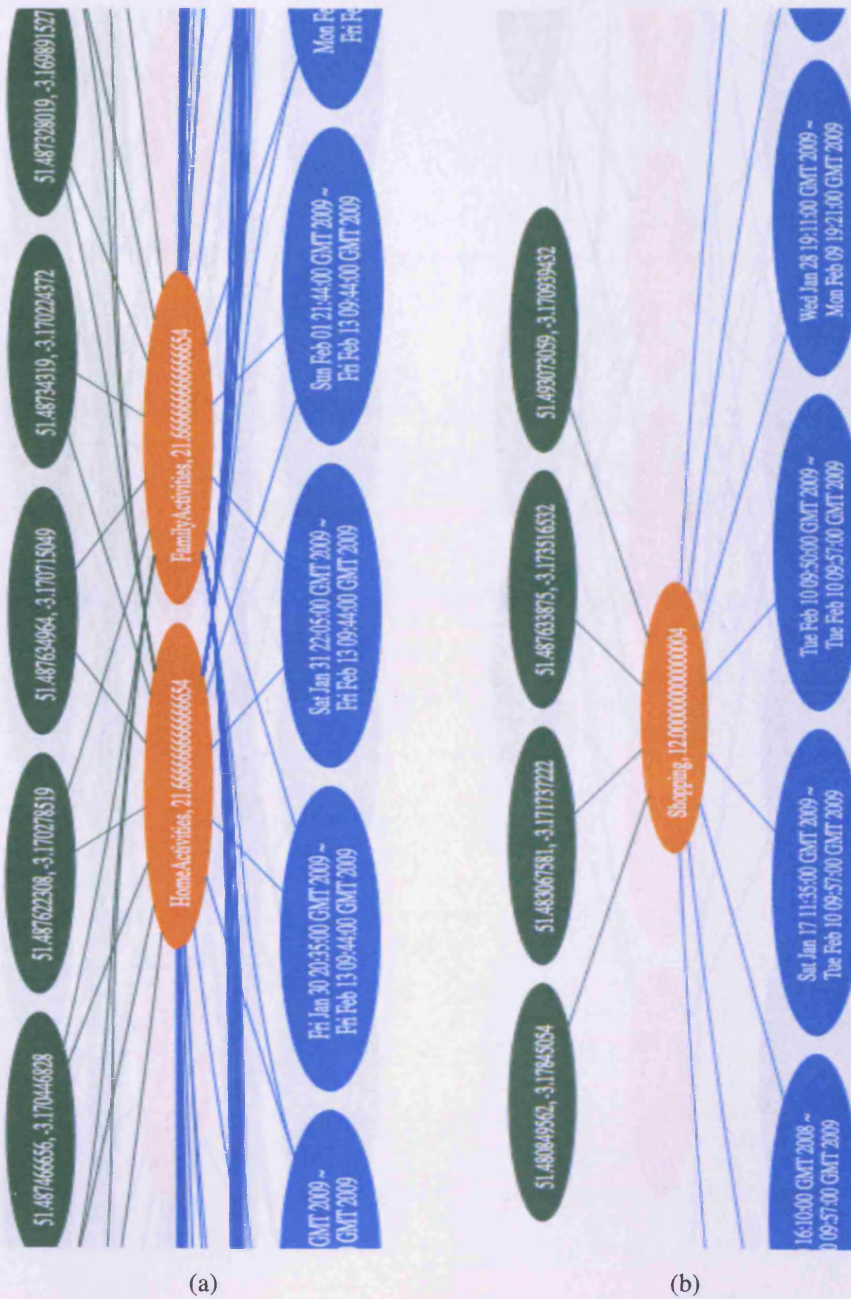


Figure 7.2: Partial visualization of the PMG for user 1

Profiles

The four people who provided their GPS logs will be referred to as *user 1*, *2*, *3*, and *4* respectively. Table 7.1 enlists the place types visited by each of the data providers.

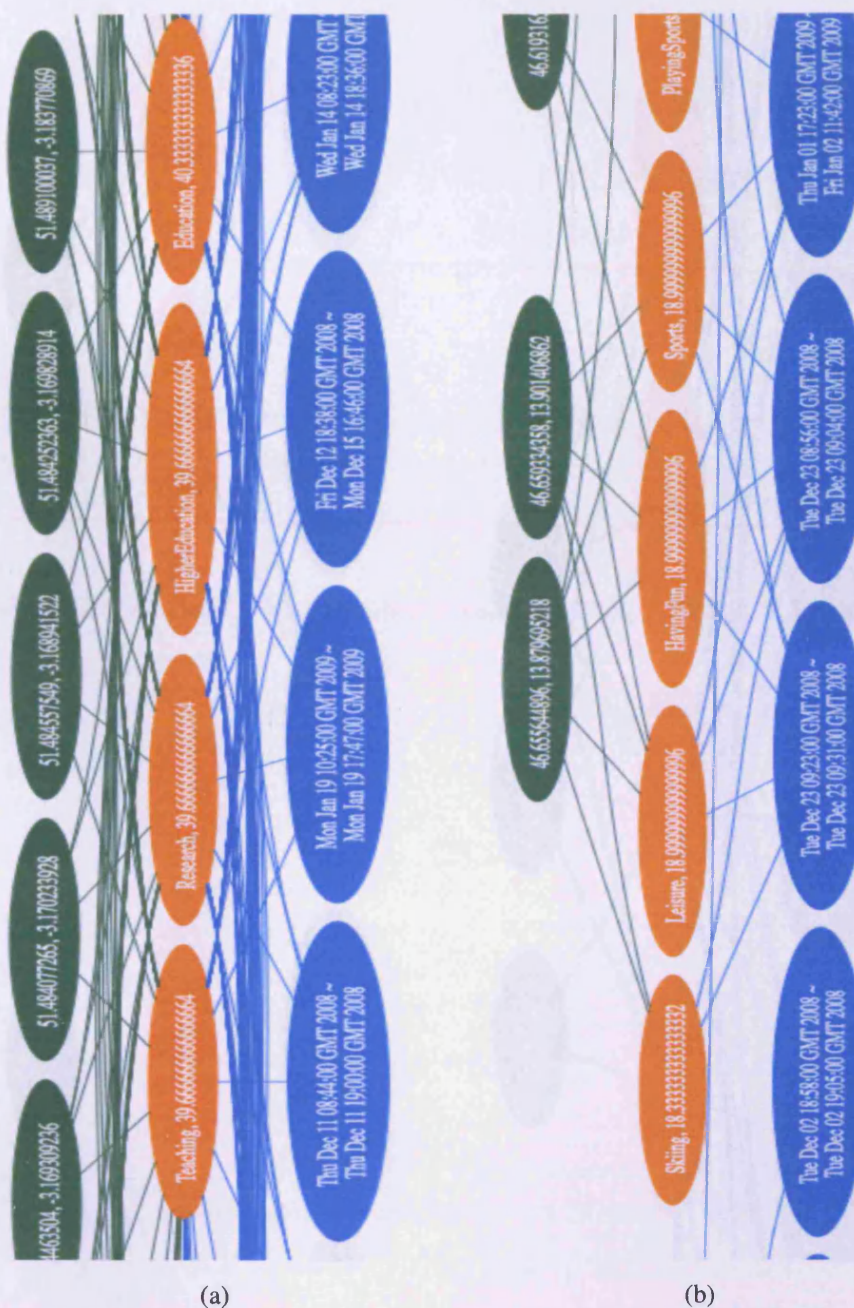


Figure 7.3: Partial visualization of the PMG for user 2

Figures 7.2, 7.3, 7.4 and 7.5 show a partial visualisation of the PMG for users 1, 2, 3 and user 4.

