

Knowledge Media Technologies

First International Core-to-Core Workshop

Klaus P. Jantke & Gunther Kreuzberger (eds.)

Nr. 21

July 2006

Herausgeber: Der Rektor der Technischen Universität Ilmenau
Redaktion: Institut für Medien- und Kommunikationswissenschaft,
Prof. Dr. Paul Klimsa
ISSN 1617-9048
Kontakt: Klaus P. Jantke, Tel.: +49 3677 69 47 35
E-Mail: klaus-peter.jantke@tu-ilmenau.de

Table of Contents

Knowledge Federation and Utilization

Yuzuru Tanaka
A Formal Model for the Proximity-Based Federation among Smart Objects.....5

Hidetoshi Nonaka and Masahito Kurihara
Protean sensor network for context aware services.....21

Hosting Team Presentations

Klaus P. Jantke
Pattern Concepts for Digital Games Research.....29

Anja Beyer
Security Aspects in Online Games Targets, Threats and Mechanisms.....41

Jürgen Nützel and Mario Kubek
Personal File Sharing in a Personal Bluetooth Environment47

Henrik Tonn-Eichstädt
User Interfaces for People with Special Needs.....54

Knowledge Clustering and Text Summarization

Jan Poland and Thomas Zeugmann
Clustering the Google Distance with Eigenvectors and Semidefinite Programming61

Tetsuya Yoshida
Toward Soft Clustering in the Sequential Information Bottleneck Method.....70

Masaharu Yoshioka and Makoto Haraguchi
Multiple News Articles Summarization based on Event Reference Information with Hierarchical Structure Analysis.....79

Core-to-Core Partners I

Roland H. Kaschek, Heinrich Mayr, Klaus Kienzl
Using Metaphors for the Pragmatic Construction of Discourse Domains.....88

Gunar Fiedler, Bernhard Thalheim, Dirk Fleischer, Heye Rumohr
Data Mining in Biological Data for BiOkIS.....104

Christian Reuschling and Sandra Zilles
Personalized information filtering for mobile applications.....114

Data and Text Mining

Hiroki Arimura and Takeaki Uno

Efficient Algorithms for Mining Maximal Flexible Patterns in Texts and Sequences.....125

Takuya Kida, Takashi Uemura, and Hiroki Arimura

Application of Truncated Suffix Trees to Finding Sentences from the Internet.....136

Shin-ichi Minato

Generating Frequent Closed Item Sets Using Zero-suppressed BDDs.....145

Kimihito Ito, Manabu Igarashi, Ayato Takada

Data Mining in Amino Acid Sequences of H3N2 Influenza Viruses Isolated during 1968 to 2006.....154

Core-to-Core Partners II

Nigel Waters

The Links Between GIScience and Information Science159

Rainer Knauf and Klaus P. Jantke

Storyboarding – An AI Technology to Represent, Process, Evaluate, and Refine Didactic Knowledge.....170

Jürgen Cleve, Christian Andersch, Stefan Wissuwa

Data Mining on Traffic Accident Data.....180

Core-to-Core Partners III

Torsten Brix, Ulf Döring, Sabine Trott, Rike Brecht, Hendrik Thomas

The Digital Mechanism and Gear Library – a Modern Knowledge Space.....188

Donna M. Delparte

The Use of GIS in Avalanche Modelling.....195

Jochen Felix Böhm and Jun Fujima

A Case Study for Knowledge Externalization using an Interactive Video Environment on E-learning.....199

Preface

Knowledge Media Technologies has been chosen as a headline for the present workshop proceedings to focus the audience's attention to the central problem domain around which a variety of investigations have been grouped.

The Japanese Core-to-Core cooperation programme has been providing the framework for this workshop. This farsighted programme does encourage and support Japanese centers of research excellence to widen and deepen their scientific connections to other research centers around the world. For this purpose, scientists from Europe, from Northern America, and from other regions such as New Zealand decided to go to Dagstuhl Castle, Germany, to meet a larger group of Japanese scientists headed and coordinated by Dr. Yuzuru Tanaka, professor and director of the Meme Media Laboratory at Hokkaido University Sapporo, Japan.

Tanaka's research center is strong in and world-wide known for its work on knowledge media technologies.

For almost half a century, the transformation into a knowledge society has been recognized world-wide. Science and technologies as well as politics strive hard to act appropriately.

With the advent of global communication networks, the development and exchange of information is boosted. Combinations of knowledge media sources open unpredictable potentials.

But most sources on the Web are not prepared to be effectively combined with other sources. What is known as the enterprise application integration problem locally shows even more seriously and with higher potential impact globally.

The Japanese programme on *Knowledge Media Technologies for the Advanced Federation, Utilization and Distribution of Knowledge Ressources* addresses one of the most important scientific and engineering issues of the present and the near future. The work belongs to the wide field of information and communication technologies, but is surely relevant to all other disciplines.

The workshop programme is completed by three sessions on *Digital Games*. This field is a rather new discipline with high economic, scientific, and social potentials. Issues of game knowledge lead to a natural integration of these sessions into the workshop program, although the *Digital Games* sessions are not reflected in the present proceedings.

Preparing the proceedings of this *First International Core-to-Core Workshop* has been a great pleasure. We are looking forward to invaluable cooperations of an unforeseeable reach and impact.

Klaus P. Jantke
Gunther Kreuzberger

Ilmenau, July 2006

A Formal Model for the Proximity-Based Federation among Smart Objects

Yuzuru Tanaka
Meme Media Laboratory
Hokkaido University
N13, W8, Sapporo, 060-8628 Japan
{tanaka, fujima, ohigashi}@meme.hokudai.ac.jp

Abstract

This paper proposes a new formal model of autonomic proximity-based federation among smart objects with wireless network connectivity and services available on the Internet. Federation here denotes the definition and execution of interoperation among smart objects and/or services that are accessible either through the Internet or through peer-to-peer *ad hoc* communication without previously designed interoperation interface. This paper proposes a new formal model for autonomic federation among smart objects through peer-to-peer communication, and then extends this model to cope with federation among smart objects through the Internet as well as federation including services over the Web. Each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object. Here we consider the matching of service-requesting queries and service-providing capabilities that are represented as service-requesting ports and service-providing ports, instead of the matching of a service requesting message with a service-providing message. Our model extracts the federation mechanism of each smart object as its interoperation interface that is logically represented as a set of ports. Such an abstract description of each smart object from the view point of its federation capability will allow us to discuss both the matching mechanism for federation and complex federation among smart objects in terms of a simple mathematical model. Applications can be described from the view point of their federation structures. This enables us to extract a common substructure from applications sharing the same typical federation scenario. Such an extracted substructure works as an application framework for this federation scenario. This paper first proposes our mathematical modeling of smart objects and their federation, then gives the semantics of our federation model, and finally, based on our model, shows several application frameworks using smart object federation.

1. Introduction

This paper proposes a new formal model of autonomic proximity-based federation among smart objects with wireless network connectivity and services available on the Internet. Smart objects here denote computing devices such as RFID tag chips, smart chips with sensors and/or actuators that are embedded in pervasive computing environments such as home, office, and social infrastructure environments, mobile PDAs, intelligent electronic appliances, embedded computers, and access points with network servers. Federation here denotes the defini-

tion and execution of interoperation among smart objects and/or services that are accessible either through the Internet or through peer-to-peer *ad hoc* communication without previously designed interoperation interface. Federation is different from integration in which member objects involved are assumed to have previously designed standard interoperation interface.

The Web works not only as an open publishing repository of documents, but also as an open repository of services represented as Web applications and/or Web services. Pervasive computing denotes an open system of

computing resources in which users can dynamically select and federate some of these computing resources as well as some of those computing resources accessible only through peer-to-peer *ad hoc* wireless communication to perform their jobs satisfying their dynamically changing demands. Such computing resources include not only services on the Internet, but also functions provided by embedded and/or mobile smart objects connected either to the Internet through WiFi communication or directly to the user's device through peer-to-peer *ad hoc* wireless connection.

Federation over the Web is attracting the attention not only for user-oriented integration of mutually related legacy Web-based business applications and services, but also for interdisciplinary and international advanced reuse and interoperation of heterogeneous intellectual resources especially in scientific simulations, digital libraries, and research activities. Federation may be classified into two types: *ad hoc* federation defined by users and autonomic federation defined by programs. We have already proposed our approach based on meme media technologies [5] for *ad hoc* federation over the Web [8, 7, 9], and its extension to aggregate *ad hoc* federation over the Web [6]. Here in this paper, we will propose a new formal model for autonomic federation among smart objects through peer-to-peer communication, and then extend this model to cope with federation among smart objects through the Internet as well as federation including services over the Web.

Every preceding studies on federation basically proposed two things, i.e., a standard communication protocol with a language to use it, and a repository-and-lookup service that allows each member to register its service-providing capabilities, and to request a service that matches its demand. For each request with a specified demand, a repository-and-lookup service searches all the registered service capabilities for those satisfying the specified demand. A repository-and-lookup service

matches service-requesting queries with corresponding service-providing capabilities. The origin of such an idea can be found in the original tuple space model Linda [1] and its extension Lime [4] that copes with mobile objects by providing each of them a dedicated tuple space. Linda and Lime are languages that use tuples to register service-providing capabilities and to issue service-requesting queries to a repository-and-lookup service called a tuple space. Java Space [2] and Jini [3] are Java versions of Linda and Lime architectures.

In these preceding studies, each service-requesting message is matched with a service-providing message in a common repository of service requests and service-providing capabilities. Both each service-requesting message and each service-providing message are represented as tuples, and they are matched with each other based on the tuple matching algorithm. Each tuple consists of a single or more than one attribute-value pair in which some attributes may take variables. Such a match of tuples is temporarily used to associate two objects, one issuing a service-requesting message and the other issuing a service-providing message. There may be more than one repository-and-lookup service in such a system. Each of these services can delegate such a query that is not matched in itself to one or more than one neighboring repository-and-lookup service, and obtain a match result.

Here in this paper, we focus on the interface of smart objects for their federation with each other and with services over the Web. We will hide any details on how functions of each smart object are implemented, and focus on abstract level modeling of its federation interface. Each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object. Here, we consider the matching of service-requesting queries and service-providing capabilities that are represented as service-

requesting ports and service-providing ports, instead of the matching of a service requesting message with a service-providing message. One of these functions requests the service provided by the other function. Therefore, each port represents an interface of either a service-requesting function or a service-providing function.

In the preceding research studies, federation mechanisms were described in the codes that define the behaviors of participating smart objects, and were not separated from these codes to be discussed independently from them. Here in this paper, we will extract the federation mechanism of each smart object as its interoperation interface that is logically represented as a set of ports. Such an abstract description of each smart object from the view point of its federation capability will allow us to discuss both the matching mechanism for federation and complex federation among smart objects in terms of a simple mathematical model. Applications can be described from the view point of their federation structures. This enables us to extract a common substructure from applications sharing the same typical federation scenario. Such an extracted substructure may work as an application framework for this federation scenario.

In the following sections, we will first propose our mathematical modeling of smart objects and their federation, then give the semantics of our federation model, and finally, based on our model, show several application frameworks using smart object federation.

2. Smart Object and its Formal Modeling

Each smart object communicates with another smart object through a peer-to-peer communication facility, which is either a direct cable connection or a wireless connection such as Bluetooth and IrDA wireless connection. Some smart objects may have WiFi communication facilities for their Internet connection. These different types of wireless connections are all proximity-based

connections, i.e., each of them has a distance range of wireless communication. We model this distance range by a function $scope(o)$, which denotes a set of smart objects that are currently accessible by an smart object o . This set $scope(o)$ may also include other smart objects that are accessible through the Internet if the smart object o is currently connected to the Internet. Therefore, the function $scope$ does not directly correspond to the wireless communication range of each object, but defines its current accessibility to other smart objects.

Each smart object provides and/or requests services. A service executed by a smart object may request another service provided by another smart object. For a smart object to request a service running on another smart object, it needs to know the id and the interface of the service at least. We may assume that each service is uniquely identified by its service type in its providing smart object. Therefore, each service can be identified by the concatenation of the object id of its providing smart object and its service type. The interface of a service can be modeled in various different ways. Here we model it as a set of attribute-value pairs without any duplicates of the same attribute. We call each attribute and its value respectively a signal name and a signal value.

Now we need to consider the pluggability between smart objects. If a smart object needs to explicitly specify, in its code, the identifier of the service it wants to access, then this smart object cannot federate with any other smart object providing the same type of services. The substitutability of each smart object in an arbitrary federation with another one providing the same type of service is called the pluggability of smart objects. Pluggable smart objects cannot specify its service request by explicitly specifying the service id. Instead, they need to specify the service they request by its name, by its type, or by its property. The conversion from each of these three different types of reference to the service id is called 'resolution'. Service-name resolution converts a

given service name to a corresponding service id. Service-type resolution converts a given service type to the service ids of services of this type. Service-property resolution converts a given service property to the service ids of services satisfying this property.

When a smart object can access the Internet, it can ask a repository-and-lookup service to perform each resolution. The object id and service type obtained as a resolution result are used to access the target smart object and its service. This requires another resolution service that converts each object id to its global network address, such as url and mobile phone number. This resolution is outside of our mathematical model. Here we will not distinguish object ids and urls. When a smart object can access others only through peer-to-peer network, it must be able to ask each of them to perform each resolution. In this case, if a target smart object can be accessed only through peer-to-peer network, then it needs to perform each resolution by itself. Therefore, such a service-providing smart object that has no Internet accessibility and might be accessed by another without Internet accessibility needs to have a resolution mechanism for its services. Here we assume that every smart object has a resolution mechanism in itself. Some smart objects can perform only object-name resolution and service-type resolution for their own services. Service-property resolution is too heavy to implement in some primitive smart objects.

In our modeling of smart objects, we take a minimalist approach. In other words, we will start with a minimum model of primitive smart objects and primitive federation operations, and then try to describe complex federation mechanisms and a wide range of applications by combining primitive smart objects and primitive federation operations.

In practical situations, the same smart object may provide more than one service of the same function through different access protocols. In order to hide the difference of access protocols from our model,

we treat services of the same function accessible through different protocols as services of different service types. Now as minimalists, we can model each primitive smart object to have only the following three functions, i.e., recognition of accessible smart objects, object-name resolution, and service-type resolution. The first one is represented by the function $scope(oid)$, the second by $onameRes(oid, oname)$, and the third by $stypeRes(oid, stype, mode)$. The last two functions are defined as follows:

$$onameRes(oid, oname)$$

= if $oname$ is the name of the smart object identified by oid then oid else nil.

$$stypeRes(oid, stype, mode)$$

= when $mode = 1$, if the smart object identified by oid provides a service of $stype$ then oid else nil, otherwise if the smart object identified by $oids$ requests a service of $stype$ then oid else nil,

When a service-requesting smart object o requests a service, it sends an object name $oname$ or a service type $stype$ to each smart object oid in its proximity represented by $scope(o)$. Each recipient smart object oid , when receiving $oname$ or $stype$ from o , performs either $onameRes(oid, oname)$ or $stypeRes(oid, stype, 1)$. The sender smart object receives either the recipient's oid or nil depending on the success of the resolution. Such resolution is called providing resolution since the resolution is performed by smart objects providing a service to discover. When a service-providing smart object o with $oname$ as its name searches for another smart object requesting a service of type $stype$ that is provided by o , it sends its object name $oname$, or the service type $stype$ to each smart object oid in its proximity represented by $scope(o)$. Each recipient smart object oid , when receiving $oname$ or $stype$ from o , performs either $onameRes(oid, oname)$ or $stypeRes(oid, stype, 0)$. Such resolution is called requesting resolution since the resolution is performed by smart objects requesting a specific service. Our minimalist's model represents each resolution mechanism as a port. Each smart object is modeled as a set of ports. Each port consists of

a port type and its polarity, i.e., either a positive polarity ‘+’ or a negative polarity ‘-’. Providing resolution is represented by a positive port, while requesting resolution is represented by a negative port. Object-identifier resolution and object-name resolution are respectively represented by port types *oid* and *oname* with polarities. A smart object with *oid* and *oname* has ports *+oid* and/or *+oname* if it has respectively an object-identifier-resolution function and/or an object-name-resolution function. A smart object has ports *+oid* and/or *+oname* if it requests another smart object identified by *oid* and/or *oname*. Service-type resolution is represented by a port with a port type *stype* and a polarity. If it is a providing resolution, then it is represented by *+stype*, else by *-stype*. Each port represents both a request and the capability of corresponding resolution. Therefore, if a smart object enters the proximity of another smart object, and if they have ports with the same port type and different polarities, one can send a resolution request to another to obtain the partner object id successfully. This means that each request for a specific smart object or a specific service is satisfied by the partner smart object. This step is called the port type matching in our model. Ports are mathematically defined as follows. Let *O*, *N*, and *S* denote respectively the set of object identifiers, the set of object names, and the set of service types. The set *P* of ports is defined as follows:

$$P = \{+, -\} \times (O \cup N \cup S).$$

Each smart object *o* has a set of ports *port(o)* which is a subset of *P*. A smart object with *oid* and *oname* may have *+oid* and *+oname* as its ports, but neither of them is mandatory. Federation of a smart object *o* with another smart object *o'* in its scope *scope(o)* is initiated by a program running on *o*. This program detects either a specific user operation on *o*, a change of *scope(o)*, or some other event on *o* as a trigger to initiate federation. The initiation of federation by a smart object *o* with *o'* performs the port matching between the ports of *o* and the ports of *o'*. As its result, every paired ports are connected by a channel identified

by their shared port type. We define a set *channel(o, o')* of port types as a set of all the channels established by a federation of a smart object *o* with *o'*.

The same smart object may be involved in more than one different federation, which implies that the same port may be involved in more than one different channel.

3. Port Signals

When a channel is established between two smart objects, they can communicate with each other through this channel. Each of these smart objects assumes a constraint on a set of I/O signals going back and forth through this channel. Their assumed constraints on signal sets must be compatible with each other. Otherwise, they cannot communicate with each other through this channel. The assumed constraint on a set of signals for a channel, and hence for a port, in each smart object is called the port signal constraint. Mathematically, a port signal constraint is represented by a set of signals, each of which consists of a signal mode, a signal name and a signal domain. By *signal(p)* is denoted a set of signals of the port *p*. Without loss of generality, we may assume that each signal is given its value either by a service-providing port or a service-requesting port. Bilateral signals can be treated as a pair of unilateral signals. A signal mode is ‘+’ if the signal value is given by the service-providing port. Such a signal is called a providing signal. Otherwise, i.e., if the signal value is given by the service-requesting port, the signal mode is ‘-’. Such a signal is called a requesting signal. The domain of a signal is a set of allowed signal values of this signal. For a signal *s*, we denote its mode, name and domain by *sigMode(s)*, *sigName(s)*, and *sigDomain(s)*.