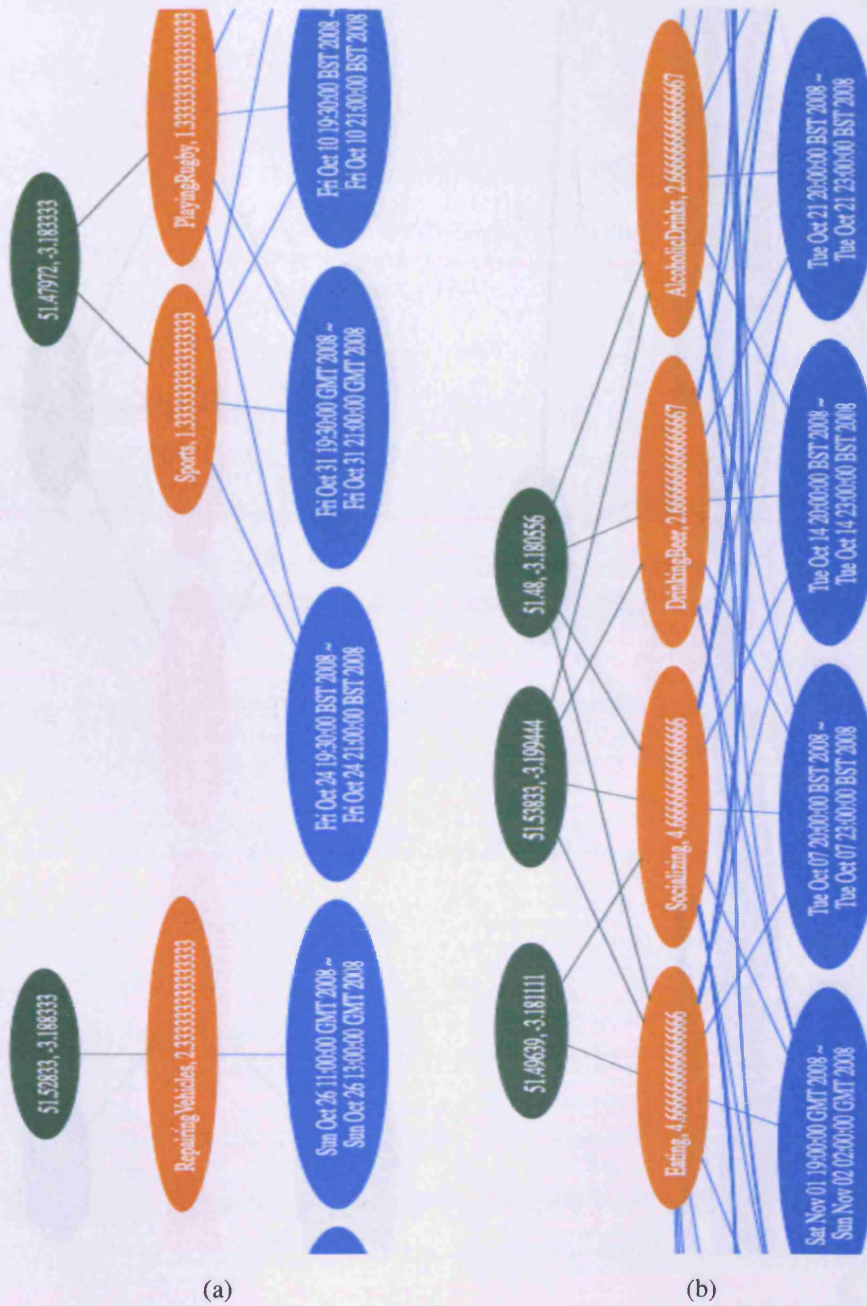


Figure 7.4: Partial visualization of the PMG for user 3

## 7.2 Evaluation

### 7.2.1 Approach

Relying on a place-type allowance that vector to achieve place semantics – via the place type – certainly characterise the need to land-park place-specific services. In-

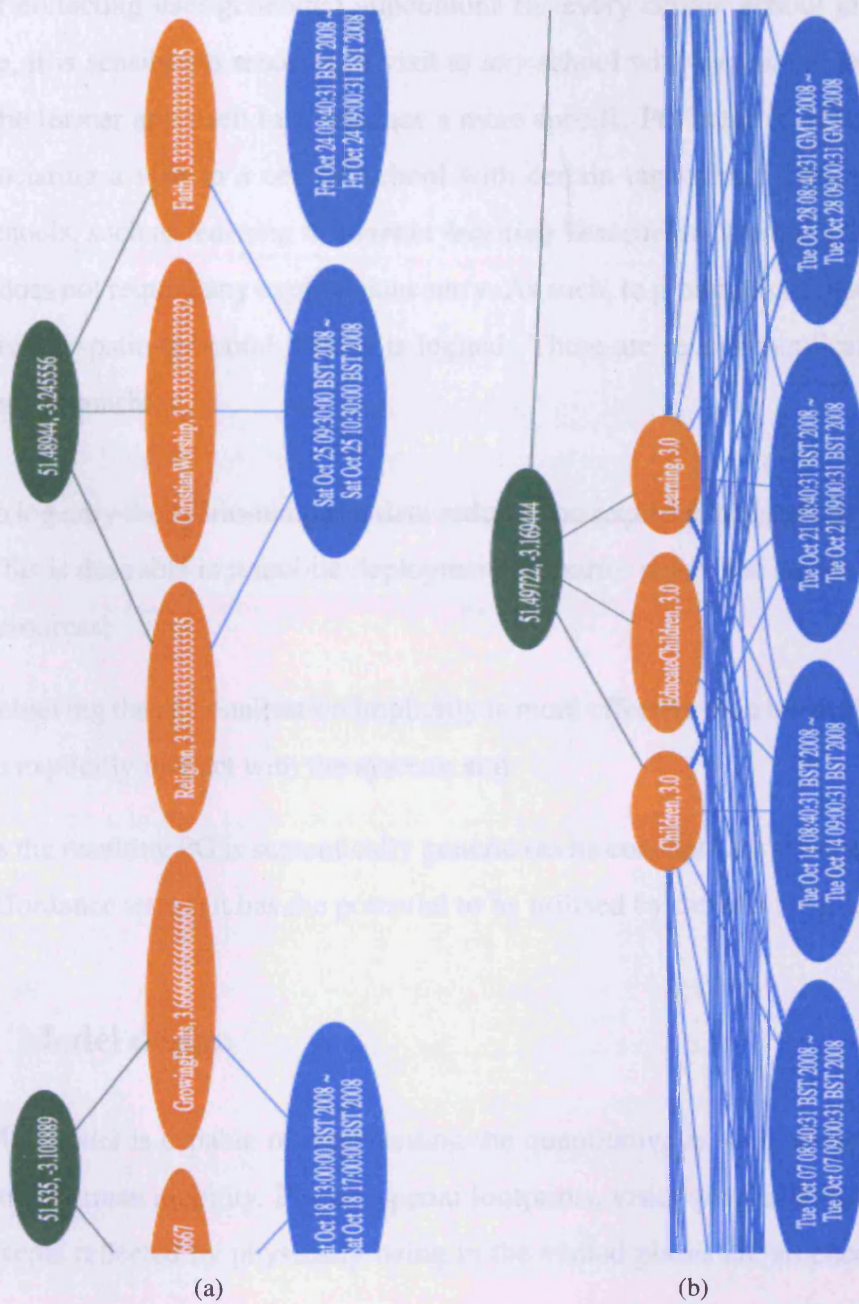


Figure 7.5: Partial visualization of the PMG for user 4

## 7.2 Evaluation

### 7.2.1 Approach

Relying on a place-type affordance data source to retrieve place semantics – via the place type – certainly eliminates the need to hard-code place-specific semantics. In-

stead of collecting user-generated annotations for every certain school in a city, for example, it is sensible to associate a visit to *any* school with *education* and *learning*. While the former approach may produce a more specific PG in terms of its semantics (by associating a visit to a certain school with certain tags which may not apply to other schools, such as *learning to swim* or *learning Venezuelan-Spanish*), the latter approach does not require any explicit data entry. As such, to produce a semantically-rich PG, only the spatio-temporal history is logged. There are several implications of the proposed approach:

- to log only the spatio-temporal data reduces the required data storage for the PG. This is desirable in a mobile deployment scenario, where the device has limited resources;
- achieving the personalisation implicitly is more effective than to require the users to explicitly interact with the system; and
- as the resulting PG is semantically generic (as its concepts are generic place-type affordance terms) it has the potential to be utilised by different applications.

### 7.2.2 Model design

The PMG model is capable of representing the quantitative as well as the qualitative elements of human mobility. Places' spatial footprints, visits' temporal footprints, and the concepts reflected by physically being in the visited places are all encoded in the model.

Information retrieval beyond the scope of the concepts encoded in a PMG is possible by linking the spatial footprints to their relative data sets, of any kind. For example, the spatial footprint of a visited university may be employed to derive its name, current events, courses taught, and so on.

The model is based on a sound mathematical theory (graph theory), which allows to utilise various mathematics of graph theory to analyse a PMG. Specifically, the notion of GED is employed to measure the relevance of a place to a PMG. Moreover, various statistics such as the set of five most common places to a certain person or the most visited place (in terms of the number of visits, the total time of visits, or both) are all measurable. By simply counting the spatial or temporal edges, the number of places which reflect a certain concept or the number of times a certain concept was visited by a certain person, all can be found. The common mobility spatial zone for a certain person can be found by retrieving the spatial footprints of the places in a PMG. This is a spatial area which probably includes home, work place, and regularly visited shops and social places. Accordingly, other irregular places or events maybe detected, such as travelling outside a city.

The model is capable of differentiating multiple concepts of the same geographical place, for different people (see step 6 of Algorithm 5.1). Assume that two individuals visited a place which is associated with *medicine* and *research*, and as such both of those concepts are encoded in their PMG. Later on, as the first individual kept visiting other places which are related to *research*, the concept *research* accumulated more weight than *medicine* in that person's PMG. The second individual may have the opposite: visiting other places which are more into *medicine* will increment the weight of *medicine* in her PMG. Subsequently, the two concepts retrieved from visiting the same place ended up to be weighted differently in different PMG, depending on the other places visited.

### 7.2.3 Pilot profile

The pilot PMG was processed to produce different lists of work-related, weekend-related, and general concepts associated with the user. Simple spatio-temporal heuristics were used to cluster the PMG into those lists. Namely, work related concepts were defined by a temporal-window of the typical working hours (09:00-17:00) during the

typical working days (Monday to Friday). The weekend is defined by its typical days (Saturday and Sunday), and any concept which does not belong to either of those lists is considered a general interest.

The person was asked to provide feedback for each concept of each list, as either i) "1" for "not related at all or not applicable" ii) "2" for "not related to the context of list" iii) "3" for "somewhat related" iv) "4" for "related" or; v) "5" for "very related" . The resulting lists and their relative feedback are shown in Table 7.2.

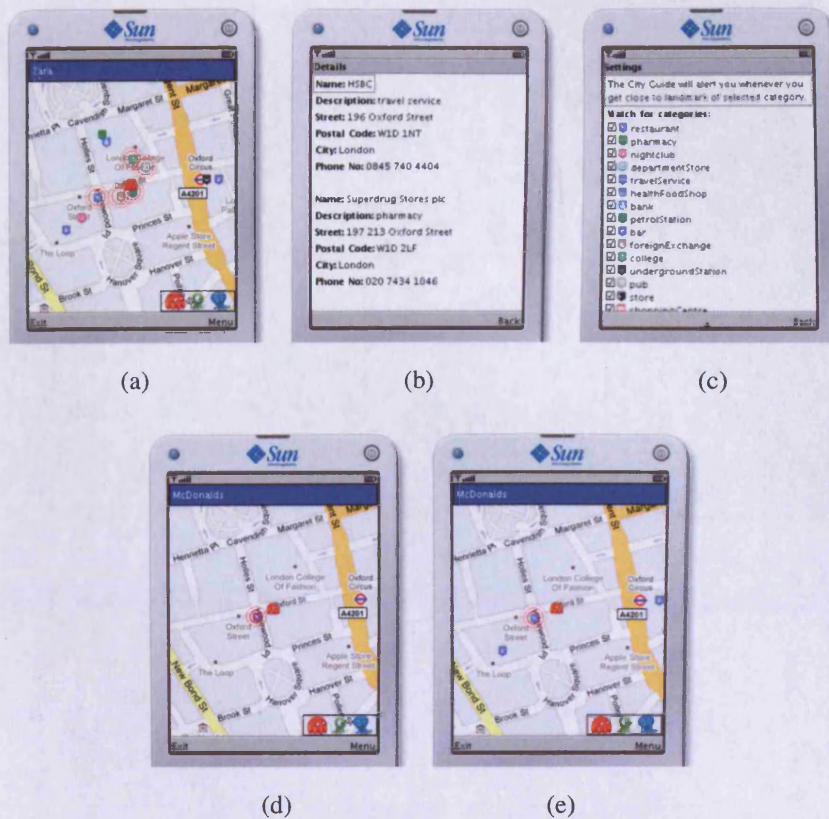
The results show that 65% of the concepts were rated between *somewhat related* and *very related*. The results show that all the work-related concepts were perfectly inferred. The weekend-related list had two *very related* concepts (*home activities* and *shopping*) while the other two were described as *somewhat related* and *not related* (*education* and *secondary education*, respectively). Those concepts were inferred from short visits to a secondary school during Saturday mornings. The person explained that she made those visits to deliver her children to a private secondary school to learn a foreign language. While *somewhat related* for *education* is an accepted result, *not related* for *secondary education* shows a limitation. Namely, some places are used for purposes other than – or as a specialization of – the typical services associated with their place-types. In this case, a private secondary school provided lessons of a foreign language outside the normal working hours – not part of a typical *secondary education*.

The general interests list show a range of feedback; while some concepts had good or accepted feedback, others had poor feedback. Discussing the results with the data provider, a conclusion is made that the quality of the ontology employed to derive the profile – in particular the semantic granularity of its concepts – can significantly affect the results. *perusing activity* seemed to be quite uninformative, while *playing* would have been rated higher if it was, for example, *playing sports*. The same argument applies to *studying of music*, which would have been rated lower if it was only *studying*.

**Table 7.2: Pilot profile concepts and their relative feedback**

Concept	feedback	
education	5	work-related
higher education	5	
research	5	
teaching	5	
postgraduate education	5	
undergraduate education	5	
education	3	weekend-related
home activities	5	
shopping	5	
secondary education	1	
education	2	general interests
secondary education	1	
swimming	5	
meeting	1	
water-based recreation	5	
training	1	
studying of music	5	
free education	2	
keeping fit	3	
sport	3	
relaxing	5	
playing	1	
pursuing activity	1	





**Figure 7.6: Simulated mobile navigation guide showcasing the utilisation of PMG to adapt the selection of POI on the map.**

### 7.2.4 Group profiles

As mentioned, mobile devices have limited resources in terms of their communication bandwidth, data storage, and computational resources. As such, mobile applications need to utilise their available resources to optimise their performance. Mobile navigation guides (Nivala and Sarjakoski, 2003b; Kray and Baus, 2003; Krüger et al., 2007) employ different techniques to personalise their service to their users.

In this experiment, an emulated mobile navigation guide utilises each of the four PMG produced to filter-out unrelated Points of Interest (POI) on the map based on algorithm 5.2. The resulting maps are given for evaluation to each the users.

**Table 7.3: Finding relevant POI in Oxford Street based on PMG.**

Place	Type	user 1	user 2	user 3	user 4
		Shown on map			
McDonalds	Restaurant	Y	Y	Y	Y
Boots	Pharmacy	N	Y	N	N
Molton House	Nightclub	N	Y	N	Y
Debenhams	Department Store	N	Y	N	N
Jane Asher	Restaurant	Y	N	N	Y
HSBC	Travel Service	N	N	N	N
Excellent Health	Health Food Shop	N	N	N	N
Abbey	Bank	N	N	N	N
The Argyll Arms	Bar	Y	N	Y	Y
Total	Petrol Station	N	N	N	N
The Loop	Bar	N	N	N	N
Jalouse	Nightclub	N	N	N	N
Boots	Pharmacy	N	N	N	N
Zara	Department Store	N	N	N	N
Oxford Circus	Underground Railway Station	N	N	N	N
London College Of Fashion	College	N	Y	Y	N
The Royal Bank Of Scotland	Bank	N	N	N	N
Fortnum & Mason	Department Store	N	N	N	N
100 Club Ltd	Music Club	N	N	N	N
Uniqlo Oxford Street	Shopping Centre	N	N	N	N
Boots	Pharmacy	N	N	N	N
House Of Fraser	Department Store	N	N	N	N
TTT Foreign Exchange Corporation	Foreign Exchange	N	N	N	N
Muji	Store	Y	N	Y	N
Liberty	Department Store	N	N	N	N
Superdrug Stores plc	Pharmacy	N	N	N	N
HSBC	Travel Service	N	N	N	N

The prototype<sup>4</sup> mimicked a hypothetical walk taken by the user at Oxford Street, London, see Figure 7.6. It alerts the user whenever a nearby POI is available, see Figure 7.6a. The user can optionally view further details about that POI, see Figure 7.6b. The user can optionally select which type of POI types to be alerted for, see Figure 7.6c. For example, in Figure 7.6d, only the nearby restaurants are shown on the map. Considering that there are over 600 place types to select from<sup>5</sup>, manual selection of POI types is not convenient. It would require a considerable effort from the user to choose a selected set of POI, as well as to maintain that selection (users preferences may change with time). Another option is to show only the previously visited places. In this case, the limitations include missing any related POI which maybe of interest but haven't been visited before. Moreover, in a visiting a new town scenario, all the available POI will be missed.

Instead, the available POI on the map are filtered-out based on their relevance to the PMG for each user, see Algorithm 5.2. To simplify the processing, only three dimensions were considered to calculate the GED, namely:  $\alpha$ ,  $\chi$ , and  $\epsilon$  (see Equation 5.6).

Each user was given two cases to evaluate. The first case shows all the available POI on the map while the user is walking along the street, and the second one shows only the most related 4 POI to that person (based on his or her PMG). For clarity, each user was given:

- a list of the previously visited places and their respective types and times of visits; and
- a list of all the available POI on the map with their respective place-types. See Table 7.3.

Each user was asked two questions:

---

<sup>4</sup>Based on Sun's NetBeans CityGuide sample application, see <http://www.netbeans.org>.

<sup>5</sup><http://www.ordnancesurvey.co.uk/oswebsite/products/pointsofinterest/classifications.html>

1. whether the POI shown are related to the places that person typically visits – and as such to the person; and
2. what other POI which the user thinks are related to him – based on his PM log, i.e. previously visited places – but were not shown on the map.

The users were asked to observe the simulation before and after applying the personalisation, and to give their answers considering the lists provided. The answers were as follows:

- all the users agreed that the POI selected on their relative maps are related to the places they typically visit;
- the approach is more convenient than displaying alerts for all the available POI – in particular as it requires no explicit input from the user (such as selecting which POI type to show on the map); and
- some users mentioned that some of the filtered-out POI are actually related to places which they visit but weren't in their GPS logs.

Place of financial services (banks, money exchange offices, and so on) and pharmacies were successfully filtered-out as these are not related to the visited places in the profiles. Although user 1 and user 3 visited banks before, their association with those places is weak (judging by their spatio-temporal log). Achieving an acceptable degree of personalisation of POI from a spatio-temporal log only demonstrates the feasibility of the proposed methodology. A set of POI was ranked based on its relevance to a certain individual employing an integrated similarity measure of space, time, and concepts of interest. The limitations of the approach will be given in the next chapter.

Having more users will not alter the results, as the PG is extracted for each user regardless of the other users. For any user, the visited places will be extracted from the pre-processed GPS logs, and then associated with their relevant concepts of place affordance to be build semantically rich profile based on the visited places, according to

the model, algorithms and methods introduces in this thesis. In the next chapter, some suggestions are given for future work considering the relationships between multiple users by comparing their PG.

### **7.3 Conclusion**

This chapter presented the results of extracting semantically rich PG according to the methodology introduced in this thesis. The extracted PG from the collected GPS logs were semantically enriched with the concepts of place affordance extracted earlier, and the resulting semantically rich PG were modelled as a graph, based on the model introduced in this thesis. The resulting PG were shown to be useful in a simulated map navigation prototype, as it helped to filter-out irrelevant places on the map.

The next chapter concludes the thesis with a review of the thesis contributions, limitations of the work, and the suggested directions for future work.