Example 1

PDA1 = {-DBaccess}
signal(-DBaccess)
= {(-query: SQLstandard), (-search:Boolean), (+result:RecoList)}

A pair (-query:SQLstandard) denotes that

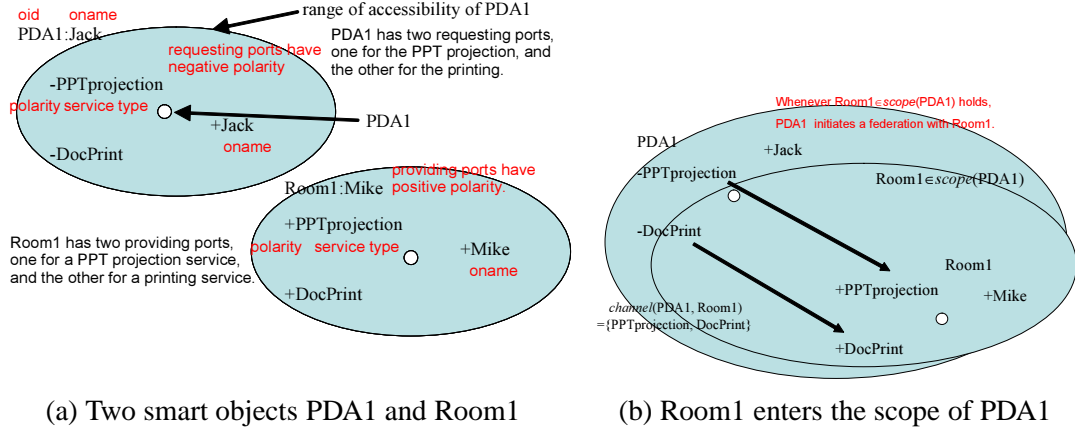


Figure 1: Federation of PDA1 with Room1 and the channels established between them.

this signal value is given by a service-requesting port, and that it has ‘query’ and SQLstandard (the domain of SQL Standard queries) as its name and domain. This service-requesting port -DBaccess issues a query in SQL Standard and a Boolean search command to obtain a search result as a record list.

```

Room1 = {+DBaccess}
signal(+DBaccess) =
    {(-query,: SQL),
     (-search:Boolean),
     (+result:RecorList),
     (+currentRecord:Record),
     (-nextRecord, Boolean),
     (-previousRecord, Boolean)}

```

This service-providing port +DBaccess receives a query in Oracle SQL, a Boolean search command, and a Boolean next-record or previous-record command, and returns a search result as a record list and a current record pointed to by a record cursor that is moved back and forth by the next-record and previous-record command. Now, we will define the compatibility of ports. A service-requesting port -p and a service-providing port +p are compatible with each other and denoted by $-p \sim +p$ if the following holds:

$$\begin{aligned}
 &(\forall(-x,y) \in \text{signal}(-p) \\
 &\quad \exists(-x,z) \in \text{signal}(-p) \ y \subseteq z) \\
 &\forall(\forall(+x,y) \in \text{signal}(-p) \\
 &\quad \exists(+x,z) \in \text{signal}(+p) \ y \supseteq z).
 \end{aligned}$$

In example 1, the port -DBaccess and the port +DBaccess are compatible with each other, i.e., $-DBaccess \sim +DBaccess$ since it holds that $\text{SQLstandard} \subseteq \text{OracleSQL}$.

4. Applications Running on a Smart Object

As shown in Figure 2, each smart object has some ports, internal variables corresponding to internal registers, and HCI variables corresponding to the console input and output variables for its user to input commands and to observe its response. Applications running on a smart object can input values from internal variables, HCI input variables, providing signals of its service-requesting ports, and requesting signals of its service-providing ports. They can output values to internal variables, HCI output variables, requesting signals of its service-requesting ports, and providing signals of its service-providing ports. Figure 2 shows two applications in a smart object and their I/O values.

Each application program is invoked independently from others by the requesting signals of its dedicated service-providing port. Application programs in the same smart object are therefore asynchronously interoperate with each other through the internal variables of the smart object. In general, a smart object requests values from other smart objects through its requesting ports, and provides values computed by its applications to other smart objects through its providing

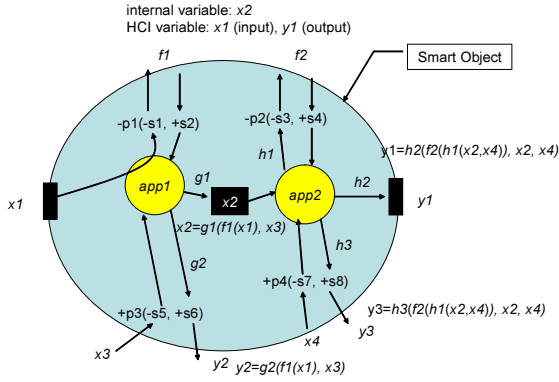


Figure 2: Applications running on a smart object.

ports. Two important classes of smart objects are providing-only smart objects called supplier objects and requesting-only smart objects called consumer objects. Supplier objects have providing ports only. For example, bar codes and RFID tag chips in consumer products, Web applications, or Web services are supplier objects. Consumer objects have requesting ports only. For example, remote controllers of all kinds are consumer objects. In general, a smart object cannot be copied. However, there are important classes of objects where this is virtually possible by copying their proxy objects. Most notably this is possible for software smart objects we will mention later.

A service-requesting port $p1$ is replaceable with another service-requesting port $p2$ if the following holds:

$$\begin{aligned}
 & (\forall (-x,y) \in \text{signal}(-p1) \\
 & \quad \exists (-x,z) \in \text{signal}(-p2) \ y \supseteq z) \\
 & \vee (\forall (+x,y) \in \text{signal}(-p1) \\
 & \quad \exists (+x,z) \in \text{signal}(-p2) \ y \subseteq z).
 \end{aligned}$$

A service-providing port $p1$ is replaceable with another service-providing port $p2$ if the following holds:

$$\begin{aligned}
 & (\forall (-x,y) \in \text{signal}(+p1) \\
 & \quad \exists (-x,z) \in \text{signal}(+p2) \ y \subseteq z) \\
 & \vee (\forall (+x,y) \in \text{signal}(+p1) \\
 & \quad \exists (+x,z) \in \text{signal}(+p2) \ y \supseteq z).
 \end{aligned}$$

5. Semantics of Federation

Let us consider a federation among three smart objects O_0 , O_1 , and O_2 as shown in Figure 3. O_1 and O_2 perform addition and multiplication, while O_0 compose these two functions provided by O_1 and O_2 to calculate $(s1+s2) \times s2$ as the value of $s3$. We need to define the semantics of smart objects and their federation so that we can calculate the composed function from the functions of participating smart objects. Here we will describe the federation semantics using a Prolog like description of the function of each smart object.

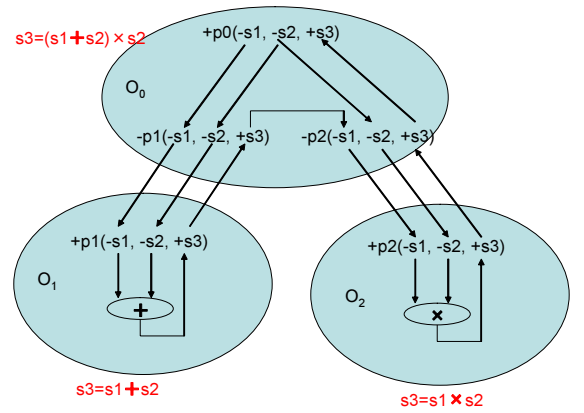


Figure 3: Federation among O_0 , O_1 , and O_2 .

The function of each smart object in Figure 3 is described as follows in our description.

$$\begin{aligned}
 O_0: & \text{p0}(\{-s1:x, -s2:y, +s3:z\}) \\
 & \leftarrow \text{p1}(\{-s1:x, -s2:y, +s3:w\}), \\
 & \text{p2}(\{-s1:w, -s2:y, +s3:z\})
 \end{aligned}$$

($\text{p0}(\{-s1:x, -s2:y, +s3:z\})$ is implied by $\text{p1}(\{-s1:x, -s2:y, +s3:w\})$ and $\text{p2}(\{-s1:w, -s2:y, +s3:z\})$.)

$$O_1: \text{p1}(\{-s1:x, -s2:y, +s3:z\}) \leftarrow [z=x+y]$$

($\text{p1}(\{-s1:x, -s2:y, +s3:z\})$ is implied by the evaluation of $z=x+y$.)

$$O_2: \text{p2}(\{-s1:x, -s2:y, +s3:z\}) \leftarrow [z=x \times y]$$

The functions of a smart object is described by a set of rules. Each rule has zero or one literal on the left-hand side, and zero or an arbitrary number of literals on the right-hand side. Each literal on the left-hand side corresponds to a service-providing port, while each literal on the right-hand side corresponds to either a program code to evaluate

or a service-requesting port. Each port may have a set of pairs, each of which consists of a signed signal name and its value. A signal value may be represented by a variable. Similarly to the logical resolution in Prolog, we start with the evaluation of a given goal with arbitrary number of literals only on the right-hand side:

$$\leftarrow L1, L2, \dots, Ln.$$

Such a rule without any left-hand side literal is called a goal. The first literal Li that is not a program code in this goal is logically resolved with some literal with the same port type that appears on the left-hand side of some other rule, and is replaced with the right-hand side of this rule with all the variables changed to the expressions that unifies these two literals. Let a rule

$$Li' \leftarrow L1', L2', \dots, Lm'$$

be such a rule. Each port literal has a set of signal pairs, each of which consist of a signal name and a value, whereas a predicate in Prolog has parameters, the number of which is fixed for the same predicate name. A right-hand side literal $Li = p(S)$ can be unified with a left-hand side literal $Li' = p(S')$ if the following holds:

$$\exists \sigma \exists \sigma' (\forall (x:y) \in S \exists (x:z) \in S' \sigma y = \sigma' z),$$

where σ and σ' are substitutions of variables. A pair (σ, σ') is called a unifier. Generally, such a unifier is not uniquely determined. In such a case, there exists the most general unifier (σ_0, σ'_0) . Using these substitutions, the current goal is replaced with

$$\begin{aligned} \leftarrow & \sigma_0(L1), \sigma_0(L2), \dots, \sigma_0(L(i-1)), \\ & \sigma'_0(L1'), \sigma'_0(L2'), \dots, \sigma'_0(Lm'), \\ & \sigma_0(L(i+1)), \dots, \sigma_0(Ln). \end{aligned}$$

This operation is the extension of Prolog resolution mechanism to the resolution of our rules. When the service-providing port $p0$ of O_0 is accessed with two signal values a

and b from outside, O_0 starts to evaluate a goal

$$\leftarrow p0(\{-s1 : a, -s2 : b, +s3 : z\}),$$

which is resolved by the rule in O_0 , and is reduced to

$$\begin{aligned} \leftarrow & p1(\{-s1 : a, -s2 : b, +s3 : w\}), \\ & p2(\{-s1 : w, -s2 : b, +s3 : z\}). \end{aligned}$$

Then the first literal is resolved by the rule in O_1 , and the second by the rule in O_2 , since the port $-p1$ and $-p2$ in O_0 are respectively connected to the ports $+p1$ in O_1 and $+p2$ in O_2 through channels, i.e.,

$$\begin{aligned} \leftarrow & [w := a + b], p2(\{-s1 : w, -s2 : b, +s3 : z\}) \\ \leftarrow & [w := a + b], [z := w \times b]. \end{aligned}$$

This result has only program codes, and can be partially evaluated to obtain the following:

$$\leftarrow [z := (a + b) \times b].$$

This is the function composed by the federation of these three smart objects.

Example 1 revisited

The semantics of PDA1 and Room1 in Example 1 is described as follows:

PDA1

App(x) \leftarrow

DBservice($\{-query:q, -search:true, +result:x\}$).

Room1

DBservice($\{-query:x, -search:y, +result:z, +currentRecord:u, -nextRecord:v, -previousRecord:w\}$)

$\leftarrow [PDB(x, y, z, u, v, w)].$

App(x) is an application running on PDA1, while PDB(x, y, z, u, v, w) is a program code executed by Room1 to access its database to answer a given query. Whenever PDA1 starts to evaluate a goal

$\leftarrow App(x),$

it is reduced as follows:

$$\begin{aligned} \leftarrow & DBservice(\{-query:q, -search:true, +result:x\}) \\ \leftarrow & [PDB(q, true, x, u, v, w)]. \end{aligned}$$

Example 2

Let us consider federations among a PDA PDA2,

a video cassette recorder VCR1, and a monitor display MonitorDisplay1, where PDA2 works as a remote controller of VCR1, and the audio and video signals of VCR1 are sent to MonitorDisplay.

```
VCR1={-VideoDisplay, +VideoControl}
  signal(-VideoDisplay)
    ={(video: {NTSC}),(audioL: Audio),
      (audioR: Audio)}
  signal(+VideoControl)
    ={(play: Boolean), (ff: Boolean),
      (rewind: Boolean), (record: Boolean),
      (stop: Boolean)}
```

```
MonitorDisplay1={+VideoDisplay}
  signal(+VideoDisplay)
    ={(video: {NTSC, PAL}),
      (audioL: Audio),
      (audioR: Audio)}
```

```
PDA2={-VideoControl}
  signal(-VideoControl)
    ={(play: boolean), (ff: boolean),
      (rewind: boolean),(stop: boolean)}
```

Their semantics is described as follows:

VCR1

```
VideoControl({-play:x1, -ff:x2, -rewind:x3,
  -record:x4, -stop:x5})
  ← [PVCR(x1, x2, x3, x4, x5, y1, y2, y3)],
  VideoDisplay({-video:y1, audioL:y2,
  audioR:y3})
```

When VCR1 receives a service request through the port +VideoControl, it performs its operation depending on the input command signals. If the command signal is 'play', it plays the loaded video tape and sends out its audio-video output signals through the service-requesting port -VideoDisplay. If the command is 'ff' i.e., 'fast forward', then it fast-forward the tape until another command signal is input, and sends out the fast-forward mark as its video output signal through the port -VideoDisplay. It similarly treats other command signals, 'rewind', and 'stop'. If the command is 'record', it starts the recording of its input video signal onto the loaded tape, and sends out the input video signal combined with the recording mark as its audio-video output signals through the port -VideoDisplay. The first literal [PVCR(x1, x2, x3, x4, x5, y1, y2, y3)] on the right-hand side of this rule represents the internal mechanism of VCR1.

MonitorDisplay1

```
VideoDisplay({-video:y1, audioL:y2,
  audioR:y3})
  ← [PMonitor(y1, y2, y3)]
```

The literal on the right-hand side represents the video displaying function of this monitor display.

PDA2

```
HCIPDA({-key1:x1, -key2:x2,
  -key3:x3, -key5:x4, -key5:x5})
  ← VideoControl({-play:x1, -ff:x2,
  -rewind:x3, -stop:x5})
```

The literal on the left-hand side represents the human-computer interaction interface of this PDA. This rule uses only five keys of PDA2, key1, key2, key3, key4, and key5 for its user to input the five commands, i.e., 'play', 'ff', 'rewind', 'record' and 'stop'. When one of these keys is pushed, PDA2 sends a service request with the corresponding command signal through the port -VideoControl.

Let us assume that the two federations, i.e., PDA2 with VCR1, and VCR1 with MonitorDisplay1, are both established. In order to know the composed function of each key input to PDA2 in these federations, we can evaluate the following goal:

```
← HCIPDA({-key1:x1, -key2:x2,
  -key3:x3, -key5:x4, -key5:x5})
```

This is reduced as follows:

```
← VideoControl({-play:x1, -ff:x2,
  -rewind:x3, -stop:x5})
← [PVCR(x1, x2, x3, x4, x5, y1, y2, y3)],
  VideoDisplay({-video:y1,
  audioL:y2, audioR:y3})
← [PVCR(x1, x2, x3, x4, x5, y1, y2, y3)],
  [PMonitor(y1, y2, y3)].
```

6. Service-Property Resolution Service

As mentioned in Chapter 2, our model does not treat service-property resolution as a primitive function. Instead, it treats this type of resolution as a service. Therefore, smart objects with this type of resolution function are modeled to have +sPropertyResolution port, whereas those capable to issue a query to another smart object for service-property resolution are modeled to have -sPropertyResolution port. These ports may

have arbitrary number of signals specifying conditions on different attributes of the service property since there may be a large variety of different types of service-property resolution. The service-property resolution mechanism is defined as follows:

```
sPropertyResolution({attr1:x1, ..., attrk:xk,
  stype:y, signal1:y1, ..., signalh:yh})
← [find({attr1:x1, ..., attrk:xk,
  stype:y, signal1:y1, ..., signalh:yh}),
  openNewPort(+y)]
```

When a channel is set up between `-sPropertyResolution` and `+sPropertyResolution`, a smart object with `-sPropertyResolution` port sends the service property information $\{attr1:x1, attr2:x2, \dots, attrk:xk\}$ through this channel to another smart object with `+sPropertyResolution` port to find out a service satisfying this property, and to create a new port with y as its service type, and with $signal_1, signal_2, \dots, signal_h$ as its signals. This newly created port enables the service-requesting smart object to set up another channel from its preset port `-y` with h signals $signal_1, signal_2, \dots, signal_h$ to the newly created port of the service-providing smart object.

A smart object with a service-property resolution service can search all the services it provides for the services satisfying the specified property condition, and provides one of these services through the API presumed by the requesting smart object. Figure 4 shows an implementation of service-property resolution as a service provided by a smart object Room. PDA requests Room for a service through a presumed port `-r`. Room does not initially provide a compatible service-providing port `+r`. PDA asks Room for service-property resolution, which makes Room create a new port `+r`.

7. Aggregate Federation

When a service-requesting smart object with a port `-p` federates with more than one service-providing smart object with `+p` port, more than one channel of the service type p are simultaneously established. The service-

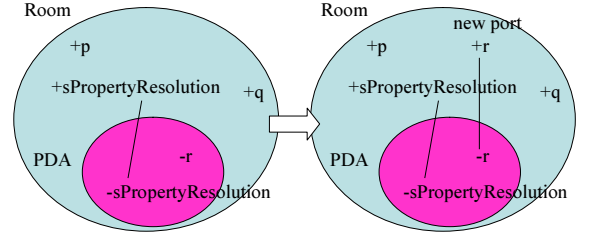


Figure 4: Service-property resolution implemented as a service.

requesting port may use one of these channels or scan these channels to access each of these services, depending on the application program using this port. Figure 5 shows a case of scanning all the channels, where PDA with a sensor reader simultaneously federates with more than one sensor.

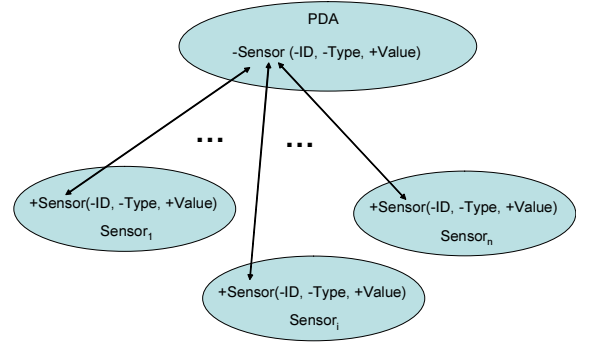


Figure 5: Aggregate federation between a PDA and more than one sensor.

The semantics of these ports are described as follows. The PDA has an application to retrieve values from all the sensors of a specified type, and to analyze this value set to obtain some index value such as the average of them.

PDA

```
Analysis({type:x, index:y})
← [y:=analysis(z), z:=bag{u}]
  Sensor({ID:v, type:y, value:u})
```

$Analysis(type:x, index:y)$ is implied by the evaluation of $y:=analysis(z)$ and $z:=bag\{u\}$, where $Sensor(\{ID:v, type:y, value:u\})$ is presumed. The vertical bar denotes that the followings constitute a where clause.

Each sensor, on the other hand, first checks if its sensor type is equal to the specified

type, and, if ‘yes’, return its current sensor value.

Sensor

Sensor($\{ID:v, type:y, value:u\}$)

$\leftarrow [u:= \text{if } y=\text{sensorType}(v) \text{ then}$
 $\quad \text{currentSensorValue}(v) \text{ else nil}]$

Sensor($\{ID:v, type:y, value:u\}$) is implied by the evaluation of $[u:= \text{if } y=\text{sensorType}(v) \text{ then currentSensorValue}(v) \text{ else nil}]$.

The composed function of Analysis($type:x, index:y$) is obtained by evaluating the following goal:

$\leftarrow \text{Analysis}(\{type:x, index:y\})$.

This can be reduced as follows:

$\leftarrow [y:=\text{analysis}(z), z:=\text{bag}\{u\}]$
 $\quad \text{Sensor}(\{ID:v, type:y, value:u\})$
 $\leftarrow [y:=\text{analysis}(z), z:=\text{bag}\{u\}]$
 $\quad [u:= \text{if } y=\text{sensorType}(v) \text{ then}$
 $\quad \quad \text{currentSensorValue}(v) \text{ else nil}]$.

This composite function accesses every sensor to check if its type is the specified one, gets the values of all the sensors of the specified type, and analyzes this value set to obtain the value y .

8. Delegation of Access Resolution

Some smart object can delegate an access resolution it is requested to other smart objects within its scope. The requesting smart object in such a case needs to explicitly specify if it actually request such delegation. Therefore, we can model such access-resolution delegation mechanism as a service. A smart object with a resolution delegation service is modeled to have a special port $+ResDelegation$, while one requesting a resolution delegation service is modeled to have $-ResDelegation$ port. Once a channel is established between these two ports, the latter smart object can ask the former smart object to generate a pair of $-p$ and $+p$ ports in itself for a specified $-p$ port in the latter. Figure 6 shows such an example.

This mechanism is defined as follows.

PDA

$\text{Appl}(x) \leftarrow L1, \text{ResDelegation}(\{ptype:p,$
 $\quad \text{signals:signal_information_about_p}\})$,

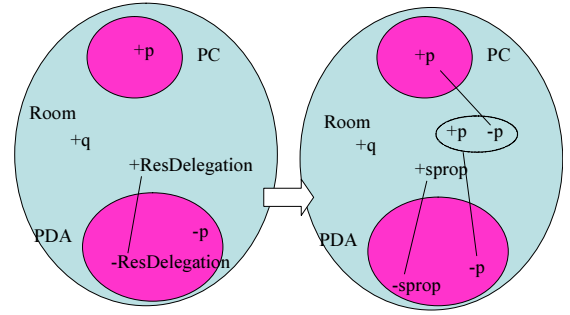


Figure 6: Resolution delegation.

$p(S(x)), L2$.

Room

$\text{ResDelegation}(\{ptype:x, signals:y\})$

$\leftarrow [\text{openPortPair}(x, y)]$

$p(x) \leftarrow p(x)$.

An application Appl in PDA invokes $L1$ and requests ResDelegation for the port type p showing the signal information about p . Using the ResDelegation result, it invokes $p(S(x))$, and $L2$. $S(x)$ here denotes the set of signals of the port p which may depend on a variable x . Room, on the other hand, delegates a resolution request for a port p with a port signal information about p by creating, in itself, a port pair $-p$ and $+p$ having the signals specified by the port signal information about p . The last rule above has $p(x)$ on both sides. However, different from Prolog, $p(x)$ on the left-hand side and the $p(x)$ on the right-hand side are different from each other. The former is a service-providing port, while the latter is a service-requesting port.

Room in this example may also ask another smart object for resolution delegation. In such a case, its semantics becomes as follows:

$\text{ResDelegation}(\{ptype:x, signals:y\})$

$\leftarrow [\text{openPortPair}(x, y)]$,

$\quad \text{ResDelegation}(\{ptype:x, signals:y\})$

$p(x) \leftarrow p(x)$.

9. Web Services and Software Smart Objects over the Web

Our model treats each WiFi access point as a smart object. Once a smart object federates with some access point that provides an

Internet access service, it can access whatever Web services this access point is given permission to access if it knows their url addresses. Now we will extend our model so that smart objects may access Web services through an access point. We will model such a function of each access point *apoint* as follows:

```
ResDelegation({ptype:x,signals:y})
  ← isURL(x), permitted(apoint, x),
    [createProxy(x, y)]
```

The procedure `createProxy(x, y)` creates a port `+x` with signals specified by `y` as a port of the access point. Such a port is called a proxy port since it works as a proxy object to communicate with a remote Web service. Another way for an access point to provide the accessibility to such a Web service is to provide its proxy port from the very beginning.

Now let us expand our model so that we can deal with distributed software objects over the Internet as software smart objects. Each software smart object has its scope including one or more than one repository-and-lookup service on the Internet. Such a repository-and-lookup service available at an address `url0` is denoted by `RepositoryLU(url0)`. The semantics of such a service is described as follows:

```
RepositoryLU(url)
url0({operation : 'register', ptype : x,
      signals : y, smartObject : z})
  ← [register(x, y, z)]
url0({operation : 'disenroll',
      ptype : x, smartObject : y})
  ← [disenroll(x, y)]
url0({operation : 'lookup', ptype : x,
      signals : y, smartObject : z})
  ← [lookup(x, y, z)]
url0({operation : 'proxy', ptype : x,
      signals : y, smartObject : z})
  ← [createProxy(x, y, z)]
```

A smart object with `url0` in its scope can directly access this service. When this service is accessed with 'register' as its operation, it registers the requesting smart object with its specified port type and port signals.

These registered information is used to create a proxy port of the same type and signals in this repository-and-lookup service to access this port in the registering smart object. When the service is accessed with 'disenroll' as its operation, it disenrolls the specified port and smart object. When the service is accessed with 'lookup' as its operation, it looks up the repository for registered ports of the given port type `x` and with the given signals `y`, and finds out a software smart object `z`. When the service is accessed with 'proxy' as its operation, it creates, in itself, a proxy port of type `x` that works as a proxy to access a software smart object `z` using signals `y`.

A repository-and-lookup service may ask another repository-and-lookup service to find out an appropriate software smart object and its port. In this case, the semantics of this repository-and-lookup service needs to add the following rule as the last rule for

```
url0({operation : 'lookup', ptype : x,signals : y}) :
  url0({operation : 'lookup',
        ptype : x, signals : y})
  ← url1({operation : 'lookup',
        ptype : x, signals : y}),
```

where `url1` is the url of another repository-and-lookup service. If there are more than one such service, we just need to add a similar rule for each of them. In order to utilize such a service, the semantics of each software smart object at `url2` is defined as follows.

Software Smart Object at `url2`

```
registration(x, y)
  ← url0({operation : 'register',
        ptype : x, signals : y,
        smartObject : url2})
```

```
Appl(x)
  ← L1,
    url0({operation : 'lookup',
        ptype : p, signals : S,
        smartObject : z}),
    url0({operation : 'proxy',
        ptype : p,signals : S,
        smartObject : z}),
    p(S(x)),L2.
```

The semantics of the operation ‘register’ is obvious. The application invokes $,url_0(\{operation : 'lookup', ptype : p, signals : S, smartObject : z\})$ and $,url_0(\{operation : 'proxy', ptype : p, signals : S, smartObject : z\})$ to create a proxy port of type p that communicates with a software smart object z using signals S , where S denotes a set of pairs, each of which consists of a signal name and a signal domain. $S(x)$ is a set of pairs, each of which consists of a signal name and its value, and denotes that it has the same set of signal names as S , and their values somehow depends on x .

10. Application Frameworks using Smart Objects

10.1. The downloading of a software smart object

A smart object may presume a standard API to request a service of another smart object that does not provide the compatible API but provides a downloadable driver to access this service through the presumed API. Such a mechanism is described as follows. Suppose that a smart object o has a service to download a software smart object o' . This is described as follows:

$$SOdownload(\{query : x, SO : y\}) \\ \leftarrow [find(\{query : x, SO : y\})]$$

A requesting smart object o'' with the download request facility can ask this smart object o to download a software smart object y that satisfies a query x , and install this software smart object y in itself, i.e.,

$$SOdownloadInstall(x) \\ \leftarrow SOdownload(\{query : x, SO : y\}), \\ [install(y)]$$

The installation of a software smart object y by o'' adds y in the scope of o'' , initiates a federation between o'' and y , and makes the scope of y equal to the scope of o'' . Figure 7 shows an example application of the

downloading of a software smart object. In this figure, the downloaded software smart object may work as a driver to access the service $+p$ of the smart object Room through a port $-r$ of the smart object PDA.

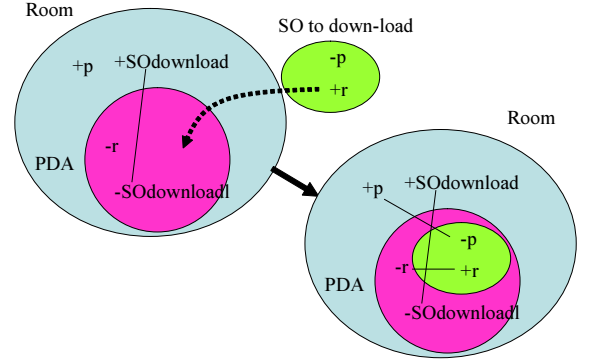


Figure 7: Smart object downloading

If a downloaded smart object is a proxy object to a Web service or another smart object, the recipient smart object can access this remote service through this proxy smart object. This case is shown in Figure 8, where PDA is accessing through its port $-r$ a remote service using a downloaded proxy smart object.

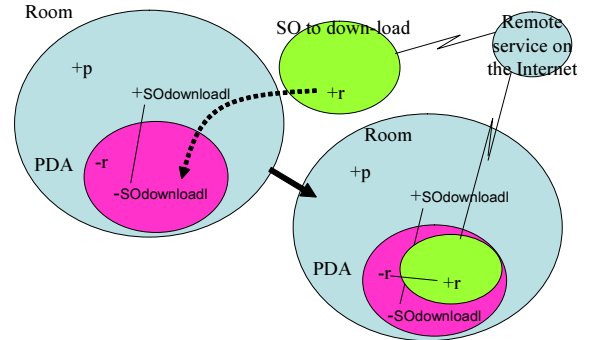


Figure 8: The downloading of a proxy smart object to access a remote service.

10.2. Location-transparent service continuation

Suppose that PDA1 has federated with Office, an access point with a server. Office provides services available at the user’s office. Let Home be another access point with a server providing services available at his or

her home. The location transparent continuation of services means that he or she can continue the job that was started at the office using PDA1 even after he or she goes back home carrying PDA1. This is realized by the following mechanism. When he or she carries out PDA1 from Office environment, he or she just needs to make PDA1 download the proxy smart object of Office as shown in Figure 9. This operation is called federation suspension. When arriving at home, PDA1 is WiFi connected to Home, then federates with Home. At the same time, the proxy smart object installed in it resumes the access to Office. Therefore, they set up one additional channel to +Print port in Home smart object as shown in the right-hand side of this figure. Now the PDA can access the database service of Office, and the two printing services of Office and Home. When PDA1 requests a print service, it asks its user or the application which of these services to choose.

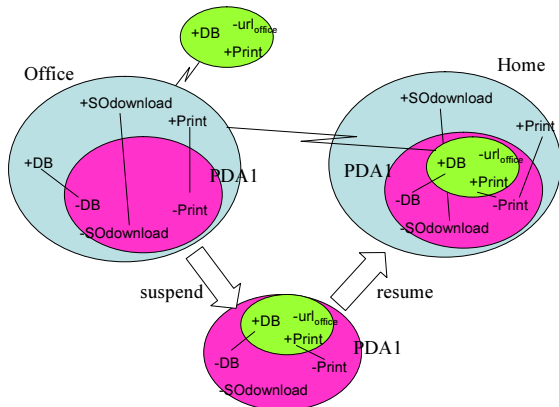


Figure 9: Location-transparent service continuation using the downloading of a proxy smart object.

10.3. Import and export of ports and glue smart objects

Here we consider a special type of port 'export'. A service-providing port +export exports proxies of all the ports except those of export type to the service-requesting smart object that accesses this port through -export. A proxy of a port +p or -p is represented as p^* , and works as a relay

station of +p or -p. A proxy p^* in a smart object o forms a channel with each of +p, -p, and p^* in a different smart object.

A glue smart object is a special smart object with no port other than a service-requesting port -export. A smart object with -p port cannot communicate with another smart object with +p if neither of them is located within the proximity of the other. Let us assume that each of them has a port +export. In this case, if we bring in a glue smart object with a sufficiently large WiFi access range to cover these two smart objects, this glue object can federate with these two, sets up channels between its -export and the +export port of each of the two objects, and obtain a port proxy p^* . This proxy works as a relay station for -p and +p in these two objects, and sets up channels from itself to each of these two ports. Through these channel, the port -p becomes able to communicate with the port +p. This situation is shown in Figure 10.

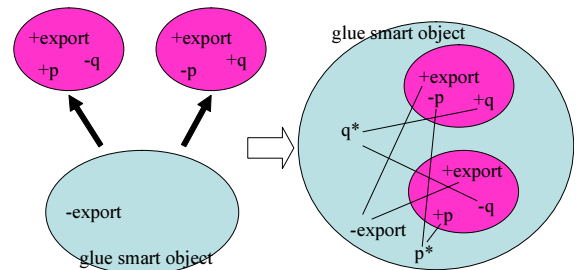


Figure 10: A glue smart object and its usage.

Their semantics is defined as follows. Each smart object that allows the export of its port proxies has the following rule.

$$\text{export}(\{\text{portProxyList} : y\}) \leftarrow [\text{CreatePortProxy}(y)].$$

Each glue smart object has the following rules.

$$\text{federation}() \leftarrow \text{export}(\{\text{portProxyList} : y\}), [\text{createPortProxy}(y)]$$

For each created port proxy p^* , the following rules are added to the glue smart object.

$$p * (x) \leftarrow p(x)$$

$$p * (x) \leftarrow p * (x)$$

Glue smart objects can be hierarchically used to make widely spread smart objects to interoperate with each other as shown in Figure 11. It should be noticed that each glue smart object does not have +export port.

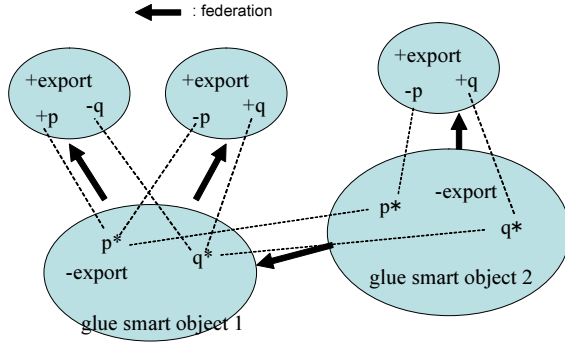


Figure 11: Hierarchical usage of glue smart objects.

10.4. Proxy smart objects for passive tags

Passive tags are accessed by a smart object with a tag reader. When such a smart object o accesses a passive tag, we consider that a new software smart object that works as a proxy of this tag is generated. The scope of this proxy smart object only includes o , while the object o updates its scope to include this proxy. This proxy of a passive tag has a port +tag with signals including the tag id, the tag value, the read signal, and the write signal. The smart object o evaluates the following goal

$$\leftarrow \text{tag}(\{\text{id} : \text{tagid}, \text{value} : v, \text{write} : \text{true}\})$$

to write a new tag value to the tag with *tagid* as its tag id. It reads this tag by evaluating

$$\leftarrow \text{tag}(\{\text{id} : \text{tagid}, \text{value} : v, \text{read} : \text{true}\}).$$

Each proxy of a passive tag has the following rule:

$$\text{tag}(\{\text{id} : x, \text{value} : y, \text{write} : \text{true}\})$$

$$\leftarrow [\text{if } x \text{ is my tag id then write}(y)]$$

$$\text{tag}(\{\text{id} : x, \text{value} : y, \text{read} : \text{true}\})$$

$$\leftarrow [\text{if } x \text{ is my tag id then read}(y)]$$

11. Concluding Remarks

This paper has proposed a new formal model of autonomic proximity-based federation among smart objects with wireless network connectivity and services available on the Internet. Smart objects here may include computing devices such as RFID tag chips, smart chips with sensors and/or actuators that are embedded in pervasive computing environments such as home, office, and social infrastructure environments, mobile PDAs, intelligent electronic appliances, embedded computers, and access points with network servers. Federation here denotes the definition and execution of interoperation among smart objects and/or services that are accessible either through the Internet or through peer-to-peer *ad hoc* communication without previously designed interoperation interface.

This paper first proposed a basic formal model for federation among primitive smart objects, and then extended this model to cope with federation among smart objects through the Internet as well as federation including services over the Web. Each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object.

Here we focused on the matching of service-requesting queries and service-providing capabilities that are represented as service-requesting ports and service-providing ports, instead of focusing on the matching of a service requesting message with a service-providing message. Our abstract modeling of each smart object from the view point of its federation capability has allowed us to discuss both the matching mechanism for federation and complex federation among

smart objects in terms of a simple mathematical model. Our model has enabled us to describe applications from the view point of their federation structures. This has enabled us to extract a common substructure from applications sharing the same typical federation scenario as an application framework for this federation scenario. This paper has also given the semantics of our federation model based on a Prolog like logical framework.

Federation of smart objects also require security mechanism to protect each smart object from unexpected or evil federation. Secure federation requires the integration of some security mechanism with our federation model, on which we are also currently working.

References

- [1] David Gelernter. Generative communication in linda. *ACM Trans. Program. Lang. Syst.*, 7(1):80–112, 1985.
- [2] Sun Microsystems. Javaspaces service specification, version 1.2, 2001.
- [3] Sun Microsystems. Jini technology core platform specification, version 1.2, 2001.
- [4] Gian Pietro Picco, Amy L. Murphy, and Gruia-Catalin Roman. Lime: Linda meets mobility. In *ICSE '99: Proceedings of the 21st international conference on Software engineering*, pages 368–377, Los Alamitos, CA, USA, 1999. IEEE Computer Society Press.
- [5] Yuzuru Tanaka. *Meme Media and Meme Market Architectures: Knowledge Media for Editing, Distributing, and Managing Intellectual Resources*. Wiley-IEEE Press, 2003.
- [6] Yuzuru Tanaka. Knowledge federation over the web based on meme media technologies. In *Lecture Notes in Computer Science*, 3847, pages 159–182, 2006.
- [7] Yuzuru Tanaka, Jun Fujima, and Makoto Ohigashi. Meme media for the knowledge federation over the web and pervasive computing environments. In *ASIAN 2004, Lecture Notes in Computer Science*, 3321, pages 33–47, 2004.
- [8] Yuzuru Tanaka and Kimihito Ito. Meme media architecture for the reediting and redistribution of web resources. In *FQAS 2004: Lecture Notes in Computer Science*, 3055, pages 1–12, 2004.
- [9] Yuzuru Tanaka, Kimihito Ito, and Jun Fujima. Meme media for clipping and combining web resources. *World Wide Web*, 9(2):117–142, 2006.

Protean sensor network for context aware services

Nonaka, H. and Kurihara, M.

Graduate School of Information Science and Technology
Hokkaido University, Sapporo 060 0814, Japan
{nonaka, kurihara}@main.ist.hokudai.ac.jp

Abstract

We present a wearable sensor network system. In our study, we refer a personal area network as “sensor network.” The main part of our system is eye and head movement tracking module based on visual sensorimotor integration. Eye-head cooperation is considered, especially head gesture accompanied with vestibulo-ocular reflex is used for a cue of person’s intention. Each sensor is responsible for respective roll, which varies according to the modality. And the modality depends on the person’s context. In order to change the function of each node, respective firmware is re-programmed by each other using in-circuit serial programming.

Keywords: sensor network, context-aware computing, eye movement tracking, wearable computer, gesture recognition

1. Introduction

Recently, the term “sensor network” has been often used for an ad-hoc wireless network or a multi-hopping network, especially in the research field of wireless communication networks. Huge numbers of inexpensive sensor nodes are deployed geographically or attached to mobile robots [2] [6] [18]. Sensor nodes acquire only local information from their surroundings. However each sensor node has limited communication and computation ability, large scaled sensor network can be constructed by organizing them into collaborative, without a priori sensing infrastructure. Now such a network system is commercialized [3] and becoming available for information acquisition including target tracking, area search, environmental monitoring, infrastructure maintenance, feature localization, border monitoring, battlefield monitoring, disas-

ter relief, and so on.

“Sensor network” is also used as that for extracting or inferring a person’s context, especially in the research field of ubiquitous computing or pervasive computing. The term “context” is defined, for example, as follows:

“Context: any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects.” by Dey et al. [4]

In order to acquire and gather information available for extracting a person’s context, variety of sensors are installed, as is often the case, by invisible way.

In the case of a room, they are embedded in furniture, electronic appliances, the ceiling, the floor, and so on. The range extends to a building, public facilities, station premise, airport, city, and so forth. Various systems and utilities for context-aware computing have been proposed based on such sensor networks [1] [4] [7] [17].

In our study, we refer a personal area sensor network as “sensor network,” however it is used for context-aware services. Context resides intrinsically in each person’s mind, therefore, context-aware services are preferable to be provided by personal belongings, rather than by surroundings. Some people may feel uncomfortable to be taken pictures wherever they go, even if they take pictures wherever. They should prefer GPS as a positioning service to that with satellite images or surveillance cameras. If an ID-tag system is available for a context-aware service, they might desire to take belong a tag reader rather than a tag.

From these considerations, we proceeded to develop a personal area sensor network close to a person. The sensor network is constructed by heterogeneous sensor nodes, for example, pressure sensors [9] [10], acceleration sensors [11], ultrasonic sensors [12]. The main part of our system is eye and head movement tracking module based on visual sensori-motor integration. Eye-head cooperation is considered, especially head gesture accompanied with vestibulo-ocular reflex is used for a cue of person’s intention. We adopted a position sensitive device (S7848, Hamamatsu) for eye tracking. Accelerometer (ADXL311E, Analog Devices), gyro sensor (ADXRS300, Analog Devices), and magnetic sensor (AMI201, Aichi Micro Systems) are used for head tracking. Each sensor is responsible for respective roll, which varies according to the modality.