## Conclusions

Extracting semantically rich profiles from human mobility data traces is a challenge that naturally emerges, as we – as humans – make sense of spatial data in terms of our everyday understanding of the places these data represent, rather than in their numeric format (i.e. sets of geographic coordinates). Our approach to derive such profiles relies on the common sense understanding of the typical human activities and services provided at certain place types – or place affordance. We handle various problematic issues with personal mobility traces, collected by GPS devices, by simple pragmatic pre-processing procedures. These procedures were shown to improve the quality of the extracted places, in terms of automatically discovering PG. A framework was introduced for extracting place affordance that is based on mining linguistic patterns of affordance. The resulting patterns were applied to multiple data sources to extract place affordance terms, which in our case consist of multiple-words forms rather than simple verbs. These terms were overlapped with the spatio-temporal elements in a graph model of a profile, which allows to employ notions of graph theory as a measure of relatedness of a place to a certain person, for personalisation purposes. We introduce a new mark-up language based on the profile model as a general encoding format for profiles of personal mobility.

In section 8.1, we discuss our contributions in terms of our approach to extract semantically rich mobility profiles. In section 8.2, we discuss the problems involved in extracting such profiles and our approach to solve those problems, as well as the limitations we were faced with along the course of this work. Finally, in section 8.3, we

elaborate on several suggestions of how to extend our work in future research.

## 8.1 Contributions

- I. The increasing prevalence of personal and spatially aware devices, and the resulting mobility logs make it possible to create PG that store places that are relevant to the individual device owner. These PG are mined from the mobility logs, which are in most cases generated from the GPS receiver in the device. The classical heuristic to detect places is based on the assumption that places are located where the GPS signal is lost. In heavily built-up areas, especially in inner-city urban canyons, a GPS signal is frequently lost even though the mobile person had not entered a place. This leads to a large number of incorrectly identified places. We investigated data quality of personal mobility logs collected by GPS devices and found several problems. These problems affected the results of automatically extracting PG. To overcome those problems, a set of simple procedures were introduced to pre-process the GPS data streams. These procedures proved to improve the results of analysing GPS data streams, for the purpose of automatically extracting PG. The results also affected the visualisation of the mobility traces, and inherently any applications which are built employing PG.
- II. An approach for extracting semantically rich profiles from solely the spatio-temporal history of a person was introduced. It was shown that this approach is applicable and feasible. While other approaches of extracting such profiles rely on other data sources, such as the Web resources related to a certain place, our approach is based on the association of a geographic place with its generic functions, from a human understanding perspective. This approach allows to associate a visit to a certain place with the typical activities and services provided at that place, which adds the semantic enrichment to the mere spatio-temporal co-ordinates which represent the visit to that place.

- 
- III.** Large part of this work involved the study of various models and notions of representing geographical places on the Semantic Web (SW). In particular, we were interested to investigate the possible natural language analysis techniques which allow to enrich the current representations of place with the notions of place affordance, human activities, and the services provided at certain place types – from a casual, common sense, perspective. Our approach is based on mining multiple data sources of functional descriptions of geographic place types, beginning with a typology of place types. A semantically rich resource of place types service descriptions is used to identify the frequent linguistic patterns of place affordance. These patterns are then applied to several other data sources – including ones from the Web – to extract place type specific services and concepts. A user-based evaluation experiment was conducted to evaluate the quality of the identified concepts. In addition, service concepts defined in an ontology created by domain experts were also used for comparison. In both cases, a high degree of relevance of the automatically derived concepts was witnessed. The results of the experiments are encouraging and show the potential value of the approach to support the development of semantically rich place ontologies.
- IV.** This work introduced a graph-theoretical model to represent quantitative as well as qualitative elements of human mobility. The model is based on graph theory, where multiple types of vertices and edges were proposed to represent spatial, temporal, and concepts elements related to a person. The model allows several statistical measure to be found from its content, as well as it allows to be queried for its concepts.
- V.** A personal similarity measure of a person and a geographical place was introduced. This measure includes integrated dimensions of space, time, and concepts. The measure is based on several heuristics from derived from common sense understanding of personal mobility in typical urban areas.
- VI.** This work presented a suggested model and an XML-based language specification



for a common representation of PG. The aim is to avoid a foreseen interoperability problem between different PG-based applications. The PGML introduced is based on a graph-theoretical model which is designed as light as possible to allow a scalable deployment, yet allows for various analytical processes to be performed.

## 8.2 Discussion

During the course of this work, several theoretical as well as technical issues were encountered. The following is a summary.

- I. The proposed GPS streams pre-processing heuristics are evaluated against four weeks test data-sets. The results show that employing the developed pre-processing techniques nearly halves the number of the incorrectly identified places. Further analysis of the results also reveals that a large proportion of the remaining incorrect places is caused by premature signal losses that cannot be corrected by pre-processing. While the quality of the GPS data streams improved, there remain some errors the effects of which cannot be overcome by simply pre-processing the data. How the remaining problems are critical to the end application depends on the type of the application, namely how accurate are the spatio-temporal coordinates need to be for that application purposes.
- II. Our approach produces a generic profile of an individual. This is of use in particular when the profile is intended to be shared between multiple personalisation applications. It showed a nice utilisation in terms of a personalised navigation application, but it may fall short for other types of applications.
- III. During the course of work on extracting place affordance, the model we developed to represent a geographical place included the roles of a *service provider* and a *service consumer*. Actual implementation to automatically extract these notions

from natural language text, however, showed a problem. To our current understanding, it is not achievable to distinguish between these notions by natural language processing techniques. Namely, a *teacher* and a *learner* are both in the linguistic form of a *doer*. Further exploration is needed to solve this problem. One possible direction for the solution may rely on adding additional data sources to the learning process, which in turn can improve the identified patterns or even suggest new ones.

As for the patterns, most of these were found to provide satisfactory results, except for pattern 8. The results of this pattern were found less promising, and as such it is applied only if no other pattern can be applied. A more effective POST tool would be expected to lead to more accurate results.

There were some negative results, some of which were already included in the results shown earlier. For some place types, no service types were extracted. This is caused by either having no matching pattern, or the fact that the place type description is poor, containing not enough data to process. An example of the former is a *Government Office*, which is defined as *buildings occupied by government officials to carry out their duties. see building.*, where there is no matching pattern. An example of the latter is a *Steading*, with its definition as *see farm*, being not informative for our purposes.

Another issue is the ranking of the extracted service types which are associated with a certain place type. While we decided to leave that for the application which employs the place affordance data source, other design decision may be to specify the data to a certain application domain, in other words to produce biased copies of place affordance toward to satisfy the need of certain application domains.

- IV. The maintenance of a PMG is an issue which needs to be addressed for a real-world application deployment. While we suggest several maintenance measures, such as a temporal threshold; a certain time period before which all the visits

are deleted from the profile, or a number of visits to a place based on which the profile is maintained, it is still an issue which is to be decided based on application requirements.

- V. The PGML is light, easily interoperable, and can be converted into, and from, other encoding mark-up languages for geographical features or user models, such as Geographic Mark-up Language (GML) and User Model Mark-up Language (UMML).

### 8.3 Future work

There are several research directions to extend the work presented in this thesis. Namely the following.

- I. Personal GPS devices allow to set the frequency of logging the spatio-temporal coordinates on the device. Future work is planned to evaluate whether increasing the frequency of the logging affects the quality of the process of automatically extracting PG, in particular if it will reduce the number of incorrectly identified places as well as how would it affect the pre-processing of the GPS data streams. Moreover, in our work the spatial scale of mobility has been restricted to that of a single city. In future work, it can be investigated whether places detected at different spatial scales can be employed to improve the overall quality of the results.

Another possible future work in terms of improving the quality of GPS data streams is to include the optimisation of the pre-processing procedures introduced in this work to work on mobile devices. The aim is to have a “cleaner” GPS data on the mobile, enabling the provision of clean GPS data to support LBS applications.

**II.** The interoperability of the resulting profiles from our approach between multiple personalised systems can be investigated. In other words, it would be interesting to measure how feasible are the produced profiles in different types of personalised applications such as personalised Web search, LBS such as location based advertising, and so on. Having the place affordance as the only source of semantics in the profile may be sufficient in terms of geographical place-based personalised systems, but it is currently unknown whether it will be satisfactory for other types of applications or not.

**III.** There are several directions to extend the work presented in this thesis on the automatic extraction of place affordance from textual data sources. In one of the carried out experiments, the results of expanding the place keywords list with semantically equivalent notions of the word *place* was investigated. For example, the use of the keyword *location* in the description of an Observatory in Wikipedia is “a location used for observing terrestrial and celestial events”, can lead to the identification of service concepts *ObservingTerrestrialEvents* and *ObservingCelestialEvents*, which are otherwise missed in the current system.

The temporal dimension of place affordance is not explored in the current work. One possible extension is the extraction of the temporal boundaries of place affordance. For example, it would be beneficial if the approach could not only associate a *school* with *teaching* and *learning*, for example, but to extend this association to include that these service types are offered during the typical working hours and days, of a school term time. This would, for example, improve the results in our profiling approach by avoiding associating a visit to place with its service types, if the visit was out of the service hours.

One possible direction of work is to gauge the value of a place affordance enriched place ontology in terms of Geographic Information Retrieval (GIR) precision and recall results. Query expansion techniques can be utilised to employ such a data source to answer queries such as *where can I change my watch bat-*

*teries in Cardiff?*. In particular, such a scenario is foreseen to show its full capabilities in a spatial Web search engine implementation. It would allow people to search and advertise geographical places in terms of their everyday life usage – rather than spatial co-ordinates or even place names. Current state of the art can not answer such a query unless the textual description of the Web resource representing the place, such as a Webpage, contains the words *watch*, *batteries*, and *Cardiff*, for example. Note that in our approach one would not need these keywords, as the spatial co-ordinates of a place can be associated with its place type, and as such its affordance.