And the modality depends on the person’s context, for example, eye-gaze is used for not only seeing or looking, but also signaling or directing a partner. In order to change the function of each node, respective firmware is re-programmed by each other using in-circuit serial programming.

2. Network configuration

Heterogeneous sensor nodes are interconnected by both wired and wireless communication links. Neighboring nodes are mainly linked single-wire (two-core) connections overlapped with power supply. For inter-PCB connection, we adopted four-core equilibrium half-duplex transmission that is compliant with RS-422, with the view of reducing radiated electromagnetic interference. Wireless links are used for external communications with surroundings. For simplicity, only one node is allowed to drive the transmission line in any moment, and the other nodes can only receive data, therefore, carry sense or collision detection are not required. The network performs at a time in one of four modes: *normal-packet-mode*, *cascaded-packet-mode*, *pulsed-network-mode*, and *programming-mode*.

In the *normal-packet-mode*, each packet is exchanged between host node and a target node with corresponding addresses (Figure 1). The network initially starts in this mode, and it enters the other modes using packet exchange in this mode.

In the *cascaded-packet-mode*, each packet is transmitted from node to node successively in a predefined order with the subsequent address (Figure 2). While the period of transmission, every traffic is observed by host node. Broadcasting message and gathering sensor data are usually achieved in this mode.

In the *pulsed-network-mode*, the transmission line is allowed in a given period

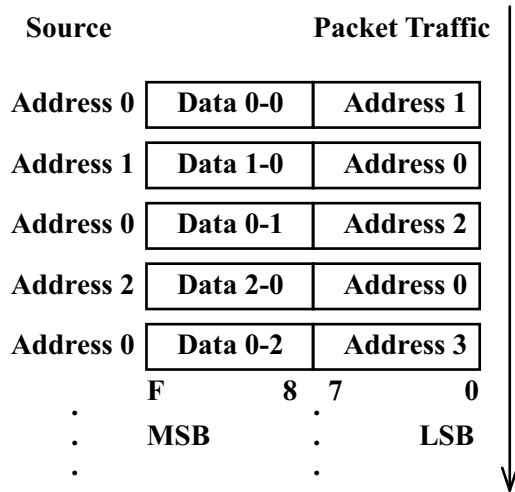


Figure 1: An example of packet traffic in normal-packet-mode. Each packet is exchanged between host node and a target node.

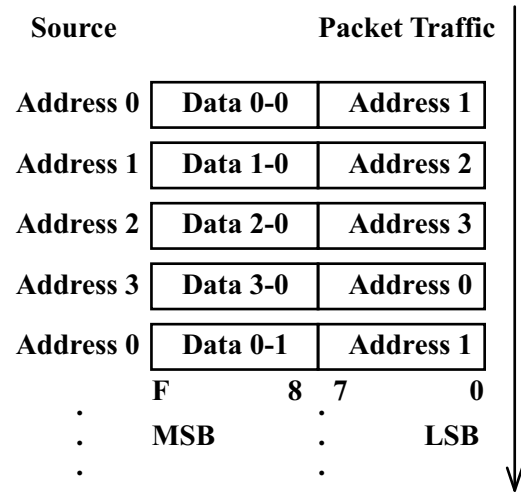


Figure 2: An example of packet traffic in cascaded-packet-mode. Each packet is transmitted successively in a predefined order with the subsequent address.

to use for multi-directional links of pulse frequency data. It is a simple representation of pulsed neural network or spiking neural network[8]. This mode is used especially for extraction of user’s intention from multi-modal gesture, such as eye-head cooperative motion. An example of operations in pulsed-network-mode is presented in section 4.

In the *programming-mode*, the processes of specific nodes or whole of the network are temporarily stopped and its firmwares are reprogrammed by the capability of in-circuit serial programming. It becomes possible by using dual-MCU (Micro Controller Unit) in the constitution of each sensor node. The one (main MCU) is assigned to gathering sensor data and executing signal processing, while the other (communication MCU) is in charge of communication with other nodes and reprogramming the main MCU in obedience to the demand of other node. This mode is used for upgrade of system version, change of user, modification of the role of the system, addition or removal of certain nodes, transition of user’s context

or environment, and so on.

3. Configuration of sensor node

Examples of hardware configurations of sensor nodes are shown in Figures 3 - 6. The common part of each node is mainly constituted of dual-MCU (Micro Controller Unit): main MCU, and communication MCU. The former is used for obtaining sensor data and signal processing, and the latter is used for communication and in-circuit serial programming, as is mentioned in the last section.

The first example is the node with 2-axis acceleration sensor (Figure 3). By each analog value of x-y axis, the gradient from the direction of gravitational acceleration is measured. It is mainly used for measuring tilt of head and a motion of nodding head. The measured data also involve dynamic acceleration, e.g. horizontal acceleration and vibration. Theoretically, it is impossible to eliminate such dynamic noises from the static acceleration derived from gravity, but under the condition that the motion is human motion, it is possible to some extent, by us-

ing the method of time-frequency analysis. The detail of the processing is mentioned in Section 4.

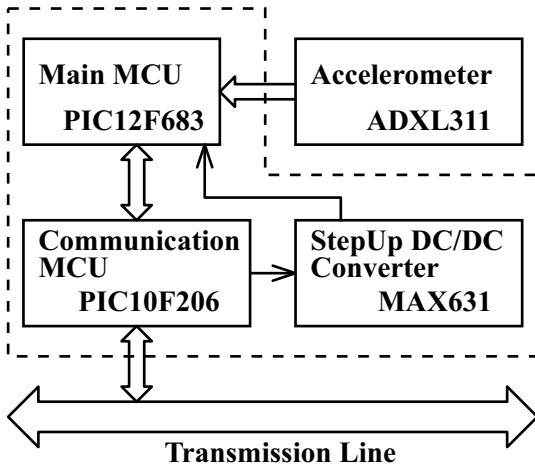


Figure 3: Hardware configuration of sensor node with 2-axis acceleration sensor.

The second example is the node with single-axis yaw-rate gyro sensor (Figure 4). Using the value of angular velocity about the axis normal to the top surface, relatively quick response of rotation can be obtained. It is mainly used for measuring turn of head and a motion of shaking head.

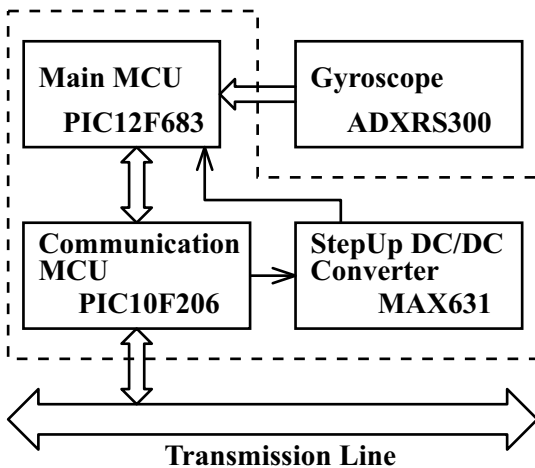


Figure 4: Hardware configuration of sensor node with 1-axis yaw-rate gyro sensor.

The third example is the node with 2-dimensional magnetic sensor or elec-

tronic compass (Figure 5). Measuring the direction of earth magnetism, static direction can be obtained. It is used for measuring the orientation of head and the compensation of the data acquired by gyro sensor.

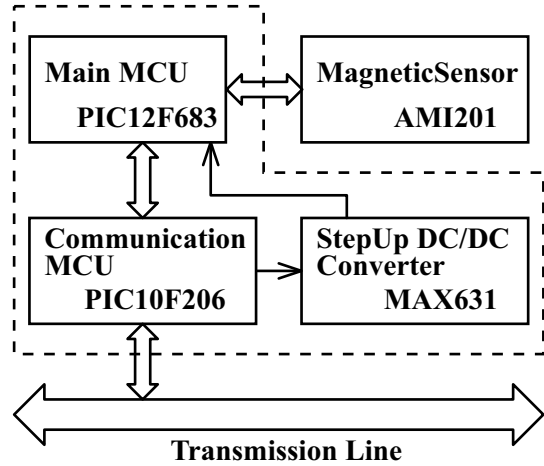


Figure 5: Hardware configuration of sensor node with 2-dimensional magnetic sensor.

The last example is the node with 2-dimensional position sensitive device and Infrared LED (Figure 6). It is used for eye movement tracking. Various methodologies for eye movement measurement have been proposed since early times, for example, electro-oculography (EOG), video-based oculography (VOG), corneal reflection method, combination of corneal reflection and video image, and so on [5]. In our purpose, where it is used with head tracking and mainly used for detecting eye-head cooperative motion in ubiquitous environment, the eye tracking system should be lightweight and small enough. Our prototype system is mounted on 5 mm × 8 mm × 3 mm flexible printed circuit board and the weight is 3.4 g, which is sufficiently compact for attaching glasses, however the precision and the accuracy are meager than those of other elaborate eye tracking systems.

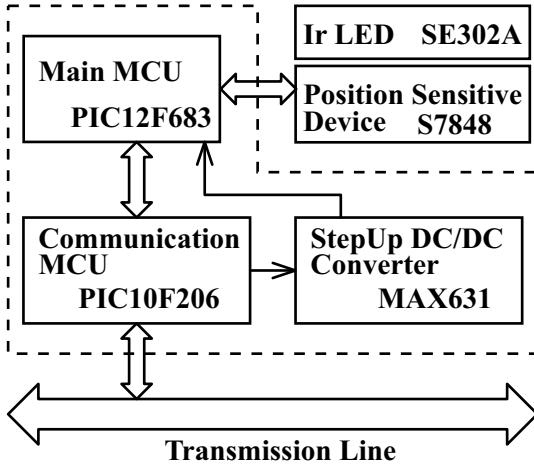


Figure 6: Hardware configuration of sensor node with 2-dimensional position sensitive device and infrared LED.

4. Detecting of eye-head cooperative gesture

Eye-head cooperative gesture is detected autonomously in a decentralized way by heterogeneous sensor nodes: acceleration sensor, gyro sensor, magnetic sensor, and position sensitive device.

According to the user's context, the network changes the mode into the *pulsed-network-mode*.

In each sensor node, for a time series $x_0(t)$, wavelet coefficients $y_i(t)$ and scaling coefficients $x_i(t)$ are calculated as follows.

$$\begin{aligned}
 y_1(t) &= \frac{x_0(t) - x_0(t-1)}{2} \\
 x_1(t) &= \frac{x_0(t) + x_0(t-1)}{2} \\
 y_2(t) &= \frac{x_1(t) - x_1(t-2)}{2} \\
 x_2(t) &= \frac{x_1(t) + x_1(t-2)}{2} \\
 &\vdots \\
 y_j(t) &= \frac{x_{j-1}(t) - x_{j-1}(t-2^{j-1})}{2} \\
 x_j(t) &= \frac{x_{j-1}(t) + x_{j-1}(t-2^{j-1})}{2}
 \end{aligned}$$

$$\begin{aligned}
 &\vdots \\
 y_J(t) &= \frac{x_{J-1}(t) - x_{J-1}(t-2^{J-1})}{2} \\
 x_J(t) &= \frac{x_{J-1}(t) + x_{J-1}(t-2^{J-1})}{2}
 \end{aligned} \tag{1}$$

This time-frequency decomposition is equivalent to the maximal overlap discrete Harr wavelet transform [16], but the expression is modified in order to be calculated only by additions, subtractions, and shift operators. Especially $x_J(t)$ represents the direct-current component, and the original time series $x_0(t)$ can be reconstructed as follows.

$$x_0(t) = x_J(t) + \sum_{j=1}^J y_j(t) \tag{2}$$

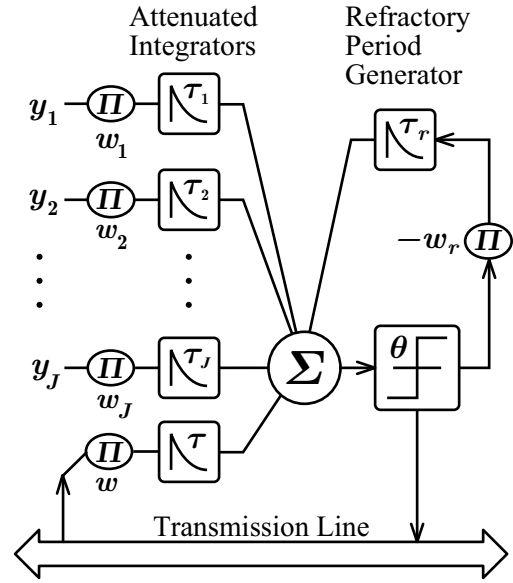


Figure 7: The block diagram of spiking neural network.

Each component of Haar transform $y_j(t)$ is multiplied by respective weight w_j and accumulated by attenuated integrator, and summed up. To inhibit an infinite activation, refractory period is generated by negative feedback.

In the simulated spiking neural network, each component $y_j(t)$ with weight w_j

```

int integrator(int value)
{
    int a = attenuation_factor;
    int temp = value;
    temp += y(t)*weight;
    value = 0;
    for(int i=16;i>0;i--){
        temp >>=1;
        a <<=1;
        if(carry is detected)
            value += temp;
    }
    return value;
}

```

Figure 8: A pseudo-code of attenuated integrator in Figure 7.

weight and attenuation_factor are predefined and $y(t)$ are input value from the sensor.

is accumulated by attenuated integrator, and summed up. Some of them are in excitatory connections and others are in inhibitory connections, according to the kind of reference signal pattern. Both of input and output are connected to the transmission line, therefore, refractory period is generated by negative feedback to inhibit an infinite activation.

Figure 7 depicts the block diagram of spiking neural network. Figure 8 shows a pseudo-code of attenuated integrator in the spiking neural network.

5. Calculation of gaze line

Using the data from the node with position sensitive device, the gaze line is also roughly calculated with the compensation of head orientation measured by the nodes with accelerometer, gyroscope, and magnetic sensor. It is achieved in the *normal-packet-mode* or *cascaded-packet-mode*. The direction vector of gaze line (ξ_x, ξ_y, ξ_z) is obtained in the same manner of our previous work [14].

Let (θ_x, θ_y) represents the direction vec-

tor of gaze line measured by position sensitive device, and (ϕ_y, ϕ_p, ϕ_r) denotes the orientation (yaw, pitch roll) of head measured by accelerometer, gyroscope, and magnetic sensor, then the vector of gaze line is given by (3).

$$\begin{aligned}
 \xi_x &= \cos \phi_y \cdot 0 - \sin \phi_y \cdot 0 \\
 \xi_y &= 0 \cdot 1 + 0 \cdot 0 \\
 \xi_z &= \sin \phi_y \cdot 0 + \cos \phi_y \cdot 0 \\
 &= 1 \cdot 0 + 0 \cdot 0 \\
 &\cdot 0 \cdot \cos \phi_p + \sin \phi_p \cdot 0 \\
 &= 0 \cdot -\sin \phi_p + \cos \phi_p \cdot 0 \\
 &= \cos \phi_r \cdot \sin \phi_r + 0 \cdot 0 \\
 &\cdot -\sin \phi_r \cdot \cos \phi_r + 0 \cdot 0 \\
 &= 0 \cdot 0 + 1 \cdot 1 \\
 &= -\sin \theta_x \cdot \cos \theta_y \\
 &\cdot \sin \theta_y \\
 &= \cos \theta_x \cdot \cos \theta_y \quad (3)
 \end{aligned}$$

The function of detecting eye-head cooperative gesture in *pulsed-network-mode* (Chapter 4) and calculating gaze line in *normal-* or *cascaded-packet-mode* (Chapter 5) are summarized in Figure 9.

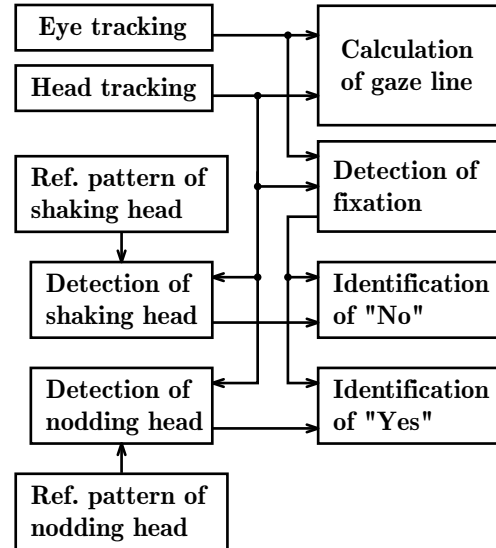


Figure 9: Logical configuration of the system for detection of eye-head cooperative gesture and calculation of gaze line.

6. Conclusions

This paper proposed a wearable sensor network system based on a capability of in-circuit serial programming. The main part of our system is eye-head movement tracking, and the roll of these movement depends on the modality and user's context. In order to cope with such variety of modality and context, we introduced four network modes: *normal-packet-mode*, *cascaded-packet-mode*, *pulsed-network-mode*, and *programming-mode*. We presented several applications of these modes for acquisition of user's intention.

At present we have only made a preparation for context-aware services with general-purpose equipment. Now we are starting on development of context-aware system with present network system combined with nonverbal interaction protocols. In addition, the consideration for individual difference is needed for further improvement, as a necessary part of our future work.

References

- [1] Benerecetti, M., Bouquet, P., and Bonifacio, M.: Distributed Context-Aware Systems, *Human-Computer Interaction*, **16**(3):213-228, 2001.
- [2] D'Costa, A., Sayeed, A.M.: Collaborative Signal Processing for Distributed Classification in Sensor Networks, *ISPN2003, LNCS 2634*:193-208, 2003.
- [3] Mote System, Crossbow Technology Inc., <http://www.xbow.com>.
- [4] Dey, A.K., Abowd, G.D., and Salber, D.: A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications, *Human-Computer Interaction*, **16**(2):97-166, 2001.
- [5] Duchowski, A.T.: Eye Tracking Methodology — Theory and practice, *Springer Verlag*, 2003.
- [6] Grocholsky, B., Makarenko, A., Kaupp, T., and Durrant-Whyte, H.F.: Scalable Control of Decentralised Sensor Platforms, *ISPN2003, LNCS 2634*:96-112, 2003.
- [7] Itao, T., Nakamura, T., Matsuo, M., and Aoyama, T.: Context-Aware Construction of Ubiquitous Services, *IEICE Transactions on Communications*, **E84-B**(12):3181-3188, 2001.
- [8] Jahnke, A., Roth, U., and Schönauer, T.: Digital Simulation of Spiking Neural Networks, "Pulsed Neural Networks", ed. Maas, W. and Bishop, C.M., *MIT Press*, 237-257, 1998.
- [9] Nonaka, H., and Kurihara, M.: Sensing pressure for authentication system using keystroke dynamics, *International Journal of Computational Intelligence*, **1**(1):19-22, 2004.
- [10] Nonaka, H., and Kurihara, M.: Anticipation of Mouse Pointer Movement Using Pressure Sensors, *International Conference on Computing Anticipatory Systems*, **7**, Liege, 2005
- [11] Nonaka, H., and Kurihara, M.: Time-Frequency Decomposition in Gesture Recognition System Using Accelerometer, *International Conference on Knowledge-Based Intelligent Information & Engineering Systems, LNAI, 3213*:1072-1078, Wellington, 2004.
- [12] Nonaka, H., and Kurihara, M.: Pulse-Based Learning for Object Identification using Ultrasound, *International Conference on Neural Information Processing*, **9**, Singapore, 2002.

- [13] Nonaka, H., and Kurihara, M.: Anticipatory Matching Method for Query-Based Head Gesture Identification, *International Journal of Computing Anticipatory Systems*, **15**:279-287, 2004.
- [14] Nonaka, H.: Communication Interface with Eye-gaze and Head Gesture using Successive DP Matching and Fuzzy Inference, *Journal of Intelligent Information Systems*, **21**(2):105-112, 2003.
- [15] Nonaka, H., and Kurihara, M.: Eye-Contact Based Communication Protocol in Human-Agent Interaction, *International Workshop on Intelligent Virtual Agents, LNAI*, **2792**:106-110, Irsee, 2003.
- [16] Percival, D.B., and Walden, A.T.: Wavelet Methods for Time Series Analysis, *Cambridge Univ. Press*, 2000.
- [17] Yamaguchi, A., Ohashi, M., and Murakami, H.: Autonomous Decentralized Control in Ubiquitous Computing, *IEICE Transactions on Communications*, **E88-B**(12):4421-4426, 2005.
- [18] Liu, J., Reich, L.J., Chung, P., and Zhao, F.: Distributed Group Management for Track Initiation and Maintenance in Target Localization Applications, *ISPN2003, LNCS* **2634**:113-128, 2003.

Pattern Concepts for Digital Games Research

Klaus P. Jantke

Technical University of Ilmenau

Institute for Media and Communication Science

Department of Multimedia Applications

Am Eichicht 1, 98693 Ilmenau, Germany

Abstract

The digital games industry has grown to the size of the film industry and is going to leave it far behind. Computers with advanced graphics capabilities have contributed to the immersive interactive experience that attracts many to spend more of their leisure time playing digital games than watching television. The available CPU power of current home computers and notebook PCs is setting the stage for game AI; console development shows a similar trend. And the expectations of human players are high.

Digital games and their potential social impact are subject to a heated debate worldwide which is fueled by tragic events such as the 1999 deadly shooting at the Columbine Highschool, Littleton, CO, USA, or the 2002 amok run at the Gutenberg Highschool in Erfurt, Germany. This debate is getting even more controversial when games such as the Super Columbine Massacre Role Playing Game enter the stage.

There are several good reasons—economically, socially, politically—to engage in digital games research. But digital games research has a high demand of interdisciplinarity. Digital games are at the same time entertainment media and IT systems. Even more specifically, they are turned more and more into complex AI systems. Digital games research is on its way to establish a digital games science. This discipline will have its language and its methodologies. And patterns are key concepts of understanding game playing, of understanding digital game reception and, thus, the potential social impact of particular digital games. Similarly, pattern concepts are crucial to anticipating the effects of game mechanics and, hence, play a key role in digital games design and development.

This paper aims at a contribution to the formation of interdisciplinary pattern concepts for digital games.

1. Patterns in Digital Games:

The Author's Initial Approach

However to generalize, we need experience, is a saying attributed to George Grätzer, one of the leading scholars in universal algebra. It applies to digital games science as well.

Where do we find in digital games instances of what might be a pattern?

When playing AGE OF EMPIRES II, e.g., most players experience certain difficulties and find particular ways to overcome them.

AGE OF EMPIRES II (figure 1) is a fantasy strategic development game with a rather substantial amount of fighting involved.

When being busy with developing their own civilization, players need to defend them-

selves against intruding NPC adversaries. A certain standard behavior—a game pattern—



Figure 1: AGE OF EMPIRES II in progress proves successful: *walling up* all entrances to their own settlement, esp. closing bridges.

2. The Underlying Perspective at Digital Games

Digital games are—no doubt—entertainment media and as such they are subject to media research. This is quite far beyond the limits of IT and AI.

When digital games are accepted as an art form, they deserve all the rights other arts enjoy such as free speech, to say it in brief.

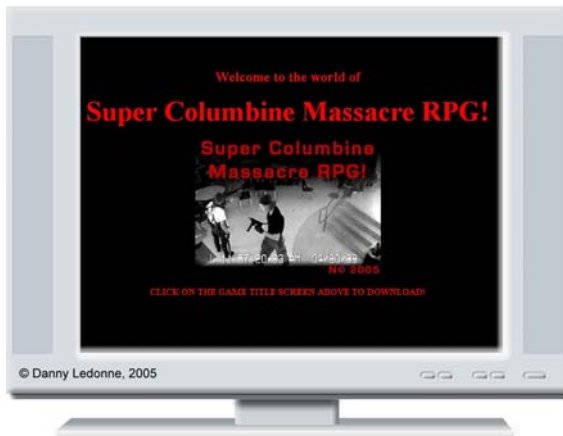


Figure 2: Download Page of a Free Game

Does this truly apply to all games including the SUPER COLUMBINE MASSACRE ROLE PLAYING GAME (see figure 2) ...? In this simple point and click shooting game, the player takes the role of one of the two teens who shot a dozen of their school mates and a teacher before they finally shot themselves.

Digital games reflect reality to some extent. In turn, playing digital games might have impact on some players' behavior in the surrounding real world (see [4, 17] for some comprehensive approaches and [11] for an interdisciplinary discussion).

Digital games have to be taken seriously. For the class of massively multi-player online role playing games (MMORPG, for short), Jeff Strain, co-founder of ArenaNet, in an interview¹, characterized those games as *just polished incarnations of a traditional MMO that was laid down ten years ago with Ultima Online. Its a game mechanic that requires you to grind away, in many cases, trying to level up. The whole goal in many of those games is to just get higher and higher*

¹www.just-rpg.com, April 28, 2006

leveled so that you can eventually, after a thousand hours, get to see the cool stuff. You dont take on those games lightly. If you're going to play a traditional MMO, not only do you have to be willing to commit to it financially in terms of an ongoing subscription, but also in terms of your life. You dont play casually—you're either completely immersed in it or you quit. Its kind of like people kicking a drug addiction.

Immersion or flow [7] is crucial to playing games [10]: *Games create a cybernetic system between you and the machine, with your senses eventually expanding to possess your avatar when you've sufficiently mastered the control system. This is the absolute magic of the form, where you stop thinking, "I need to press X to jump", and start thinking, "I'll jump". Just look at the language people use to talk about games to show how much their sense of identity has merged with their in-game character. If someones enjoying a game, it's, "It hit me", never, "It hit my character", ... Videogames are the simulator which swallows your consciousness alive and takes you to another place.*

In addition to the entertainment media perspective, digital games are clearly—usually complex—IT systems. Going even further, they should be seen as systems of artificial intelligence (AI). Computer scientists might be surprised, but in gamer communities, it still is a rather 'unusual perspective' to see digital games as IT systems ([16], p. 16).

To the author, it is fundamental to consider digital games at the same time as entertainment media and as systems of AI [11, 12].

When being interested in the way in which digital games might have impact on some human brains and, thus, have an impact of social relevance, one needs a holistic view at digital games: IT systems that affect the human brain by rewarding experiences [14].

Next to the analytical perspective, digital games science deals with principles and problems of digital game synthesis: design, development, and implementation of entertaining IT systems.

3. Digital Games as Generators of Game Playing Experience

After stressing the entertainment media perspective in the preceding section, the present one will put more emphasis on considering digital games as IT systems.

Digital games like all other types of games are generators of human playing experience. When humans play a game, sequences of activities are unfolded. What sequence may be potentially generated is determined by the game mechanics. But usually only a small amount of those potential sequences unfold in real game playing.

A more formal terminology helps to clarify the issues under discussion.

With a game G , there is given a finite set of potential activities that may take place. In formally simple cases such as playing CHESS, EINSTEIN WÜRFELT NICHT², or JOSTLE², every game play π appears as a finite sequence of moves of the alphabet of possible activities, i.e. $\pi \in A^*$.

In many games, especially in beat'em up games such as SOUL CALIBUR (see figure3)



Figure 3: SOUL CALIBUR II – Just a Few Keys Forming Large Numbers of Patterns

the necessity is arising to press certain keys simultaneously. This may be formally represented by either introduced extra moves for those combinations or, more elegantly,

²EINSTEIN WÜRFELT NICHT is a game developed by Ingo Althöfer, Jena, Germany. JOSTLE has been designed by the present author. Both games have been used for in-depth discussions of game patterns (see [12] for more details).

by inventing a particular partial operator of key combinations. The alphabet A of actions is supposed to be closed under this operator, i.e. those combinations of keys introduce extra actions into A .

Based on those technicalities, every game G may be seen as generating a formal language $\Pi(G) \subseteq A^*$ of potential game experiences.

So far, the approach laid out lies completely within IT terminology, what directly leads to a clarity outperforming all utterances of media sciences by far. But the limitations are obvious as well.

For complex games G and under realistic conditions of game playing, large parts of $\Pi(G)$ are never played. It remains open whether or not the part of $\Pi(G)$ really played within the *cybernetic system* of the players and the game [10] may be described in sufficiently clear terms. Let's call the set of really played action sequences within A^* $\Psi(G) \subseteq \Pi(G)$. What about $\Psi(G) \subset \Pi(G)$?

That this is not a scholastic discussion, shall be explained by means of a practical case.

Let us express it somehow formally first: In all cases of play observed by the author, playing the game AGE OF EMPIRES II involved playing frequently the subsequence of actions necessary to wall up some area, especial building walls to close bridges that lead to the player's settlement. (Note that those actions can usually be performed by three subsequent clicks, a structure that can be easily identified in any string $\pi \in \Pi(g)$.)

Let us express the problem secondly in more semantic terminology: Is it necessary that this pattern of players' behavior occurs in successfully playing AGE OF EMPIRES II? Or is it just easier to play that way? It is the author's suspicion that very fit players might be able to play the game without using the pattern mentioned. Being fast enough, they might be able to defeat invading hordes of NPCs instead of avoiding battles.

It remains open whether or not the discussed behavioral pattern occurs in all $\Pi(G)$, in all $\Psi(G)$, but not necessarily in $\Pi(G) \setminus \Psi(G)$, or just in some part of $\Psi(G)$.

4. Patterns in Game Play –

Key to Fun and Immersion

The author’s model of human game playing behavior was inspired by [9] and has been introduced in [11] in much detail.



Figure 4: Human Game Playing Behavior

Playing a game means to gain control over the balance of indetermination and self-determination. When you play a game, you learn to master certain difficulties. Several regularities of successful playing are learned unknowingly.

A game without regularities is unplayable. To say it the other way around, every playable game exhibits a lot of regularities.

Our human brains have developed over millions of years to become highly qualified mechanisms of identifying regularities [17]. Identifying regularities wherever occurring, whether consciously or unconsciously, is pleasant and gives us a good feeling—the key to understand fun in game playing [14].

A crucial point for understanding the fascination of games is to see their potentials of immersion. Players who are able to identify with their avatars may experience deepest satisfaction.

At the persona level of immersion, the virtual world is just another place you might visit, like Sydney or Rome. Your avatar is simply the clothing you wear when you go there. There is no more vehicle, no more separate character. It's just you, in the world. (Richard Bartle, cited after [18])

This assumes a highly unconscious mastery of patterns—key to digital games science.

5. Pattern Concept Case Studies

Koster [14] who nicely motivates patterns as key to fun does not mention a single particularly interesting case, whereas Björk and Holopainen [6] list more than two hundred of what they call patterns. They simply wrote down every regularity in games that came to their minds. In the author’s opinion, it does not make sense to call literally everything a pattern; *pattern* should not become another word for *everything*.

We need to find pattern concepts neither too narrow nor too wide that may serve as fundamentals of an interdisciplinary digital games science. Those concepts have to support a communication between scientists from disciplines as diverse as mathematics, computer science, artificial intelligence, psychology, sociology, and media and communication science.

According to Christopher Alexander [1, 2] who introduced the pattern concept to the sciences, to engineering, and to the arts, patterns are something that relates to human behavior.

Only when mathematicians adopted and adapted pattern concepts, they stripped them *from* the essentials.

The present section, with Grätzer’s initially cited words in mind, aims at a collection of patterns in digital games that are of different quality. The understanding of those case studies is deemed a basis for a subsequent systematization.

We did already discuss patterns such as the walling up pattern in AGE OF EMPIRES II which may be expressed as substrings in certain (or possibly all) game plays $\pi \in \Psi(G)$. Those are patterns on the level of actions³.

Conventional mathematical investigations mostly speak about patterns of this type [5].

Those patterns may have a more complex appearance due to the richness of the virtual game world, but may be reduced to their essential structure more or less directly. Such a

³“Moves” is the traditional term preferred in many publications; we are open to accept both expressions.

case shows in HALF-LIFE 2: EPISODE ONE as illustrated by figure 5.

In the survival game Half-Life 2 the human player is accompanied by his fellow NPC named Alyx. When the player solves some key tasks, Alyx is taking care of the rest of the work by eliminating all other attacking adversaries.



Figure 5: HALF-LIFE 2: EPISODE ONE – Cars for Blocking Exits of Ant Lions’ Nests

During mission 3, the player has to pass some underground parking area as depicted in figure 5. Certain adversary monsters named ant lions live in nests in the underground. The player has to avoid or to fight them. A quite useful tactics is to block the nest exits by pushing cars to block the holes. The same pattern of player activity solves the problem repeatedly five or six times.

Those are examples of patterns in sequences of actions—repeatedly occurring substrings.

In a larger number of point & click criminal story adventures such as BLACK MIRROR, AGATHA CHRISTIE: AND THEN THERE WERE NONE, and THE DA VINCI CODE solving riddles is essential to playing the game successfully.

Riddles (puzzles, ...) in games establish bottlenecks in the state space illustrated in figure 6. To reach certain sets of game states, one needs to pass a comparably small class of other states before.

Those bottlenecks may be characterized in different ways. Usually, the cardinality of bottlenecks is a secondary, but characteristic property. In many cases, it helps to look at

game states in an attribute-value-way adopting terminology from relational databases. There are bottlenecks characterized by some finite number of attributes a_1, \dots, a_n and particular values c_1, \dots, c_n . Game states s in a bottleneck are defined by the property $\forall_{i=1, \dots, n} s.a_i = c_i$ being valid in all states of the bottleneck, but in no predecessor state.

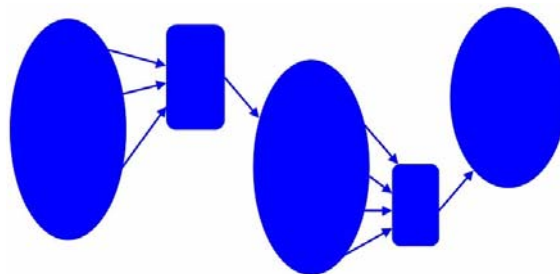


Figure 6: Bottlenecks in a Game State Space

Bottlenecks decompose the game state space into subspaces as depicted in figure 6.

The game THE DA VINCI CODE provides an extreme case study of a rigid game space structure with narrow bottlenecks. On the game phase from which the screenshot in figure 7 is taken, the human player is stuck until a certain number of actions has been performed to set certain attribute values.

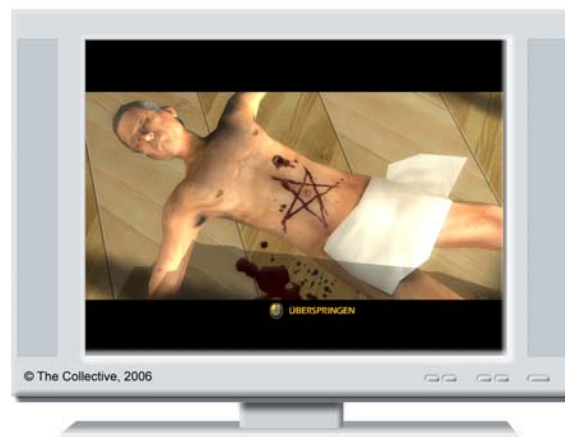


Figure 7: THE DA VINCI CODE – Bottlenecks in the Game State Space Determining Patterns of Human Game Playing Behavior

The whole game THE DA VINCI CODE consists of subspaces like that. The description of the pattern identified requires to speak about sets of game states and, perhaps, properties of states in a set.

Compared to the pattern found above in AGE OF EMPIRES, this is a higher level pattern.

Games such as BLACK MIRROR and AGATHA CHRISTIE: AND THEN THERE WERE NONE show the same pattern, but not as rigid as in THE DA VINCI CODE.

Let us conclude the present case study by some speculation. The strength of the player guidance through the game's state space makes those games good candidates for implementing something like a dramaturgy [15] and this, in turn, makes them subject to related analyzes [8, 13].

There is a large family of games normally called ego-shooters or first person shooters; those games are usually also quite linearly organized and, therefore, potentially good candidates of employing dramaturgy.

The player gets a usually simple mission and has to shoot a number of adversaries on his way. The game play is characterized by the first person perspective of seeing the whole environment 'through the eyes of the avatar'.

Literally all fantasy adventure games contain potions (magic drinks in fancy bottles or anything like that) for recreating an avatar's life energy or, a little bit stronger, to give him extra lives.

In first person shooters, those potions have a different appearances and look much less magically, though their effect is as magic as in any fantasy game.



Figure 8: CALL OF DUTY – Widely Used Patterns between Helpful and Annoying

There are simple patterns of presentation to attract the player's attention to the potions such a first-aid kits in the ego-shooter CALL OF DUTY that are glowing or even blinking.

In other games, avatars are turning their heads, e.g., to signal something important. In BLACK MIRROR certain keys are shining.

Let us direct our attention to properly more complex patterns. CALL OF DUTY is an award winning⁴ ego-shooter distinguished by several features that make playing more successful and, thus, more fun.



Figure 9: CALL OF DUTY – Patterns of Higher Level Guidance to Human Players

In the game CALL OF DUTY, the human player fights within a group of NPC fellows. The figures 9, 10 and 11 show screenshots taken within a few subsequent seconds of game play. From the player's perspective, the experience is as follows:



Figure 10: CALL OF DUTY – Screenshots from a Playing Experience, the Second Step

I am running through a building. My NPC fellows pass me. They leave the house through some doorway and I follow them (figure 9). They turn left (figure 9 and 10) and I do so as well. Outside (figure 10) they

⁴more than 80 awards since its publication in 2003

turn left and I am rather hot on their heels. In this way, we reach the next point where we are facing our adversaries (figure 11). 35



Figure 11: CALL OF DUTY – Screenshots from a Playing Experience, the Third Step
This guidance by NPC fellows exemplified is very valuable to the human player in some respect. Firstly, the NPC guidance helps finding the way in the virtual environment. Secondly, it keeps the pace high enough to cope with the time limits. Thirdly, the forced speed of movement keeps the tension high and contributes to the overall experience. Last but not least, when the player's NPC fellows run for cover, this indicates some danger and helps the human player also to act appropriately, e.g., to take cover as well. The pattern in the NPC behavior is difficult to describe in terms of states and moves, but easy to identify in game play.



Figure 12: METROID – The Morphing Ball
In the METROID games series, the human player acts as the intergalactic bounty hunter Samus Aran. This female NPC can change her appearance drastically (figure 12).

Shapped as the so-called morphing ball, she is better prepared for speed runs and able to pass quite narrow passages.

The syntactic form of the transformation as a behavioral pattern is extremely simple—just a single key pressed, i.e., a single letter in a sequence $\pi \in \Psi(G)$.

Obviously, identifying the change from the avatar Samus Aran into the morphing ball and, vice versa, unfolding Samus Aran from her spherical form does not say much about the experience of game playing.

Identifying a pattern syntactically rarely tells much. The crux is to relate syntax and semantics.

The METROID case study has been chosen to illustrate the key problem in especially simple terms.

In other cases, both syntax and semantics are difficult, but nevertheless important patterns occur. Consider, for instance, the survival game CALL OF CTHULHU which did appear in a PC version in Germany in early 2006. How is a truly creepy atmosphere provoked to arise?



Figure 13: The Creepy CALL OF CTHULHU

The question for the principles of touching human emotions is among the most difficult and deep problems of the arts. There are no universal recipes [15], but several dozens of examples we may learn from [8, 13].

One may call some of the tricks in use in the film industry patterns. There is some hope that the digital games science will also learn from the arts.

6. Hierarchies of Pattern Concepts

The case studies of the preceding section have brought to light some central issue of patterns in digital games science: Different patterns may require quite different terms to describe them.

First of all, do we expect and work for any universal approach that applies to any genre—whatever the term genre might mean—of digital games?

Experienced readers may have recognized that the author deviates from the usage of genre terms that appear currently in the majority of publications about digital games. The current practice results in a partially inconsistent usage of terminology. There is an urgent need for clarification, but not yet any firm basis. Therefore, the author prefers to avoid misleading, partially inconsistent terms and, instead, circumscribes what he wants—as seen above—in his own words.

The progress of a digital games science will also contribute to a revision of terminology. The author reserves his contribution to some forthcoming publication.

So, back to the issues of patterns in digital games without any considerations specific to the one or to the other genre.

To be honest, what will be proposed below has been adopted and adapted from media research, to some extent.