- IV. It is interesting to investigate further notions of graph theory, and study their applicability for the model introduced in this work. Namely, the notion of independent set of a graph, the set of vertices which do not directly connect to all the other vertices in the graph. Inspired by this notion, it would be interesting to find the independent spatial, temporal, and conceptual sets of nodes. For example, we would propose the set of independent spatial nodes as those nodes in a PMG which do not share any temporal or conceptual nodes. These represent places which are conceptually isolated from the rest of the graph, for example all the places which are related to sports. Whether this clustering is useful or not, only the application of it will answer.

Another possible direction of work is to study the feasibility of matching two PMG as a measure of similarity between two individuals, in a social networking context. Given a set of PMG representing different people, it is possible to find the relatedness between those people based on the similarity of their PMG, i.e. based on the similarity of the places visited. In such a scenario, the notion of Graph Edit Distance (GED) is possibly useful to measure the similarity between a set of PMG.

- V. The heuristics introduced in this work for measuring personal place similarity are integrated with a summation function of equal weight each. More experiments

on real-world data and on a larger scale of mobility area and number of participants may lead to prioritise some of these heuristics over the others. In other words, some heuristics may be found to be more influential than the others.

- VI.** A possible large-scale employment of the model presented in this work is a freely accessed Web Service, which receives spatio-temporal coordinates from a mobile device as its input. The service would be responsible for identifying places, mapping its input to any available place semantics, and encoding the resulting PMG in PGML. The PGML file could either be kept on a server or maintained on the mobile device. Any necessary spatio-temporal middle-ware services, for translating between different systems of spatial coordinates or time zones, are assumed to be available.

## Bibliography

A.I. Abdelmoty, P.D. Smart, B.A. El-Geresy, and C.B. Jones. Supporting frameworks for the geospatial semantic web. In *SSTD '09: Proceedings of the 11th International Symposium on Advances in Spatial and Temporal Databases*, volume Lecture Notes in Computer Science 5644, pages 335–372. Springer-Verlag, 2009. ISBN 978-3-642-02981-3.

Piotr Adamczyk, Kevin Hamilton, Alan Chamberlain, Steve Benford, Nick Tandonitj, Amanda Oldroyd, Kate Hartman, Kati London, Sai Sriskandarajah, Eiman Kanjo, Peter Lanshoff, Kaoru Sezaki, Shin'ichi Konomi, Muaz A. Niazi, Hafiz F. Ahmad, Fauzan Mirza, Arshad Ali, George Roussos, Dikaios Papadogkonas, and Mark Levene. Urban computing and mobile devices. *IEEE Distributed Systems Online*, 8(7):2, 2007. ISSN 1541-4922. doi: <http://dx.doi.org/10.1109/MDSO.2007.46>.

Kiyoharu Aizawa, Datchakorn Tancharoen, Shinya Kawasaki, and Toshihiko Yamasaki. Efficient retrieval of life log based on context and content. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 22–31, New York, NY, USA, 2004. ACM. ISBN 1-58113-932-2. doi: <http://doi.acm.org/10.1145/1026653.1026656>.

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 273–280. ACM Press, 2004. ISBN 1-58113-881-4. doi: <http://doi.acm.org/10.1145/1008992.1009040>.

Sarabjot Anand and Bamshad Mobasher. Intelligent techniques for web personalization. *Intelligent Techniques for Web Personalization*, pages 1–36, 2005. doi: 10.1007/11577935\_1. URL [http://dx.doi.org/10.1007/11577935\\_1](http://dx.doi.org/10.1007/11577935_1).

Gennady L. Andrienko, Natalia V. Andrienko, and Stefan Wrobel. Visual analytics tools for analysis of movement data. *SIGKDD Explorations*, 9(2):38–46, 2007. URL <http://doi.acm.org/10.1145/1345448.1345455>.

Daniel Ashbrook and Thad Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5): 275–286, 2003. URL <http://dx.doi.org/10.1007/s00779-003-0240-0>.

Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60(1):208–221, 2007. URL <http://dx.doi.org/10.1016/j.datak.2006.01.013>.

Mauro Brunato and Roberto Battiti. PILGRIM: A location broker and mobility-aware recommendation system. In *PerCom*, pages 265–272. IEEE Computer Society, 2003. URL <http://computer.org/proceedings/percom/1893/18930265abs.htm>.

D.S. Burggraf. Geography markup language. *Data Science Journal*, 5(0):178–204, 2006.

Gilberto Câmara, Antonio Miguel, Vieira Monteiro, Argemiro Paiva, Ricardo Cartaxo, and Modesto De Souza. Action-driven ontologies of the geographical space: Beyond the field-object debate. In *Proceedings 1st International Conference on Geographical Information Science, GIScience*, pages 52–54, 2000.

Jae-Woo Chang, Rabindra Bista, Young-Chang Kim, and Yong-Ki Kim. Spatio-temporal similarity measure algorithm for moving objects on spatial networks. In Osvaldo Gervasi and Marina L. Gavrilova, editors, *Computational Science and Its Applications - ICCSA 2007, International Conference, Kuala Lumpur, Malaysia, August 26-29, 2007. Proceedings. Part III*, volume 4707 of *Lecture Notes in Com-*



*puter Science*, pages 1165–1178. Springer, 2007. ISBN 978-3-540-74482-5. URL [http://dx.doi.org/10.1007/978-3-540-74484-9\\_102](http://dx.doi.org/10.1007/978-3-540-74484-9_102).

Harry Chen, Filip Perich, Timothy W. Finin, and Anupam Joshi. SOUPA: Standard ontology for ubiquitous and pervasive applications. In *MobiQuitous*, pages 258–267. IEEE Computer Society, 2004. ISBN 0-7695-2208-4. URL <http://csdl.computer.org/comp/proceedings/mobiquitous/2004/2208/00/22080258abs.htm>.

M. Crossley, N. J. Kings, and J. R. Scott. Profiles — analysis and behaviour. *BT Technology Journal*, 21(1):56–66, 2003. ISSN 1358-3948. doi: <http://dx.doi.org/10.1023/A:1022404310934>.

Alexandre de Spindler, Moira C. Norrie, Michael Grossniklaus, and Beat Signer. Spatio-temporal proximity as a basis for collaborative filtering in mobile environments. In Moira C. Norrie, Schahram Dustdar, and Harald Gall, editors, *UM-ICS*, volume 242 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006. URL <http://ceur-ws.org/Vol-242/paper4.pdf>.

Ronald Denaux, Vania Dimitrova, Anthony G. Cohn, Catherine Dolbear, and Glen Hart. Rabbit to owl: ontology authoring with a cnl-based tool. In *Proceedings of the 2009 conference on Controlled natural language, CNL'09*, pages 246–264, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-14417-9, 978-3-642-14417-2. URL <http://portal.acm.org/citation.cfm?id=1893475.1893492>.

Anind K. Dey and Gregory D. Abowd. Towards a better understanding of context and context-awareness. In *Workshop on The What, Who, Where, When, and How of Context-Awareness, as part of the 2000 Conference on Human Factors in Computing Systems (CHI 2000)*, April 2000.

Goran M. Djuknic and Robert E. Richton. Geolocation and assisted GPS. *IEEE Computer*, 34(2):123–125, 2001. URL <http://dlib.computer.org/co/books/co2001/pdf/r2123.pdf>.

- D. Dransch. Activity and Context—A Conceptual Framework for Mobile Geoservices. *Map-based Mobile Services*, pages 31–42, 2005.
- T. D’Roza and G. Bilchev. An overview of location-based services. *BT Technology Journal*, 21(1):20–27, 2003. ISSN 1358-3948. doi: <http://dx.doi.org/10.1023/A:1022491825047>.
- Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006. URL <http://dx.doi.org/10.1007/s00779-005-0046-3>.
- J.A. Edwards and A. Templeton. The structure of perceived qualities of situations. *European journal of social psychology*, 35(6):705–723, 2005.
- Max J. Egenhofer. Toward the semantic geospatial web. In *Proceedings of the tenth ACM international symposium on Advances in geographic information systems*, pages 1–4. ACM Press, 2002. ISBN 1-58113-591-2. doi: <http://doi.acm.org/10.1145/585147.585148>.
- Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1):1–27, February 2003. ISSN 1533-5399.
- E. Emadzadeh, A. Nikfarjam, and S. Muthaiyah. A comparative study on measure of semantic relatedness function. In *The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010*, pages 94–97, 2010.
- Fredrik Espinoza, Per Persson, Anna Sandin, Hanna Nyström, Elenor Cacciatore, and Markus Bylund. GeoNotes: Social and navigational aspects of location-based information systems. *Lecture Notes in Computer Science*, 2201:2–??, 2001. ISSN 0302-9743. URL <http://link.springer-ny.com/link/service/series/0558/bibs/2201/22010002.htm>; <http://link.springer-ny.com/link/service/series/0558/papers/2201/22010002.pdf>.

S. Feng and C.L. Law. Assisted gps and its impact on navigation in intelligent transportation systems. In *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*, pages 926–931. IEEE, 2002.

Paolo Ferragina and Antonio Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Softw, Pract. Exper*, 38(2):189–225, 2008. URL <http://dx.doi.org/10.1002/spe.829>.

FireEagle. Fireeagle, 2010. URL <http://fireeagle.yahoo.net>.

Michael Flamm and Vincent Kaufmann. The concept of personal network of usual places as a tool for analysing human activity spaces. 11th World Conference on Transport Research, Berkeley, June 24-28, 2007, 2007.

A. Frank. Ontology for Spatio-temporal Databases. In M. Koubarakis and et al., editors, *Ontology for Spatio-temporal Databases*, volume Lecture Notes in Computer Science 2520, pages 9–77, 2003. URL [ftp://ftp.geoinfo.tuwien.ac.at/frank/chorochronos\\_chapter2.pdf](ftp://ftp.geoinfo.tuwien.ac.at/frank/chorochronos_chapter2.pdf).