Consider systematic film analysis [8, 13] as a sample source. A standard approach is to separate a certain film under consideration in scenes and to group scenes into phases. On this basis, you may go into depth and attribute properties to scenes for possibly identifying regularities when looking at the film representation as a whole.

Notice that, interestingly, systematic film analysis typically does not talk about the film itself, but about the film's particular representation derived. In other words, we build models and perform reasoning about the models built.

That's what we are going to do in the digital games science as well.

For this purpose, the digital games science needs its modeling language(s)—the focus of this paper.

The insights one can gain depend very much on the language in use.

Let us have a brief look at one of the classics, Byron Haskin's film of the famous book⁵ "The War of the Worlds", 1953.

The film easily decomposes into 31 scenes which may be grouped into 5 main phases.

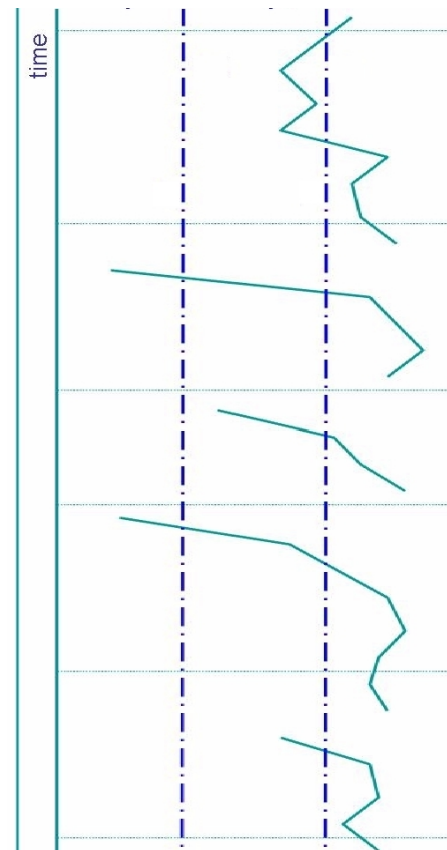


Figure 14: Formal Tension in the Film "The War of the Worlds", Byron Haskin, 1953

Based on such a structuring of the flow of activity, one might analyze whatever seems of interest. Following [8], we have simply counted the frequency of camera shots per minute. In figure 14, the scenes proceed from the top to the bottom. On the horizontal axis, the frequency of shots is shown. What the reader can see might be interpreted as some pattern(s) employed when making the film.

⁵The H. G. Wells book "The War of the Worlds" became properly famous by Orson Welles' radio feature broadcasted on October 30, 1938, by CBS in Northern America. It made Orson Wells famous.

When analyzing digital games, let us have a look at game play in a similar fashion. When game play proceeds, one scene follows the other. It depends very much on what you consider a scene.

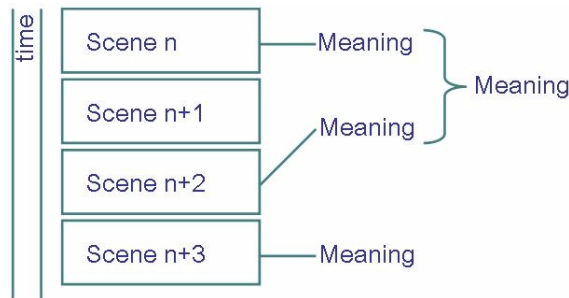


Figure 15: Some Schematic Approach to the (not only) Digital Game Play Understanding

On the atomistic level, one might consider every action in A a scene. In such a way, the sequence of scenes listed (in figure 15 from the top to the bottom) coincides with $\pi \in \Psi(G)$.

Every traditional protocol of a CHESS game, for instance, forms such a sequence. The above-mentioned sequences of actions in AGE OF EMPIRES II fit as well.

Even without any assigned particular semantics, one may go for string mining to exhibit potential patterns of game play.

Though this seems trivial because of being much too formal, it yields several results such as the preferred combinations of hits, kicks, and movements in a row a particular player of SOUL CALIBUR II performs.

The case study of AGE OF EMPIRES II was intended to point to the fact that several low-level patterns may have some higher-level meaning such as walling-up a bridge.

For understanding digital games, we need to decide about the language expressions admissible to describe meaning as indicated in figure 15.

What might seem difficult at a first glance becomes easier when having a closer look. Let us just ask what terminology we need to tell the story⁶ of a game play.

⁶This is not to be confused with storytelling as it is used in some communities: The author does *not* claim that every game tells a story. But we may tell a story about every game we are playing.

Once again, we take inspiration from the media sciences. For talking about drama and film, there does exist a rich terminology with concepts such as anagnorisis, climax, dénouement, and peripeteia, to mention a few. We need to develop words—or, better, concepts—to talk about playing a game.

The language of digital games science does not yet exist.

We need to develop a layered language, whereby layers may be visualized like the columns in figure 15; there is, naturally, no limitation to just three layers.

A few layers are immediately obvious:

- game states and moves,
- sets of game states,
- composite actions that establish some meaning which may be circumscribed on a higher language level.

Walling up a bridge is of the third type. We are still able to name the actions and have a look at every syntactic detail, but it seems clearly more convenient to communicate on a semantic level.

Other languages constructs may be introduced to talk about structural digital games features such as

- embedding of cut scenes,
- integration of puzzles,
- changing camera positions (switching between first person and third person views, e.g.).

The third point comes close to the level of formal tension (see figure 14).

Another higher level refers to information resp. knowledge of human players and NPCs;

- information systematically delivered in portions,
- indications given to make a player worrying (towards 'Eigenaffekt' à la [15]).

On the highest level according to the present approach, one might talk about the player's mental states, her/his goals and intentions and the like. Such a language level would be

necessary to address complex patterns such as those discussed with respect to the game CALL OF CTHULHU above (see figure 13). How to cause affright?

One might go even further and speak about more than the game itself. What about the embedding of game play in the surrounding world?

Important aspects are not yet covered by the approach sketched above. This fact shall be illustrated by means of a single case study.

TREASURE⁷ is a digital game developed for academic purposes. The focus of the underlying research is on communication using varying protocols such as GPS or Bluetooth.

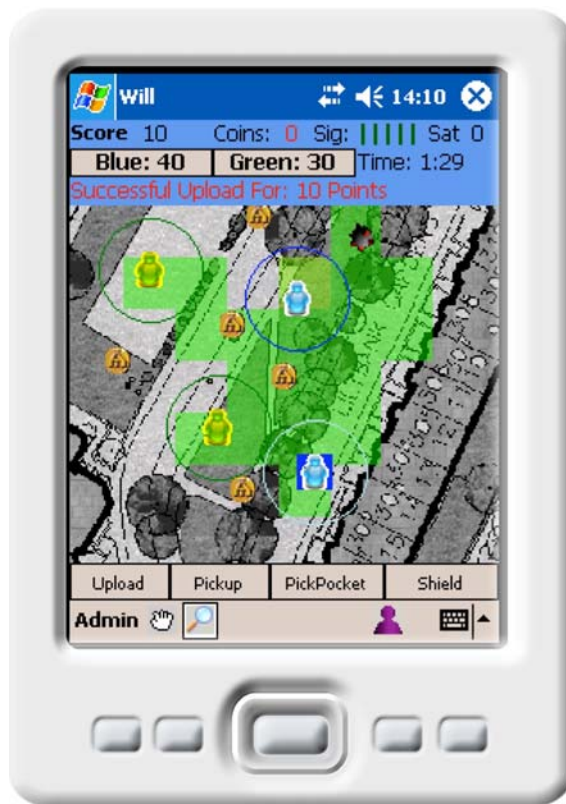


Figure 16: TREASURE – Patterns of a Higher Level – Human-System Interaction

The aim of the TREASURE game is to collect ‘coins’ scattered over an urban area such as a park on the PDA’s display in figure 16. The players’ goal is to get them in to the treasure chest (technically, a server). Two teams of two players compete against each other. A clock counts down, and the team with the most coins in their treasure chest at the end

⁷The PDA screenshot on display in figure 16 is published with permission of the authors.

wins the game. The coins can only be seen on the map on the players PDAs.

When players come close to a treasure, they can pick it up (see the related button on the PDA screen). For saving coins to the treasure chest, they have to come into the reach of the server’s wireless network.

There is a variety of more player interactions such as picking pockets (by using the button ‘PickPocket’ on the PDA screen), when one player comes close to another one. Playing the game needs obviously certain awareness of network connectivity issues.

Among many other issues, the developers of TREASURE have been interested in the players’ behavior with respect to seams. Such a seam is a break (or gap) between tools or media. Seams appear frequently when different digital systems are used in combination, or when they are used along with the other conventional media that make up our everyday environment.

The TREASURE game has been intentionally designed as a seamful one. Mastery of the game requires to perceive and master seams such as a missing connection to the server.

Patterns that show in the TREASURE game are of a quality quite different from those patterns discussed so far. To express those patterns, one may need concepts from the environment such as locations and relations from urban environments, communication protocols, and, perhaps, combinations of them.

For illustration, a certain pattern might contain a description saying that a human player with his PDA is in reach of some wireless network able to connect his PDA to the background server. The weather—really the true weather out there—might play a role in patterns.

The technology is bringing game playing back to our natural environment where it belongs⁸. The digital games science is not yet prepared to deal with those phenomena.

⁸...such as *TV brought murder back to home where it belongs*, a well-known saying attributed to Alfred Hitchcock

7. A Brief Outlook

A large variety of problems have not even been touched in the present short paper; here is just one quite complex case:

In the German manual of **BLACK & WHITE**, a strategic development game, you find a rather irritating statement on page 19: *Denke immer daran, dass dir deine Kreatur im Laufe der Zeit immer ähnlicher wird. Sie verkörpert deine Persönlichkeit und deine Spielweise.*

You play with "god's hand" and develop a creature such as the cow on display (see figure 17). You teach and train your creature such that it develops a particular behavior.

The cited German text from the manual says that players should be aware of the fact that their digital creatures' behavior is not only reflecting the way they play, but the players' personalities. This sounds disconcerting.

Indeed, some communication with players revealed concern about the extent to which the creature's behavior is telling about the player's individuality.



Figure 17: **BLACK & WHITE** – Teach Your Creature and Project Your Own Personality

Patterns in game playing are reflected by patterns in the NPC's behavior, a relationship which has not yet been subject to any systematic investigation. The issue might be of interest both to the social sciences and to IT and AI—there is exciting work to come.

The topic addressed needs expressions for qualities of an NPC's behavior and terms for potentially related features of a player's personality; we arrive at patterns in psychology.

8. Acknowledgments

Several of my students delivered substantial contributions to our recent research towards a *Digital Games Science*.

There is not sufficient space to go into detail, but one paragraph should be used to give a brief impression of each ongoing work.

Jochen F. Böhm, currently in Saarbrücken, gave me comprehensive introductions into digital games I did not know much about.

Arne Primke, currently in Berlin, recently started investigations into patterns which possibly show in the two ego-shooter series **CALL OF DUTY** and **MEDAL OF HONOR**.

Carsten Lernerz and *Sebastian Dittmann* have undertaken a systematic analysis of the **METROID** game series under the perspective of occurrence and development of patterns.

Sebastian Gaiser and *Andreas Schmidt* are investigating the development of classical jump'n run games such as **SUPER MARIO**.

Monique Scholz and *Steffen Rassler* became fascinated by the truly unique atmosphere players are experiencing when playing the survival/horror game **CALL OF CTHULHU: DARK CORNERS OF THE EARTH**. They started their own search for the patterns that stimulate such an impressive atmosphere.

Steffen Kehlert systematized the utilization of commercial modern music in particular digital games such as sports and fun sports games like, for instance, **TONY HAWKS** and **NEED FOR SPEED**.

Many more of my students in Darmstadt, in Ilmenau, and in Mannheim contributed through engaged discussions, useful hints, interesting suggestions, and by pointing to necessary corrections.

Last but not least, the author's work towards a digital games science is very much driven by considerations of memetics (see [11, 12] for a much more comprehensive discussion). It seems particularly exciting to find out how ideas of game and play evolve over time. The present author's colleagues and friends *Susan Blackmore* and *Yuzuru Tanaka* are invaluable partners in related discussions.

Appendix A: (Digital) Games that are Mentioned in this Publication

AGATHA CHRISTIE: AND THEN THERE WERE NONE
AGE OF EMPIRES II
BLACK MIRROR
BLACK & WHITE
CALL OF CTHULHU: DARK CORNER OF THE EARTH
CALL OF DUTY
CHESS
EINSTEIN WÜRFELT NICHT
HALF-LIFE 2: EPISODE ONE
JOSTLE
MEDAL OF HONOR
METROID
NEED FOR SPEED
SOUL CALIBUR II
SUPER COLUMBINE MASSACRE RPG
SUPER MARIO
THE DA VINCI CODE
TONY HAWKS
TREASURE
ULTIMA ONLINE

References

- [1] C. Alexander. *The Timeless Way of Building*. New York: Oxford University Press, 1979.
- [2] C. Alexander, S. Ishikawa, and M. Silverstein. *A Pattern Language*. New York: Oxford University Press, 1977.
- [3] R. Bartle. I was young and I needed the money. *The Escapist*, (43):4–9, 2006.
- [4] J. Berndt. *Bildschirmspiele: Faszination und Wirkung auf die heutige Jugend*. Münster: Monsenstein und Vannerdat, 2005.
- [5] J. Bewersdorff. *Luck, Logic & White Lies. The Mathematics of Games*. Wellesley, MA, USA: A K Peters, 2005.

- [6] S. Björk and J. Holopainen. *Patterns in Game Design*. Hingham, MA, USA: Charles River Media, 2004.
- [7] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, 1991.
- [8] W. Faulstich. *Grundkurs Filmanalyse*. München: Wilhelm Fink Verlag, 2002.
- [9] J. Fritz. *Das Spiel verstehen. Eine Einführung in Theorie und Bedeutung*. Weinheim & München: Juventa, 2004.
- [10] K. Gillen. Culture wargames. *The Escapist*, (1):8–13, 2005.
- [11] K. P. Jantke. Digital game knowledge media (Invited Keynote). In Y. Tanaka, editor, *Proceedings of the 3rd International Symposium on Ubiquitous Knowledge Network Environment, February 27 March 1, 2006, Sapporo Convention Center, Sapporo, Japan, Volume of Keynote Speaker Presentations*, pages 53–83. Hokkaido University of Sapporo, Japan, 2006.
- [12] K. P. Jantke. Knowledge evolution in game design – just for fun. In *CSIT 2006, Amman, Jordan, April 5-7, 2006*, 2006.
- [13] H. Korte. *Einführung in die Systematische Filmanalyse*. Berlin: Erich Schmidt Verlag, 2004.
- [14] R. Koster. *A Theory of Fun for Game Design*. Scottsdale, AZ, USA: Paraglyph Press, Inc., 2005.
- [15] P. Rabenalt. *Filmdramaturgie*. VIS-TAS media production, 2004.
- [16] T. Schreiner. Faszination Videospiele. *PLAYZONE*, (08/06, Ausg. 94):16–21, 2006.
- [17] M. Spitzer. *Lernen. Gehirnforschung und die Schule des Lebens*. Spektrum Akademischer Verlag, 2002.
- [18] M. Wallace. In celebration of the inner Rouge. *the escapist*, (30):4–6, 2006.

Security Aspects in Online Games Targets, Threats and Mechanisms

Anja Beyer

Technische Universität Ilmenau

Institut für Medien und Kommunikationswissenschaft

Am Eichicht 1, 98693 Ilmenau

anja.beyer@tu-ilmenau.de

Abstract

This paper presents a summary of possible threats in Online Games based on the fundamentals of IT-Security. The IT security is defined via its targets confidentiality, integrity, availability, privacy and non-repudiation. Possible threats are identified and protection mechanisms are discussed. An outlook on the oncoming project to security in Online Games will conclude the paper.

1. Motivation

Since "World of Warcraft (WoW)" was published in the USA in 2004 Online Games are truly booming. WoW started with about 200.000 players in the United States. In the meanwhile there are about six million people playing WoW on 40 servers around the world in one environment. For games like this, where thousands of players can play simultaneously in the same environment via the Internet, a new term was created: "Massive(ly) Multiplayer Online Role-Playing Games (MMORPG)". Before the Online Games became popular the games were designed for one player against a virtual player simulated by the game software. The Online Games allow the people playing together (with real human beings) from all over the world [16]. That this development is an important one (not only) for the future of games is invigorated through the statement of Kirmse and Kirmse. They say Online Games are the biggest revolution in games since the introduction of home computers [9]. WoW is the most popular and successful MMORPG but not the only one as there are lots of others (e.g. Dungeons and Dragons online, Guild Wars, Lineage 2).

A relatively new effect in this genre is that the virtual worlds and the reality affect each

other increasingly. That means, that playing an Online Game is not only fun but in some cases also real business [10]. For the players "the virtual worlds are as real as the physical world" [6]. They do not only have fun playing the games, they spend money and leisure time in the virtual world and experience flow [5]. The players get carried away in the games and they attach a certain importance on them. At this point the games become a serious touch and thoughts of security issues are necessary.

In the academic field a lot of work is done on graphics and hardware improvement for games but less work is done on security issues. Smed et al. [15] concentrate on the networking issues, like latency, bandwidth and computational power. Kirmse and Kirmse [9] recognize two security goals for online games: protecting sensitive data (like credit card numbers) and providing a fair playing field. These two goals are of high concern but do not cover all security needs. Other publications are only focused to cheating [17, 3, 1]. Of course honest players do not like other players cheating and so they exert pressure on the publishers to provide a fair game setting. The games industry has realized that they will lose players and money when they do not act against cheating [15]. In this paper the author wants to show that

IT-Security is an important topic and more than just protection against cheating. However this paper does not provide a full list of all possible threats, so the author focuses the most valuable assets of the player and the provider.

2. IT-Security Targets

From the daily IT news one could get an impression on how important the topic IT-Security already is. The number of "bad news" is increasing steadily. With more and more applications connected to the insecure network 'Internet', the problems will not become less and the topic IT-Security will be of even more importance in the future. To give a short introduction the author wants to give an overview on what IT-Security is.

2.1. In General ...

In [11], IT-Security is mostly defined via its protection targets confidentiality, integrity and availability. In e-commerce systems the goals non-repudiation and privacy are additionally important.

Confidentiality means the protection against unauthorized access to data and information. Communication between two partners is thought to take place secretly. That means that no third party is allowed to acquire knowledge about the communication.

Integrity refers to protection against unauthorized modification of data or information: the shown information has to be correct and presented unmodified.

Availability indicates the protection against unauthorized interference of functionality.

Non-repudiation expresses the unauthorized non-commitment, meaning the loss of bindingness. Business partners have to be sure that both stay with their proposal to sell or buy.

Privacy, also regulated by the EU Directive 95/46/EC on the protection of personal data [4], is the right of an individual person on informational self determination. It allows

individual persons to decide about the usage of their personal data.

2.2. ... and in Online Games

From [11] we have learned that it is important not to consider IT-Security from only one perspective (multilateral security). That is why this chapter shows which security requirements apply to Online Games from the players and from the publishers view.

As described above, confidentiality is necessary in interpersonal communication and in e-commerce transactions. This is especially true in Games and specifically essential in Online Games. The communication of two players or a team is done via the Internet which does not provide any protection against eavesdropping or trapping information. When the players in an Online Game arrange a strategy to fight against other players or to win a match, it is unaccepted that this information reaches the combatant because in that case the combatant would have an unfair advantage. Also payment data, like credit card numbers have to be confidential.

From the players perspective there are two main concerns that affect integrity: on the one hand the player is interested in the reached score or level, stored on the server of the publisher, not to be unauthorized modified. The most valuable asset of the player is the reached score or level of the game. On the other hand, the player's system has to be kept integer, so that no unauthorized changes are done in the system without the recognition of the user (as lately happened with the Sony root kit [14]). As the player pays a monthly fee (€ 10,99 for WoW) for playing Online Games on the publishers server, he or she will not accept an hour or daily long unavailability of the server. So the the publisher is forced to install mechanisms that assure the availability of the servers (see chapter 3).

Once agreed to a contract and having paid for the usage, the player is unwilling to accept a repudiation of the contract from the publisher. But this is exactly what regularly happens when players are proven cheating.