Jon Froehlich, Mike Y. Chen, Ian E. Smith, and Fred Potter. Voting with your feet: An investigative study of the relationship between place visit behavior and preference. In Paul Dourish and Adrian Friday, editors, *UbiComp 2006: Ubiquitous Computing, 8th International Conference, UbiComp 2006, Orange County, CA, USA, September 17-21, 2006*, volume 4206 of *Lecture Notes in Computer Science*, pages 333–350. Springer, 2006. ISBN 3-540-39634-9. URL [http://dx.doi.org/10.1007/11853565\\_20](http://dx.doi.org/10.1007/11853565_20).

Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Anal. Appl.*, 13:113–129, January 2010. ISSN 1433-7541. doi: <http://dx.doi.org/10.1007/s10044-008-0141-y>. URL <http://dx.doi.org/10.1007/s10044-008-0141-y>.

GPX. Gpx, 2009. URL <http://www.topografix.com/gpx.asp>.

I.N. Gregory, C. Bennett, V.L. Gilham, and H.R. Southall. The Great Britain Historical GIS Project: from maps to changing human geography. *The Cartographic Journal*, 39(1):37–49, 2002. ISSN 0008-7041.

Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: Parsing and modeling location histories. In Max J. Egenhofer, Christian Freksa, and Harvey J. Miller, editors, *Geographic Information Science, Third International Conference, GIScience 2004, Adelphi, MD, USA, October 20-23, 2004, Proceedings*, volume 3234 of *Lecture Notes in Computer Science*, pages 106–124. Springer, 2004. ISBN 3-540-23558-2. URL <http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=3234&page=106>.

G. Hart, S. Temple, and H. Mizen. Tales of the river bank: first thoughts in the development of a topographic ontology. In F. Toppen and P. Prastacos, editors, *Proceedings of the 7th AGILE Conference*, pages 165–168, Heraklion, 2004. Crete University Press.

Dominik Heckmann and Antonio Krüger. A user modeling markup language (userML) for ubiquitous computing. In Peter Brusilovsky, Albert T. Corbett, and Fiorella de Rosis, editors, *User Modeling 2003, 9th International Conference, UM 2003, Johnstown, PA, USA, June 22-26, 2003, Proceedings*, volume 2702 of *Lecture Notes in Computer Science*, pages 393–397. Springer, 2003. ISBN 3-540-40381-7. URL <http://link.springer.de/link/service/series/0558/bibs/2702/27020393.htm>.

Dominik Heckmann, Boris Brandherm, Tim Schwartz, and Margeritta von Wilamowitz-Moellendorff. GUMO, the general user model ontology. In *10th International Conference on User Modeling*, Edinburgh, Scotland, 2005.

J. Hightower. From position to place. In *Proceedings of The 2003 Workshop on Location-Aware Computing*, pages 10–12. Citeseer, 2003.

Linda L. Hill, James Frew, and Qi Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5, January 1999. URL <http://www.dlib.org/dlib/january99/hill/01hill.html>.

Marty Himmelstein. Local search: The internet is the yellow pages. *IEEE Computer*, 38(2):26–34, 2005. URL <http://doi.ieeecomputersociety.org/10.1109/MC.2005.65>.

Annika Hinze and Agnès Voisard. Locations- and time-based information delivery in tourism. In Thanasis Hadzilacos, Yannis Manolopoulos, John F. Roddick, and Yannis Theodoridis, editors, *Advances in Spatial and Temporal Databases, 8th International Symposium, SSTD 2003, Santorini Island, Greece, July 24-27, 2003, Proceedings*, volume 2750 of *Lecture Notes in Computer Science*, pages 489–507. Springer, 2003. ISBN 3-540-40535-6. URL <http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=2750&spage=489>.

Matthew Hockenberry and Ted Selker. A sense of spatial semantics. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 851–856, New York, NY, USA, 2006. ACM. ISBN 1-59593-298-4. doi: <http://doi.acm.org/10.1145/1125451.1125618>.

Jongyi Hong, Eui-Ho Suh, Junyoung Kim, and Su-Yeon Kim. Context-aware system for proactive personalized service based on context history. *Expert Syst. Appl*, 36(4): 7448–7457, 2009. URL <http://dx.doi.org/10.1016/j.eswa.2008.09.002>.

Masaki Ito, Jin Nakazawa, and Hideyuki Tokuda. mPATH: An interactive visualization framework for behavior history. In *AINA*, pages 247–252. IEEE Computer Society, 2005. ISBN 0-7695-2249-1. URL <http://doi.ieeecomputersociety.org/10.1109/AINA.2005.253>.

Christopher B. Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. In *Proceedings of the International Conference*

*on Spatial Information Theory, COSIT*, volume Lecture Notes in Computer Science 2205, pages 322–335. Springer-Verlag, 2001. ISBN 3-540-42613-2.

Christopher B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies an overview of the spirit project. In *SIGIR '02*, pages 387–388, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: <http://doi.acm.org/10.1145/564376.564457>.

Matt Jones, George Buchanan, Richard Harper, and Pierre Louis Xech. Questions not answers: a novel mobile search technique. In Mary Beth Rosson and David J. Gilmore, editors, *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 155–158. ACM, 2007. ISBN 978-1-59593-593-9. URL <http://doi.acm.org/10.1145/1240624.1240648>.

T. Jordan, M. Raubal, B. Gartrell, and M. J. Egenhofer. An affordance-based model of place in gis. In T. Poiker and N. Chrisman, editors, *Eighth International Symposium on Spatial Data Handling*, pages 98–109. IUG, 1998.

T. Kapler, R. Harper, and W. Wright. Correlating events with tracked movements in time and space: a GeoTime case study. In *Proceedings of the 2005 Intelligence Analysis Conference*, 2005.

Kevin Keenoy and Mark Levene. Personalisation of web search. In Bamshad Mombasher and Sarabjot S. Anand, editors, *ITWP*, volume 3169 of *Lecture Notes in Computer Science*, pages 201–228. Springer, 2003. ISBN 3-540-29846-0. URL [http://dx.doi.org/10.1007/11577935\\_11](http://dx.doi.org/10.1007/11577935_11).

P. Kikiras, V. Tsetsos, and S. Hadjiefthymiades. Ontology-based user modeling for pedestrian navigation systems. In *ECAI 2006 Workshop on Ubiquitous User Modeling (UbiqUM)*, Riva del Garda, Italy, 2006.

Ig-Jae Kim, Sang Chul Ahn, and Hyung-Gon Kim. Personalized life log media system in ubiquitous environment. In Frank Stajano, Hyung Joong Kim, Jong-

Suk Chae, and Seong-Dong Kim, editors, *ICUCT*, volume 4412 of *Lecture Notes in Computer Science*, pages 20–29. Springer, 2006a. ISBN 978-3-540-71788-1. URL [http://dx.doi.org/10.1007/978-3-540-71789-8\\_3](http://dx.doi.org/10.1007/978-3-540-71789-8_3).

I.J. Kim, S.C. Ahn, H. Ko, and H.G. Kim. PERSONE: personalized experience recoding and searching on networked environment. In *Proceedings of the 3rd ACM workshop on Continuous archival and retrieval of personal experiences*, pages 49–54. ACM, 2006b.

C. Kray and J. Baus. A Survey of Mobile Guides. In *Workshop on HCI in mobile guides, 5th Int. Symposium on HCI with Mobile Devices and Services. Udine, Italy, 2003*.

Antonio Krüger, Jörg Baus, Dominik Heckmann, Michael Kruppa, and Rainer Wasinger. Adaptive mobile guides. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 521–549. Springer, 2007. ISBN 978-3-540-72078-2. URL [http://dx.doi.org/10.1007/978-3-540-72079-9\\_17](http://dx.doi.org/10.1007/978-3-540-72079-9_17).

W. Kuhn. Ontologies from text. In M.J. Egenhofer and D.M. Mark, editors, *GIScience*, Savannah, GA, 2000.

Werner Kuhn. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science*, 15(7):613–631, 2001. doi: 10.1080/13658810110061180.

Werner Kuhn. An Image-Schematic Account of Spatial Categories. In *Spatial Information Theory*, number 4736 in *Lecture Notes in Computer Science*, pages 152–168. Springer-Verlag, 2007. 8th International Conference, COSIT 2007. Melbourne, Australia.

Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 19:191–233, 2000.

J.H. Lea, H.S. Ryu, Sa.G. Lee, Y.B. Lee, and S.R. Kim. Tracking a personalized trail effectively with a mobile phone. In *In S. Jang and K. van Laerhoven S.G. Lee and K. Mase (Eds.), Proceedings of the Second International Workshop on Personalized Context Modeling and Management for UbiComp Applications*, 2006.

Ye Lei and Hui Lin 0002. The web integration of the GPS+GPRS+GIS tracking system and real-time monitoring system based on MAS. In James D. Carswell and Taro Tezuka, editors, *Web and Wireless Geographical Information Systems, 6th International Symposium, W2GIS 2006, Hong Kong, China, December 4-5, 2006, Proceedings*, volume 4295 of *Lecture Notes in Computer Science*, pages 54–65. Springer, 2006. ISBN 3-540-49466-9. URL [http://dx.doi.org/10.1007/11935148\\_6](http://dx.doi.org/10.1007/11935148_6).

L. Liao, D.J. Patterson, D. Fox, and H. Kautz. Building personal maps from gps data. *Annals of the New York Academy of Sciences*, 1093(1 Progress in Convergence: Technologies for Human Wellbeing):249–265, 2006.

Lin Liao, Dieter Fox, and Henry A. Kautz. Location-based activity recognition using relational markov networks. In Leslie Pack Kaelbling and Alessandro Saffioti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 773–778. Professional Book Center, 2005. ISBN 0938075934. URL <http://www.ijcai.org/papers/1572.pdf>.

Hyeeun Lim and Nupur Bhatnagar. Quantifying the predictability of a personal place, 2008. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.101.1282>.

R. Lindsey, V.D. Veksler, A. Grintsvayg, and W.D. Gray. Effects of corpus selection on semantic relatedness. In *8th International Conference of Cognitive Modeling, ICCM2007*, 2007.

Juhong Liu, Ouri Wolfson, and Huabei Yin. Extracting semantic location from outdoor positioning systems. In *Mobile Data Management*, page 73. IEEE Computer



Society, 2006. ISBN 0-7695-2526-1. URL <http://doi.ieeecomputersociety.org/10.1109/MDM.2006.87>.

Natalia Marmasse and Chris Schmandt. Location-aware information delivery with commotion. In Peter J. Thomas and Hans-Werner Gellersen, editors, *Handheld and Ubiquitous Computing, Second International Symposium, HUC 2000, Bristol, UK, September 25-27, 2000, Proceedings*, volume 1927 of *Lecture Notes in Computer Science*, pages 157–171. Springer, 2000. ISBN 3-540-41093-7. URL <http://link.springer.de/link/service/series/0558/bibs/1927/19270157.htm>.

Irina Matveeva. *Generalized latent semantic analysis for document representation*. PhD thesis, University of Chicago Chicago, 2008.

Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000. URL <http://doi.acm.org/10.1145/345124.345169>.

Adriano Moreira and Maribel Santos. From gps tracks to context inference of high-level context information through spatial clustering. In *2nd International Conference and Exhibition on Geographic Information GIS Planet 2005*, Estoril, Portugal, 2005.

Yasuhiko Morimoto, Masaki Aono, Michael E. Houle, and Kevin S. McCurley. Extracting spatial knowledge from the web. In *SAINT*, pages 326–333. IEEE Computer Society, 2003. ISBN 0-7695-1872-9. URL <http://computer.org/proceedings/saint/1872/18720326abs.htm>.

D. Mountain and J. Raper. Modelling human spatio-temporal behaviour: a challenge for location-based services. In *Proceedings of the Sixth International Conference on GeoComputation, University of Queensland, Brisbane, Australia*, pages 24–26, 2001.

O.M. Mozos, P. Jensfelt, H. Zender, G.J.M. Kruijff, and W. Burgard. From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots. In *IEEE Int. Conf. on Robotics and Automation (ICRA) Workshop on Semantic Information in Robotics*, 2007.

Maurice D. Mulvenna, Sarabjot S. Anand, and Alex G. Büchner. Personalization on the net using web mining: introduction. *Commun. ACM*, 43:122–125, August 2000. ISSN 0001-0782. doi: <http://doi.acm.org/10.1145/345124.345165>. URL <http://doi.acm.org/10.1145/345124.345165>.

R. Newbould and R. Collingridge. Profiling — technology. *BT Technology Journal*, 21(1):44–55, 2003. ISSN 1358-3948. doi: <http://dx.doi.org/10.1023/A:1022400226864>.

Nam T. Nguyen, Dinh Q. Phung, Svetha Venkatesh, and Hung Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 955–960, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: <http://dx.doi.org/10.1109/CVPR.2005.203>. URL <http://dx.doi.org/10.1109/CVPR.2005.203>.

T.A.S. Nielsen and H.H. Hovgesen. GPS in Travel and Activity Surveys. *Trafikdays på Aalborg Universitet. Aalborg Universitet*, 2004.

Mihoko Niitsuma and Hideki Hashimoto. Extraction of space-human activity association for design of intelligent environment. In *ICRA*, pages 1814–1819. IEEE, 2007. URL <http://dx.doi.org/10.1109/ROBOT.2007.363585>.

A.M. Nivala and L.T. Sarjakoski. An approach to intelligent maps: context awareness. In *The 2nd Workshop on HCI in Mobile Guides*, 2003a.

Annu-Maaria Nivala and L. Tiina Sarjakoski. Need for context-aware topographic maps in mobile devices. In Kirsi Virrantaus and Håvard Tveite, editors, *Scangis'2003 - The 9th Scandinavian Research Conference on Geographical Information Science, 4-6 June 2003, Espoo, Finland - Proceedings*, pages 15–29. Department of Surveying, Helsinki University of Technology, 2003b. ISBN 951-22-6565-6. URL <http://www.scangis.org/scangis2003/papers/22.pdf>.

P. Nurmi and S. Bhattacharya. Identifying meaningful places: the non-parametric way. *Pervasive Computing*, pages 111–127, 2008.

A. O'Connor, A. Zerger, and B. Itami. Geo-temporal tracking and analysis of tourist movement. *Mathematics and Computers in Simulation*, 69(1-2):135–150, 2005. URL <http://dx.doi.org/10.1016/j.matcom.2005.02.036>.

openstreetmap. openstreetmap, 2010. URL <http://www.openstreetmap.org>.

Donald J. Patterson, Lin Liao, Dieter Fox, and Henry A. Kautz. Inferring high-level behavior from low-level sensors. In Anind K. Dey, Albrecht Schmidt, and Joseph F. McCarthy, editors, *UbiComp 2003: Ubiquitous Computing, 5th International Conference, Seattle, WA, USA, October 12-15, 2003, Proceedings*, volume 2864 of *Lecture Notes in Computer Science*, pages 73–89. Springer, 2003. ISBN 3-540-20301-X. URL <http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=2864&page=73>.

Matthew Perry, Farshad Hakimpour, and Amit Sheth. Analyzing the space, and time: an ontology-based approach. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, GIS '06*, pages 147–154. ACM, 2006. ISBN 1-59593-529-0.

P. Persson, F. Espinoza, P. Fagerberg, A. Sandin, and R. C "oster. GeoNotes: A location-based information system for public spaces. *Designing information spaces: the social navigation approach*, pages 151–173, 2002.

Adrian Popescu, Gregory Grefenstette, and Pierre-Alain Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 85–93. ACM, 2008. ISBN 978-1-59593-998-2.

S. Pradhan, C. Brignone, Jun-Hong Cui, A. McReynolds, and M.T. Smith. Websigns: hyperlinking physical locations to the web. *Computer*, 34(8):42–48, aug. 2001. ISSN 0018-9162. doi: 10.1109/2.940012.

Salil Pradhan. Semantic location. *Personal and Ubiquitous Computing*, 4(4):213–216, 2000. URL <http://link.springer.de/link/service/journals/00779/bibs/0004004/00040213.htm>.

Iko Pramudiono, Takahiko Shintani, Katsumi Takahashi, and Masaru Kitsuregawa. User behavior analysis of location aware search engine. In *Mobile Data Management*, pages 139–145. IEEE Computer Society, 2002. ISBN 0-7695-1500-2. URL <http://computer.org/proceedings/mdm/1500/15000139abs.htm>.

Alexander Pretschner and Susan Gauch. Personalization on the web. Technical report, University of Kansas, 1999.

M. Raubal and W. Kuhn. Ontology-based task simulation. *Spatial Cognition and Computation*, 4(1):15–37, 2004.

EC Relph. *Place and placelessness*. Pion Ltd, 1976.

Doug Riecken. Introduction: personalized views of personalization. *Commun. ACM*, 43(8):26–28, 2000. URL <http://doi.acm.org/10.1145/345124.345133>.

R.D. Rugg, M.J. Egenhofer, and W. Kuhn. Formalizing behavior of geographic feature types. *Geographical Systems*, 4:159–180, 1997.

Beatrice Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. University of Pennsylvania, School of Engineering and Applied Science, Dept. of Computer and Information Science., Philadelphia, 1990. ISBN 0582291496.

T. Saponas, J. Lester, J. Froehlich, J. Fogarty, and J. Landay. ilearn on the iphone: Real-time human activity classification on commodity mobile phones. *University of Washington CSE Tech Report UW-CSE-08-04-02*, 2008.

Arno Scharl, Hermann Stern, and Albert Weichselbraun. Annotating and visualizing location data in geospatial web applications. In Susanne Boll, Christopher Jones, Eric Kansa, Puneet Kishor, Mor Naaman, Ross Purves, Arno Scharl, and Erik Wilde,

editors, *Proceedings of the First International Workshop on Location and the Web, LocWeb 2008, Beijing, China, April 22, 2008*, volume 300 of *ACM International Conference Proceeding Series*, pages 65–68. ACM, 2008. ISBN 978-1-60558-160-6. URL <http://doi.acm.org/10.1145/1367798.1367809>.

Simon Scheider and Werner Kuhn. Affordance-based categorization of road network data using a grounded theory of channel networks. *International Journal of Geographical Information Science*, 24(8):1249–1267, 2010.

B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE, 2008.

Falko Schmid and Kai-florian Richter. Extracting places from location data streams. In *In A. Zipf (Eds.), Workshop Proceedings (UbiGIS, 2006)*. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.6928>.

Falko Schmid, Kai-Florian Richter, and Patrick Laube. Semantic trajectory compression. In Nikos Mamoulis, Thomas Seidl, Torben Bach Pedersen, Kristian Torp, and Ira Assent, editors, *Advances in Spatial and Temporal Databases, 11th International Symposium, SSTD 2009, Aalborg, Denmark, July 8-10, 2009, Proceedings*, volume 5644 of *Lecture Notes in Computer Science*, pages 411–416. Springer, 2009. ISBN 978-3-642-02981-3. URL <http://dx.doi.org/10.1007/978-3-642-02982-0>.