Also privacy is a goal that must not be underestimated. According to the law [4], each person has the right to decide who may save and to what extent he or she is allowed to save the personal data.

The publishers most valuable asset is of course the game itself because this earns the money. So the publisher is interested in the integrity of the game. This means that no unauthorized person should be able to change the setting of the game. In combination with that, the availability of the server is an important target for the publisher. So people will only pay for the game if they are able to play. Here the targets of the player and the publisher are the same.

Furthermore, the publisher wants to make the players satisfied to make them keep on paying. They will do if their interests are fulfilled. They will not be satisfied if other players will have an unfair advantage. So the publisher has to act against cheating. In the example of WoW the publisher scans the user's system for cheating tools which is a heavy intervention in the users system.

Also, the usability of the system is an important target because high support costs have immense economic consequences.

It is obvious that the security targets of the publishers are not always the same like the security targets of the players. They compete in some cases (system scan, data collection).

3. Security Threats in Online Games

In the following chapters, possible threats on client side and on servers/publishers side are described. The classification is according to where the attack takes place. This means that the damage can be at another place. If e.g. the publisher does not protect its servers in an adequate way, there might be the possibility that an attacker can read out the credit card payment data of the customers. The attack is on the servers side but the user is affected of the damage if the attacker uses his or her credit card number for not intended purposes.

3.1. On Clients Side

An attack on confidentiality on clients side is the eavesdropping of the communication between players. If a combatant is able to listen to their communication and finds out about their strategy, he or she would have an unfair advantage. Considering confidentiality, a player also has to be aware of phishing attacks. The term phishing derives from the two words password and fishing. It is a criminal activity belonging to social engineering [12]. An attacker pretends to be a trustworthy person or company and tries to fraudulently acquire account names and passwords by asking the players directly using mail. Furthermore on confidentiality, there is the risk of malware injection. Malware stands for malicious software and means viruses, trojan horses, worms, etc. that are programmed to spy out account data and passwords.

Considering integrity, there is the risk of unauthorized modification of the clients hardware or software configuration.

A Denial-of-Service (DOS) attack can make the client game software unserviceable so that it cannot be used anymore.

Privacy attacks are also possible, when an attacker is able to spy out personal data on the clients side.

If a player does not stay with the promise to buy or pay, we speak of the threat of non-repudiation. This is e.g. the case if players are displeased with dishonest players cheating. This once happened with the MMORPG Call of Duty 2, when players called out a boycott [2].

3.2. On Servers Side

A possible threat concerning confidentiality on server side is to spy out secret information about the combatant, e.g. the amount of gold he or she owns.

There are several attacks on integrity on the servers side. First of all, there is the threat of unauthorized modification of the games functionality. Since the game is the most valuable asset for the publisher, it might be

the worst case scenario if an attacker succeeds in changing the games setting, e.g. to deactivate a whole continent of an environment.

The games scores of the players are saved on the server. So the publisher has to protect the scores from unauthorized modification of the games scores of the players. A third attack imaginable is URL-Spoofing. This means, if an attacker succeeds in changing the URL's of the gaming servers, the players are forwarded to a third-party server with the risk of revealing sensitive information to an untrustworthy party.

There is also the risk of Non-Availability of the server, e.g. due to unscheduled maintenance or DOS attacks.

Publishing personal data of players either willfully or through inappropriate protection of the storage is an attack on Privacy.

There is the danger of an attack on Non-Repudiation on servers side, too. If the publisher decides to eliminate players from the game, e.g. because he or she suspects or proves them cheating, he steps back from the contract although the player has payed for the usage. That is why the publishers insert an clause in general terms and conditions to leave that option open.

Some examples show that it must not always be the bad attackers from the outside who threat the games. The threats can also come from the inside (e.g. unsatisfied or careless employees) or due to act of nature beyond control.

3.3. Cheating

As the above threats concern digital games they are not really special for Online Games. All other IT application systems are exposed to these threats, too. As digital games are special IT application systems [7] these threats also apply to them. Not a new but a very special threat to digital games is cheating. Yan [8] defines "any behavior that a player may use to get an unfair advantage, or achieve a target that he is not supposed to is cheating." Yan and Randell [16] classify

cheats into 15 categories:

- Cheating due to misplaced trust (A)
- Cheating due to Collusion (B)
- Cheating by Abusing Game Procedure (C)
- Cheating related to Virtual Assets (D)
- Cheating due to Machine Intelligence (E)
- Cheating via the Graphics driver (F)
- Cheating by Denying Service to Peer Players (G)
- Timing Cheating (H)
- Cheating by Compromising Passwords (I)
- Cheating due to Lack of Secrecy (J)
- Cheating due to Lack of Authentication (K)
- Cheating by Exploiting a Bug or Loop-hole (L)
- Cheating by Compromising Game Servers (M)
- Cheating Related to Internal Misuse (N)
- Cheating by Social Engineering (O)

The above list covers two aspects why cheating is possible: first because of the technical possibilities (I, J, K, L, M, N): The cheater, if he or she has enough malicious intend might exploit the games procedure and bugs to gain an unfair advantage. Second is the social aspect (D, G, O): some players are not aware of the threats and give trust to the wrong people (O).

4. Security Mechanisms

For a wide variety of threats there already exist protection mechanisms. A secret communication can be realized via encryption

and virtual private networks (VPN). Protection against malicious software can be done through firewalls, virus scanners, intrusion detection systems and an adequate configuration of hard- and software. The condition is that they are kept updated. To avoid a breakdown of the servers due to maintenance or due to unforeseen incident, back-up systems are recommended.

To prevent players from cheating the one and only solution at the moment is to scan the users system to detect additionally installed cheating software. The game producers integrate software to scan the users system for cheating software. One famous tool is called PunkBuster [13] which is used in e.g. World of Warcraft and Call of Duty 2. If once a player is proven cheating, he or she will be eliminated from the game. From the authors view, until now, there is not a good solution to fight against cheating. Especially the social aspects of cheating, like social engineering or denying service, can not be prevented.

It follows that not all of the threats can be faced with pure technical solutions. For example phishing is a real awareness problem. It is important to call the attention of the people to reduce such attacks.

Maybe not all forms of cheating can be circumvented easily, but at least a proper implementation and configuration is the minimum. To meet the social aspects of cheating takes much more effort because the awareness of the people has to grow in their heads.

5. Conclusions and Further Work

Online Games are IT application systems which have to meet security standards as usual networked IT applications do. Furthermore, the games have an additional emotional factor. As the players identify themselves with the avatars and spend much time in the games, they will have to be satisfied by the games industry who earn money with the games.

In our future work, we want to analyse the security mechanisms in current Online Games. We expect that not enough security

mechanisms are realized to meet the security requirements of these games. We will provide a proposal of which and how further mechanisms can be included.

Our aim is to raise the awareness for existing threats of the producers, publishers and players. In our "Security in Online Games" project, starting from August 2006, we want to survey how security can be implemented without the players losing the fun of gaming (e.g. by clicking away security alerts).

References

- [1] H. Banavar. Security issues in multi-player, distributed network games. <http://ww2.cs.fsu.edu/banavar/research/NSPaper.htm>, 2000.
- [2] Call of duty 2 - community ruft zum boykott auf. Winfuture - Windows Online Magazin, <http://www.winfuture.de/news/23212.html> [28/11/2005], June 2006.
- [3] Y.-C. Chen, J.-J. Hwang, R. Song, G. Yee, and L. Korba. Online gaming cheating and security issue. *International Conference on Information Technology: Coding and Computing (ITCC'05)*, 1:518–523, 2005.
- [4] Directive 95/46/EC of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995.
- [5] Moore R.J. Ducheneaut, N. The social side of gaming: A study of interaction patterns in a massively multiplayer online game. *Proceedings of ACM CSCW'04 Conference on Computer-Supported Cooperative Work*, 6(3):360–369, 2004.
- [6] M. Jakobsson. *Virtual worlds and social interaction design*. PhD thesis, Umeå University, Department of Informatics, 2006.
- [7] K. P. Jantke. Digitale Spiele Forschung, Technologie, Wirkung & Markt, Invited Plenary Talk, Wismar, Germany, June 8/9, 2006.

- [8] Yan J.J and Choi H.-J. Security issues in online games. *The Electronic Library*, 20(2):125–133, 2002.
- [9] C. Kirmse and A. Kirmse. Security in online games. Gamasutra <http://www.gamasutra.com>, originally published in *Game Developer*, July 1997.
- [10] Dr. A. Lober and O. Weber. Money for nothing? - Handel mit Spiel-Accounts und virtuellen Gegenständen. *c't*, 20:178–180, 2005.
- [11] G. Müller and A. Pfitzmann. Sicherheit, insbesondere mehrseitige IT-Sicherheit. In *Mehrseitige Sicherheit in der Kommunikationstechnik Verfahren, Komponenten, Integration*, pages 21–29. Addison-Wesley-Longman, 1997.
- [12] Phishing. Wikipedia, <http://en.wikipedia.org/wiki/Phishing> [29/06/2006], June 2006.
- [13] PunkBuster Anti-Cheating Software. <http://www.evenbalance.com>, July 2006.
- [14] M. Russinovich. Sony, rootkits and digital rights management gone too far. http://www.sysinternals.com/blog/2005_10_01_archive.html, published 31.10.2005, June 2006.
- [15] Hakonen Harri Smed J., Kaukoranta T. Aspects of networking in multiplayer computer games. In *Proceedings of International Conference on Application and Development of Computer Games in the 21st Century*, pages 74–81, 2001. available at <http://www.tucs.fi/Publications/proceedings/pSmKaHaa.php>.
- [16] J. Yan and B. Randell. Security in computer games: from pong to online poker. Technical Report Series, Published by the University of Newcastle upon Tyne, 2005.
- [17] George Yee, Larry Korba, Ronggong Song, and Ying-Chieh Chen. Towards designing secure online games. In *20th International Conference on Advanced Information Networking and Applications (AINA 2006), 18-20 April 2006, Vienna, Austria*, pages 44–48, 2006.

Personal File Sharing in a Personal Bluetooth Environment

Jürgen Nützel and Mario Kubek

Technische Universität Ilmenau

Institut für Medien- und Kommunikationswissenschaft

Am Eichicht 1

98684 Ilmenau

Abstract

This paper introduces a piconet file sharing application using Bluetooth. The described Java-based application is designed for mobile devices supporting J2ME. The focus of the paper is on the description of the needed client/server architecture, the used message protocols and the final implementation. Therefore common Bluetooth protocols and Bluetooth APIs for J2ME are explained. The paper concludes with a view on future enhancements regarding automatic program activation, the handling of DRM and the integration of user profiles.

1. Introduction

Modern mobile devices like cell phones or PocketPCs offer many different connectivity solutions. Beside WAN protocols based on W-CDMA and GSM, short-distance wireless communication technologies like IrDA, Bluetooth and WLAN are very common.

The usage of these technologies is free of charge (WLAN can be an exception) and does not require network operators or providers for data transfer, which makes them attractive for users that are currently in the same area and need to exchange data. While IrDA based connections have a lack of flexibility as they can only be established between two devices that point to one another during the data transfer, WLAN support is only available in premium class devices. The Bluetooth technology in mobile devices, which is the main focus of this paper, offers more flexibility than IrDA and is available in nearly every business class device. Data transmission is realized via radio communication.

The Bluetooth Special Interest Group (SIG) [1] with members like Ericsson, IBM, Intel and Nokia releases the Bluetooth specification. The Bluetooth standard 1.x supports data rates of 723,2 kBit/s. In 2004 version 2.0 was released. This standard enables data rates of 2.1 MBit/s. A Bluetooth network (Piconet) can contain up to 16,7 Mio participants, whereby only eight devices can be active at the same time. Application specific profiles, that represent the capabilities of a mobile device, are the structure for data transfer via Bluetooth. Common profiles among others are the Dial-up Networking Profile, the File Transfer Profile, the Headset Profile and the Fax Profile. Their signature is exchanged between the devices, so that their capabilities can be matched in order to use the different services. Bluetooth even supports security mechanisms as authentication and confidentiality.

Additionally many mobile operating systems like Windows Mobile or Symbian offer Bluetooth APIs for

development. Applications built for these devices will not run on other devices. Therefore the Java virtual machine has been developed. This way J2ME (the mobile edition of Java) applications, called MIDlets, can run on many devices with different operating systems. The J2ME is built on configurations and profiles. They are the basic classes and methods that J2ME programs can use on the devices. The most common profile is the MIDP, based on the Connected Limited Device Configuration (CLDC). These classes can be extended by the device vendors with optional APIs to support device specific features. Examples are the Mobile Media API (JSR-135), the Location API (JSR-179) and the Bluetooth API (JSR-82). JSR stands for Java specification request.

This paper describes a mobile file sharing application, called BlueMatch, which can run on devices supporting MIDP 2.0 and the Bluetooth API (JSR-82). The special characteristic of BlueMatch lies in its combined client/server architecture, which follows the peer-to-peer model, so users can search and download files, while others can download from their devices at the same time.

First we want to present the components of the Bluetooth API before we go into details of BlueMatch's system architecture and message protocol. Future enhancements and a view on the handling of DRM-encrypted files in such an application conclude the paper.

2. The Bluetooth API

The Bluetooth stack consists of several protocol layers, that are implemented in hardware and software. The low level layers radio, baseband/link controller and link manager are not covered in this paper. We will focus on higher level

protocols like RFCOMM (radio frequency communication) and SDP (service discovery protocol), as they are used in the file-sharing application BlueMatch. Therefore it is useful to introduce the Bluetooth API (JSR-82) first, because it enables the usage of these protocols in mobile Java applications.

The Bluetooth API (JSR-82) [2] consists of two packages `javax.bluetooth` and `javax.obex`, the optional Object Exchange API. The classes in `javax.bluetooth` offer methods for discovery of devices and services, the support for connections using L2CAP (logical link and adaptation protocol) and RFCOMM, as well as device and data interfaces. These packages rely on the generic connection framework (GCF) in package `javax.microedition.io`. We do not want to cover the methods offered in the API, instead we discuss the general abilities of these classes.

The following Bluetooth protocols can be used in JSR-82:

- SDP for device/service discovery and service registration
- L2CAP for packet-oriented connections
- RFCOMM for stream-oriented connections
- OBEX for transferring objects like files, images and vCards

The protocol RFCOMM, which offers two-way communication over virtual serial ports, resides on top of L2CAP and enables multiple concurrent stream-oriented connections between devices, even over one Bluetooth link. L2CAP segments the data by RFCOMM into packets before they are sent and reassembles received packets into larger messages. L2CAP is a packet-oriented protocol and is suitable for developers that wish to build custom packet-

oriented protocols. In this case flow control has to be implemented explicitly. RFCOMM supports this already and enables security parameters. Therefore the BlueMatch application relies on this protocol. OBEX is implemented on top of RFCOMM, but actually only a few devices on the market support this optional protocol in Java applications. So file transfer has to be implemented using RFCOMM or L2CAP. The service discovery protocol (SDP) is used by clients to search for Bluetooth devices and services. It is also responsible for registering custom service records in the device's service discovery database (SDDB). After registration a server can accept and open connections by clients.

Security Mechanisms

RFCOMM enables secure Bluetooth connections. These security parameters for service usage can be set:

- authenticate
- encrypt
- authorize

Authentication means, that two devices must be paired in order to exchange data. Encryption can be activated to secure a link between two devices and relies on authentication. The keys used for authentication and encryption are generated by modified versions of the 128-bit block cipher algorithm SAFER+ [3]. Authorization is used to demand a permission for remote devices to use a service. The requested device grants or denies access to a service, for instance after asking the user through a man-machine interface to grant trust. This trust can be temporary or permanent.

3. BlueMatch

Now we want to present the prototype application BlueMatch, which enables

users in passing to share files via Bluetooth. As a MIDlet BlueMatch can run on mobile devices that support MIDP 2.0 applications and implement the Bluetooth API (JSR-82) and the file connection API (JSR-75). It is based on a client/server architecture, whereby the client part runs concurrently with the server part in separate threads. That means an instance can connect to the server on other devices, but can even process connections initialized by remote devices. On incoming connections, the server accepts and opens them automatically without asking the user.

Currently these features are already implemented:

- search and display remote devices running BlueMatch
- download of file lists from remote devices
- download of files from remote devices
- serving of local file lists and files for remote devices
- leaving notification for other members when quitting BlueMatch

These functions are the use cases of this application, embedded in the general system architecture, which will now be described.

The Client Implementation

Each BlueMatch instance is starting a client thread. In this thread, remote devices and their BlueMatch services are searched and called. After a connection to a service has been established, BlueMatch specific commands can be transmitted to the server. These commands will be described in the BlueMatch message protocol (chapter 4). In theory the client and the server together can hold a maximum of seven connections to remote devices, for

instance for simultaneous downloads. But if this limit is reached, the client and the server as well can not establish or open any new connections until a connection has been closed. Actually this limit can be lower on real devices. A minimum of two connections at a time must be allowed by the device to run BlueMatch properly, as it needs at least one connection for the client and one for an incoming connection on the server.

The Server Implementation

The threaded server in BlueMatch creates and opens a service with an universally unique identifier (UUID). This service can be searched by clients on remote devices. The connections between clients and the server are stream-oriented, based on the RFCOMM protocol. For each client that connects to the server the server accepts and opens the connection, if the maximum number of allowed connections on the device has not been reached already. Every connection is processed in a separate thread. This allows to accept more than one connection at a time. The processing is done automatically without notifying the user. This is possible, because the BlueMatch service does not rely on authorization and authentication. So the MIDlet can run in the background, for instance on multitasking operating systems like Symbian and wait for other users to connect.

Data Management

The following application data have to be managed in BlueMatch:

- list of discovered remote devices
- list of BlueMatch services on remote devices
- local files
- list of files on remote devices

- incoming files from remote devices

The lists of found remote devices and BlueMatch services are used to connect to other devices also running BlueMatch. To the user only the community list is visible, which displays the friendly names of the found devices with BlueMatch services in their SDDB. To send and receive files and filenames a special class handling files has been implemented. This class, which uses the file connection API (JSR-75), is used by the client and the server to access an existing memory card or hard disk in the phone. This is done through the MIDP-property “fileconn.dir.Memorycard”, which points to the necessary drive name. A memory card is usually the best solution in a phone to store large files, whereby some phones like the Nokia N91 [4] might have an integrated hard disk. For simplicity of the prototype only the root of the card is used by BlueMatch.

4. The BlueMatch Message Protocol

BlueMatch implements a custom message protocol. The commands a client sends to a remote service represent the use cases of BlueMatch. Each command is sent as an UTF-string, using methods in the MIDP-class DataOutputStream. Within a command additional parameters are transmitted to the service depending on the use case. Now these commands will be explained.

The Presentation

After a BlueMatch instance was started, other devices also running the BlueMatch service will be searched for and put into the local community list. To each found BlueMatch service the client introduces himself with the command

"Presentation", followed by these data in UTF-Strings:

- the member's name (device's friendly name)
- the URL of the local BlueMatch service

Receiving this message, the servers on remote devices add the new member to their community list. So their clients can connect to the newly found services as well. Only new members send their identification to the BlueMatch services on the remote devices. The clients on these devices do not have to do this on their part, as their friendly names and service URLs have already been retrieved.

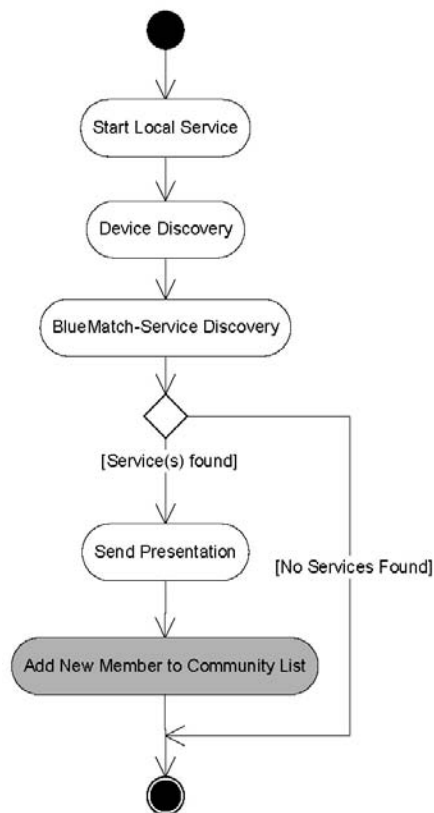


Fig. 1: Joining a BlueMatch Community

The figure 1 outlines the main steps for joining a BlueMatch community. The white activities are performed by the joining BlueMatch instance, the grey action is executed by instances in the existing community.