Schmidt-Belz, Barbara, Nick, Achim, Poslad, Stefan, and Zipf, Alex. Personalized and location-based mobile tourism services. In *Proc. of Mobile-HCI, 2002*. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8983>.

Stefan Schnfelder, Kay W. Axhausen, Nicolas Antille, and Michel Bierlaire. Exploring the potentials of automatically collected GPS data for travel behaviour analysis –, September 06 2002. URL <http://citeseer.ist.psu.edu/602061.html>; <http://rosowww.epfl.ch/mbi/papers/schoenfelder.pdf>.

S. Schonfelder and U. Samaga. Where do you want to go today?—More observations on daily mobility. In *3th Swiss Transport Research Conference, Ascona*, 2003.

J. Schuurmans and E. Zijlstra. Towards a continuous personalization experience. In *Proceedings of the conference on Dutch directions in HCI*, page 19. ACM, 2004.

Angela Schwering. Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS*, 12(1):5–29, 2008. ISSN 1467-9671. doi: 10.1111/j.1467-9671.2008.01084.x. URL <http://dx.doi.org/10.1111/j.1467-9671.2008.01084.x>.

S. Searby. Personalisation — an overview of its use and potential. *BT Technology Journal*, 21(1):13–19, 2003. ISSN 1358-3948. doi: <http://dx.doi.org/10.1023/A:1022439824138>.

T. Selker and W. Burleson. Context-aware design and interaction in computer systems. *IBM Systems Journal*, 39(3.4):880–891, 2000. ISSN 0018-8670. doi: 10.1147/sj.393.0880.

S. Sen. Two types of hierarchies in geospatial ontologies. In *GeoSpatial Semantics*, pages 1–19, 2007. URL [http://dx.doi.org/10.1007/978-3-540-76876-0\\_1](http://dx.doi.org/10.1007/978-3-540-76876-0_1).

Sumit Sen. Use of affordances in geospatial ontologies. In *Proceedings of the 2006 international conference on Towards affordance-based robot control*, pages 122–139, Berlin, Heidelberg, 2008. Springer Verlag. ISBN 3-540-77914-0, 978-3-540-77914-8. URL <http://portal.acm.org/citation.cfm?id=1787357.1787366>.

C. Shahabi and Y.S. Chen. Web information personalization: Challenges and approaches. *Databases in Networked Information Systems*, pages 5–15, 2003.

Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399, 2006. URL <http://dx.doi.org/10.1016/j.compenvurbsys.2005.08.003>.

Rainer Simon, Peter Fröhlich, and Hermann Anegg. Beyond location based - the spatially aware mobile phone. In James D. Carswell and Taro Tezuka, editors, *Web and Wireless Geographical Information Systems, 6th International Symposium, W2GIS 2006, Hong Kong, China, December 4-5, 2006, Proceedings*, volume 4295 of *Lecture Notes in Computer Science*, pages 12–21. Springer, 2006. ISBN 3-540-49466-9. URL [http://dx.doi.org/10.1007/11935148\\_2](http://dx.doi.org/10.1007/11935148_2).

P.D. Smart, C.B. Jones, and F.A. Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Sixth international conference on Geographic Information Science, GIScience 2010*, volume *Lecture Notes in Computer Science 6292*, pages 234–248, 2010.

Barry Smith and David M. Mark. Geographical categories: an ontological investigation. *International Journal of Geographical Information Science*, 15(7):591–612, 2001.

Timothy Sohn, Kevin A. Li, Gunny Lee, Ian E. Smith, James Scott, and William G. Griswold. Place-its: A study of location-based reminders on mobile phones. In Michael Beigl, Stephen S. Intille, Jun Rekimoto, and Hideyuki Tokuda, editors, *UbiComp 2005: Ubiquitous Computing, 7th International Conference, UbiComp 2005, Tokyo, Japan, September 11-14, 2005, Proceedings*, volume 3660 of *Lecture Notes in Computer Science*, pages 232–250. Springer, 2005. ISBN 3-540-28760-4. URL [http://dx.doi.org/10.1007/11551201\\_14](http://dx.doi.org/10.1007/11551201_14).

Mike Spence, Cormac Driver, and Siobhán Clarke. Sharing context history in mobile, context-aware trails-based applications. In *1st international workshop on exploiting context histories in smart environments (ECHISE 2005), Pervasive 2005*, May 2005.

S. Steiniger, M. Neun, and A. Edwardes. Lecture notes: Foundations of location based services. University of Zurich, 2006.

P. Stopher, Q. Jiang, and C. FitzGerald. Deducing mode and purpose from GPS data. In *11th TRB National Transportation Planning Applications Conference, Daytona Beach, 2007*.

Yuichiro Takeuchi and Masanori Sugimoto. Cityvoyager: An outdoor recommendation system based on user location history. In Jianhua Ma, Hai Jin, Laurence Tianruo Yang, and Jeffrey J. P. Tsai, editors, *Ubiquitous Intelligence and Computing, Third International Conference, UIC 2006, Wuhan, China, September 3-6, 2006, Proceedings*, volume 4159 of *Lecture Notes in Computer Science*, pages 625–636. Springer, 2006. ISBN 3-540-38091-4. URL [http://dx.doi.org/10.1007/11833529\\_64](http://dx.doi.org/10.1007/11833529_64).

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Potential for personalization. *ACM Trans. Comput.-Hum. Interact.*, 17(1), 2010. URL <http://doi.acm.org/10.1145/1721831.1721835>.

Michael Terry, Mynatt, Elizabeth D., Kathy Ryall, and Darren Leigh. Social net: using patterns of physical proximity over time to infer shared interests. In *Proceedings of ACM CHI 2002 Conference on Human Factors in Computing Systems*, volume 2 of *Short Talks*, pages 816–817, 2002. URL <http://doi.acm.org/10.1145/506443.506612>.

Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning, EMCL '01*, pages 491–502, London, UK, 2001. Springer-Verlag. ISBN 3-540-42536-5. URL <http://portal.acm.org/citation.cfm?id=645328.650004>.

B. Tversky and K. Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15(1):121–149, 1983.

V.D. Veksler, A. Grintsvayg, R. Lindsey, and W.D. Gray. A proxy for all your semantic needs. In *29th Annual Meeting of the Cognitive Science Society, CogSci2007, 2007*.



Jingtao Wang and John Canny. End-user place annotation on mobile devices: a comparative study. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 1493–1498, New York, NY, USA, 2006. ACM. ISBN 1-59593-298-4. doi: <http://doi.acm.org/10.1145/1125451.1125725>.

Xiaohang Wang, Daqing Zhang, Tao Gu, and Hung Keng Pung. Ontology based context modeling and reasoning using OWL. In *PerCom Workshops*, pages 18–22. IEEE Computer Society, 2004. URL <http://csdl.computer.org/comp/proceedings/percomw/2004/2106/00/21060018abs.htm>.

Norbert Weibenberg, Rudiger Gartmann, and Agnes Voisard. An ontology-based approach to personalized situation-aware mobile service supply. *Geoinformatica*, 10: 55–90, March 2006. ISSN 1384-6175. doi: 10.1007/s10707-005-4886-9. URL <http://portal.acm.org/citation.cfm?id=1116038.1116042>.

Ryan Wishart, Karen Henricksen, and Jadwiga Indulska. Context obfuscation for privacy via ontological descriptions. In Thomas Strang and Claudia Linnhoff-Popien, editors, *Location- and Context-Awareness, First International Workshop, LoCA 2005, Oberpfaffenhofen, Germany, May 12-13, 2005, Proceedings*, volume 3479 of *Lecture Notes in Computer Science*, pages 276–288. Springer, 2005. ISBN 3-540-25896-5. URL [http://dx.doi.org/10.1007/11426646\\_26](http://dx.doi.org/10.1007/11426646_26).

Hongbo Yu and Shih-Lung Shaw. Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographical Information Science*, 22(4):409–430, 2008. URL <http://dx.doi.org/10.1080/13658810701427569>.

Changqing Zhou, Dan Frankowski, Pamela J. Ludford, Shashi Shekhar, and Loren G. Terveen. Discovering personal gazetteers: an interactive clustering approach. In Dieter Pfoser, Isabel F. Cruz, and Marc Ronthaler, editors, *12th ACM International Workshop on Geographic Information Systems, ACM-GIS 2004, November 12-13,*

2004, Washington, DC, USA, *Proceedings*, pages 266–273. ACM, 2004. ISBN 1-58113-979-9. URL <http://doi.acm.org/10.1145/1032222.1032261>.

Changqing Zhou, Pamela Ludford, Dan Frankowski, and Loren Terveen. Talking about place: An experiment in how people describe places. *Proc. Pervasive, Short Paper, 2005a*. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.88.3668>.

Changqing Zhou, Pamela J. Ludford, Dan Frankowski, and Loren G. Terveen. An experiment in discovering personally meaningful places from location data. In Gerrit C. van der Veer and Carolyn Gale, editors, *Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005*, pages 2029–2032. ACM, 2005b. ISBN 1-59593-002-7. URL <http://doi.acm.org/10.1145/1056808.1057084>.

Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25(3):12, 2007. ISSN 1046-8188. doi: <http://doi.acm.org/10.1145/1247715.1247718>.

Alexander Zipf and Matthias Jöst. Implementing adaptive mobile GI services based on ontologies: Examples from pedestrian navigation support. *Computers, Environment and Urban Systems*, 30(6):784–798, 2006. URL <http://dx.doi.org/10.1016/j.compenvurbsys.2006.02.005>.