Retrieving File Lists

To download a member's file list, the user selects an entry in the community list and presses the button "Find Files". Now the client sends the command "GetFileList" to the selected service and opens an incoming connection to retrieve the remote device's file list.

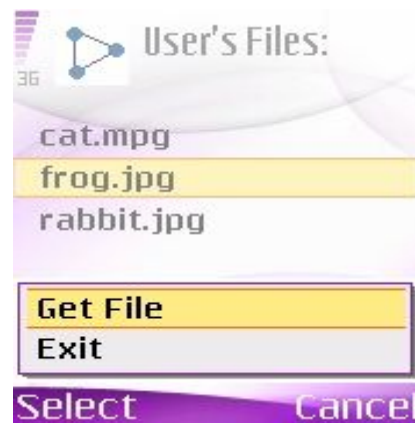


Fig. 2: User's File List

The server replies with the number of following filenames as UTF-strings. This number tells the client how many entries should be read from the server. Figure 2 shows a screenshot of a file list with further options taken from a Nokia 6680.

Downloading Files



Fig. 3: Download Process

After receiving a file list, the user can choose some files of interest to be downloaded. Therefore the client initializes a file download with the command "GetFile" and the filename as parameter. Afterwards an incoming

connection is opened to receive the file data. The requested server now returns the filename, the filesize and the file data to the client. The filesize is used on the client side to calculate the size of the download buffers and to determine, whether there is enough space left on the phone's memory card. A gauge bar visualizes the download process as shown in figure 3.

Leaving the Community

If a user decides to leave the community and closes the program, the other members should be informed, so they can delete the corresponding entry in the local community list. Therefore the client sends a "Goodbye" message with the local service URL to all found BlueMatch services. Afterwards the program quits.

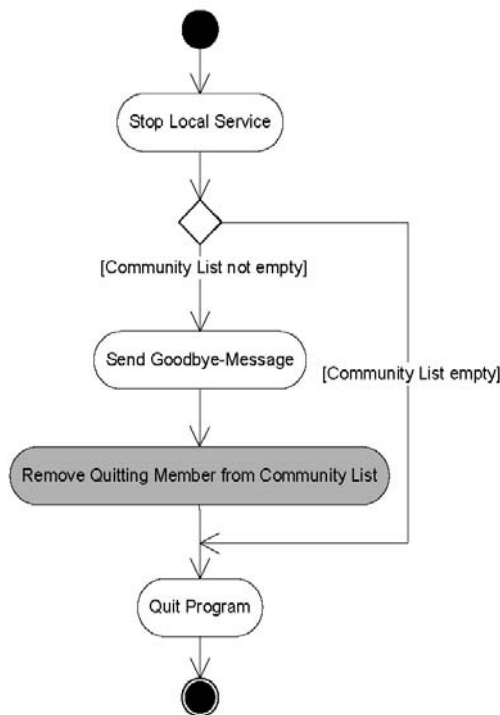


Fig. 4: Leaving the Community

Figure 4 shows these steps in an activity diagram. The white activities are performed by the leaving BlueMatch program.

5. Future Development

Now we want to list some options to improve BlueMatch. The MIDP 2.0 offers the possibility to start MIDlets automatically using timer based activation or on inbound network connections. This function is called push registry. Therefore an inbound server connection must be registered with it. This is even possible with Bluetooth server connections. This way, the program does not have to be started to be reached by remote devices. This function could be implemented in BlueMatch with some modifications.

In the presented prototype the type of the transferred files will not be checked. In future versions it is possible to restrict file transfer to encrypted DRM-files (Digital Rights Management). This will be a basic functionality to support legal superdistribution. With the transaction tracking [5] feature from OMA's (Open Mobile Alliance) DRM v2.0 it is possible to reward the content redistribution by the Rights Issuer. As one of the first devices the Nokia N91 [4] implements this new DRM standard.

Based on user transactions and file information personal profiles could be modelled to implement a mobile distributed recommendation engine to find suitable files on remote devices. The user profile could be sent as part of the presentation, so that other users have a quick overview of community members matching their taste. Papers regarding this function have been released in [6] and [7]. Furthermore the persistent storage of application data like favoured community members and recent transactions is an useful improvement.

6. References

- [1] Homepage of Bluetooth SIG; www.bluetooth.org

[2] Java APIs for Bluetooth Wireless Technology (JSR-82); www.jcp.org/en/jsr/detail?id=82

[3] Hole, K. J.; Wi-Fi and Bluetooth Course; University of Bergen; 2006; www.kjhole.com/Standards/BT/BTdownloads.html

[4] Nokia N91 Device Specification; 2006; www.forum.nokia.com/devices/N91

[5] OMA DRM v2.0 Specification; 2006 www.openmobilealliance.org/release_program/docs/DRM/drm_v2_0.html

[6] Kubek, M.; “Verteiltes Nutzer- und Content-Matching in mobilen Kommunikationssystemen im Umfeld des PotatoSystems”; Diploma Thesis; Technical University Ilmenau; 2005; www.4fo.de/de/students#kubek

[7] Nützel, J. and Kubek, M; “A Mobile Peer-To-Peer Application for Distributed Recommendation and Re-sale of Music”; 2006; AXMEDIS 2006 conference contribution

User Interfaces for People with Special Needs

Henrik Tonn-Eichstädt
Technische Universität Ilmenau
Institut für Medien- und Kommunikationswissenschaft
Am Eichicht 1, 98693 Ilmenau
henrik.eichstaedt@tu-ilmenau.de

Abstract

Assistive technologies for people with disabilities transfer information to the appropriate, usable senses and so assists the users' special abilities. Pure transformation of contents is sometimes not enough, because the transformed information only relies on a textual representation of non-text content.

At this point, it seems to be interesting how far assistive technologies can usefully go and if there is an alternative to pure transformation. This paper describes some ideas about (new) interfaces for people with special needs.

1. Introduction

In 'International Classification of Functioning and Disability' (ICIDH-1, [6]), the World Health Organization (WHO) defines three steps concerning disabilities. First, an impairment as "loss or abnormality of [...] structure or function at the organ level" is the basis which could result in a disability which is described as "restriction or lack of ability to perform an activity in a normal manner." Both, impairment and disability, can - but need not to - result in a handicap which means having a disadvantage in social life.

To mitigate such handicaps, disabled people are frequently using computers. With computers, these people learn, have assistance in real world communication (e.g. speech synthesis) or are connected to the Internet. The Internet offers an opportunity to take part in social life and to be able to act autonomously, without the help of others¹. For blind people, the Internet means a broad access to everyday life as the Internet is the first place where blind people can (inter)act

1. in unknown domains without help,
2. without being in danger of being involved in a real world accident and
3. without the danger of buying bean tins instead of pineapple tins due to the lack of vision.

Some people need to have special media to understand the presented content. Deaf and some users hard of hearing need to see text translated into sign language. Many of those people do not understand written language.

As blind and a lot of other disabled people can not rely on standard I/O-technologies, assistive devices and technology help them getting access to the computer. Foot-mice and mouthsticks are alternative devices for motor impaired persons. Screen-reader, braille displays and speech output are helpful to the blind; screen magnifier to low vision people. Screenreader transform graphical output into a textual representation which then can be rendered for perception through the sense of touch or hearing. Media that can not be transformed needs to be coded with alternative text (e.g. the alt-attribute for images in HTML). If this alternative text is missing, the content can not be

¹In Germany, this is often called 'Daseinsentfaltung' in connection with legal aspects. A translation to the English language does not exist. The term could be described as 'independent expansion of a beings' existence'.

perceived by blind users.

Though assistive technologies work fairly well, in most cases they only convert graphical output optimized for sighted users to an output of minor quality. Thus, users with special needs - especially blind users - have to cope with challenges that arise due to interfaces based on graphical paradigms. This paper deals with concepts of interfaces and interaction strategies which could be developed and optimized with regard to blind people. The basic ideas behind these presented concepts can be transferred to everyday situations where sighted people have to interact in a situation which is similar to the blind users' everyday live regarding perception prospects. These situations could be dealing with

1. audio menus of telephone applications,
2. interacting with a navigation system while driving or
3. operating a machine blindly.

2. Audio interfaces

Considering auditory output, several problems have to be taken into account. Audio is a sequential phenomenon: information has to be presented in a linear way. Parallel presentation of audio would lead to babel and results in unusable output. Further on, audio features several degrees of freedom. Sound perception is influenced by volume, pitch, timbre and loudness. A challenge is to create different sounds which vary those aspects and are doubtlessly distinguishable.

Some time ago, Edwards ([4]) tried to convey direct manipulation objects from the graphical world to the audio world. The work ends with the conclusion that much research has to be done to reach an usable interface. In principle, the approach has proven suitable. Asakawa et al. ([1]) somehow continue the work and try to transport visual effects into audio and tactile presentation. They are proposing 'background colour music' and 'foreground sound'. But

they encounter the problem of interfering sound. The authors also "[...] would like to more precisely evaluate the suitable type of information for each sense."

Röber and Masuch ([10]) investigate the sonification of objects in a 3D virtual world. By distinguishing different sounds, users can decide on moving in this world and choosing the appropriate interaction for the presented objects. This approach is demonstrated in a digital game prototype. A very similar idea is described in [8]. There, an audible space was presented to blind people who reviewed this new presentation form as usable.

The 3D approach seems to be interesting. Sighted users are arranging object icons on a 2D iconic desktop to achieve a certain order in their files and programs. Blind people cannot do that. Although screenreaders transform the visual output into a linear text form, spacial - or better surfacial - relations between the objects get lost. A clue on the distance between or the grouping of objects is not available to these users.

In order to compensate the lack of vision, blind people have evolved distinct skills in hearing. So it seems to be a good idea to think about an auditory represented room where users can place objects. [8] showed that this approach is of interest.

Looking at the sketch of such a room in Figure 1, it becomes obvious that this interface could be adequate to perceive auditory information about all existing objects. How this information can be coded is analysed in e.g. [7] for synthetic sound ('earcons') and in [12] for pieces of real sounds ('auditory icons').

At least, two main questions quickly arise:

1. **How should the objects be presented?** As a constant playing sound will counterwork usability basics and rise 'sound pollution', alternative options have to be checked. Some basic research from [11] deals with the ability of human beings to extract information from simultaneously presented sounds.

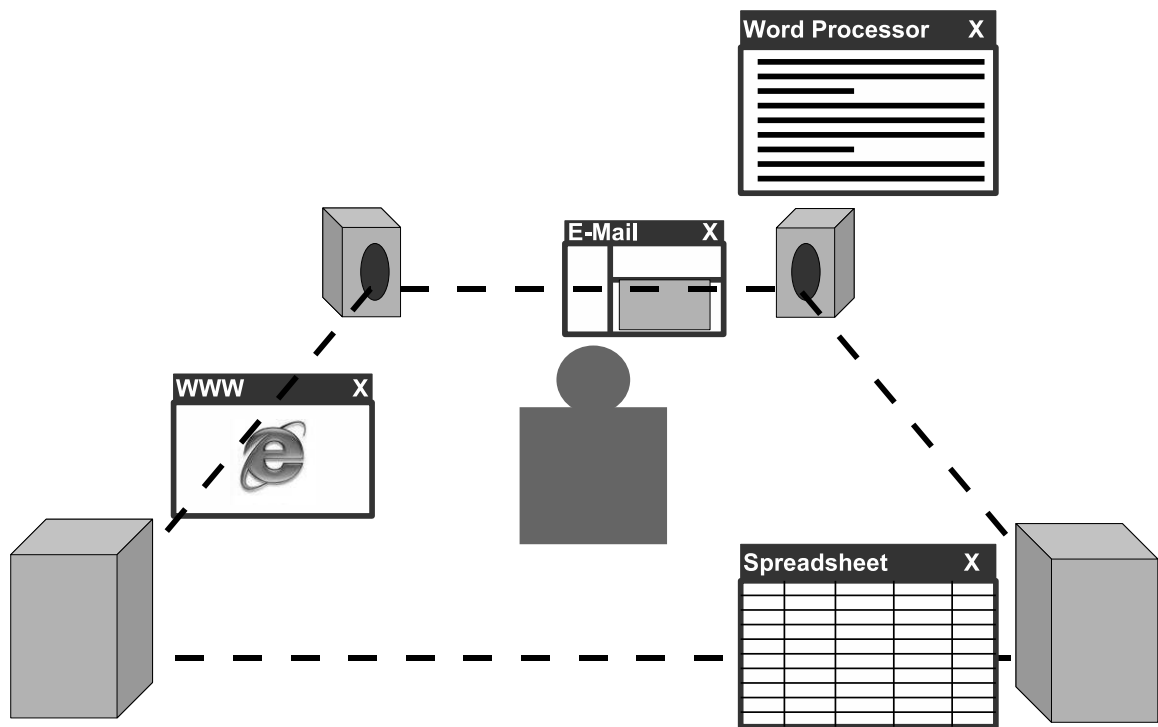


Figure 1: In a virtual auditory room, the user could place her or his applications freely in 'space'.

Based on such work, pursuing concepts can be developed.

2. **By which means can objects be placed or picked up?** There is so far no adequate everyday input device. Nowadays, the virtual room has to be turned (like in a lot of 3D-games) or the user has to wear some augmented reality device like a helmet, special glasses or a data glove. Where turning the room is a cheap escape, the described devices are vision-based and not suited for the vision impaired. A data glove as it was shown in the movie 'Minority Report' (see Figure 2) seems to be a nice idea to manipulate in a 3D environment - but will such device be affordable by the mean user?

So there is space for research in the psychological (perception research) and technical (development of input devices) field to answer these questions.

3. Tactile Interfaces

Another type of presenting information in a non-visual way are tactile interfaces. An



Figure 2: John Anderton (Tom Cruise) is scanning files in a 3D-interface. Copyright: Twentieth Century Fox und Dreamworks.

established tactile in-/output device is the braille display which is widely used by blind people. Such devices present a set of braille characters to the user (40 or 80 braille characters are common sizes). Each character is built with 6 or 8 braille pins which are arranged in a 2x3 or 2x4 matrix (see Figure 3). On those displays, users can read small portions of text at a time. After the text is read, they have to skip to the next line or element.

For reading text, the common braille displays are well usable. Users are able to read at a high word rate and can perceive

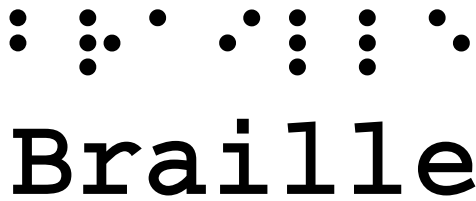


Figure 3: 'braille' written in 6-pin-braille and plaintext.

even large amounts of text quickly. But those braille displays are limited to text presentation. Graphical information can not be displayed. Even semantic information like emphasized text or headings can only be presented by adding the appropriate meta information to the text which then can be displayed: 'Heading level one: Tactile Interfaces'. So how could these 'advanced' pieces of information be considered in braille presentation?

3.1. Degrees of freedom in tactile presentation

Tactile perception is not limited to the binary information of a pin being there or not being there like in classic braille displays. So how can pressure, temperature and vibration - which are the main tactile sensations (see e.g. [3]) - be varied?

1. **Temperature.** The variation of temperature of the braille pins or the braille field itself could be one means to present information. As long as everything is in order, the temperature will be comfortable (of course, the user has to adjust her or his comfortable temperature). If an error occurs or the computer suddenly changes its state (because a backgrounded application should be focussed), the temperature could rise (or fall) to indicate this change of state to the user. Of course, the temperature has to be varied in a 'healthy' interval that the user is not hurt by a sudden change of temperature. Additionally, the difference has to be large enough to be easily

perceivable. Finally, the temperature has to be managed in a way that the user can proceed working with the computer and fix the error or pay attention to the backgrounded application. Variation of temperature seems to be useful to communicate warnings - but are there other fields of use?

2. **Pressure.** Changing the pressure of a tactile sensation can also be used to convey information. A firmly and statically presented tactile symbol has a different meaning than a softly, smoothly presented symbol. If and how these variations can be introduced in a useful way has not been researched so far. Maybe experience of force feedback devices can be used to develop new presentation forms based on pressure.
3. **Oscillation.** Brewster describes 'Tactons' that transfer structural information to oscillating braille pins ([2]). Frequency, amplitude, wave form, duration and rhythm are varied to encode structural information like emphasis, highlighting of phrases or citations. How these variations can be used in a sensible way is not clear so far. As it was the case with varying temperatures, main presentation characteristics have to be found for oscillation to obtain easily distinguishable, non-hurting and non-perturbing presentation of different information.

A variety of open questions has to be answered concerning the degrees of freedom of presented haptical information.

3.2. Two-dimensional braille fields

The second limitation of classical braille displays is the character-based presentation. This is optimal for the display of text but fails in case of any graphical output. Some graphical information could possibly be mediated by variation of temperature, pressure or oscillation as it was described above. But for graphical media, these variations alone

are not sufficient as those need more room to be displayed.

Some basic research exists in this field. In [5], graphics and charts are transferred into a tactile presentation. Rotard et al. ([9]) have developed a tactile web browser which can display text, tables and graphics on a 120 by 60 pin display. The author of this paper knows of a Japanese researcher who works on an appropriate display technique for a limited display area.

The following list shows open questions in this field.

1. Are interaction methods which are developed for size-limited, graphical mobile displays (e.g. panning or the fish-eye-view) transferable to tactile displays?
2. (How) can dynamic graphics (e.g. movies or animations) be presented on a tactile interface? In visual perception, the user can get a quick overview on an interface by just looking at it. Perceiving tactically, the user has to explore the interface with for example his or her finger tips as the areas of the skin which are sensitive enough are limited to a few sensible points.
3. How can the user be lead across a new, two-dimensional interface to make exploration of the interface easier? In visual interfaces, parts of it are coded in colour or similar functions are grouped. Using the laws of (visual) perception, the users' eye is guided through the interface. What laws of tactical perception can be used to do the same for tactile interfaces?

3.3. Input in two-dimensional tactile interfaces

Talking again about direct manipulation as an established and usable interaction paradigm: How can direct manipulation techniques be transferred to two-dimensional tactile displays? Most classic

braille displays have a 'routing key' assigned to each braille character. Pressing this routing key, the user can activate a link or place the cursor in a form field directly.

One way could be to add an input function to the braille pins. The user could then feel the information and just press the focussed area to activate an element. This of course holds some danger: as the user has to press at some strength to feel the pins, he or she could accidentally press and activate an element. Another problem could be to display different elements in a way that they are separated from each other to prevent parallel activation. And a third point: How is it presented to the user that the focussed-on element is clickable?

A second idea is to have buttons placed in a spot separated from the output braille area. These buttons would have to be connected to the displayed information which is a design challenge. But the buttons forestalls activation by accident.

So, concerning input, some research could be done.

3.4. Challenges

Summarized, what challenges are hidden in tactile interfaces?

1. Presentation of information in a (very) limited space.
2. Appropriate presentation regarding temperature, pressure and oscillation.
3. Finding usable input options.
4. Mechanical construction. Concerning the presentation and input options, the device itself has to be built:
 - (a) The device has to connect the desired functionality with the braille pins. The pins themselves have to be positioned at small distances and are tiny size.
 - (b) The device has to be produced at reasonable costs to enable users to

afford it (standard braille displays cost around 5.000 EUR and up at the moment which makes it difficult for users to afford such devices).

4. Summary

This paper gave a brief overview on possible research on user interfaces for people with special needs. What is common to both scenarios is that users could not rely on graphical output. To even this out, alternative senses are used: the senses of hearing and touch. For each of these senses, an appropriate interface is imaginable. Parts of the presented ideas are already integrated in established products. Nevertheless, these products lack some functions. Adding these functions or building enhanced devices is the aim of the described ideas. Possibly, findings from this future research can help to improve everyday devices for all users.

References

- [1] Chieko Asakawa, Hironobu Takagi, Shuichi Ino, and Tohru Ifukube. Auditory and tactile interfaces for representing the visual effects on the web. In *Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies*, pages 65–72, New York, NY, USA, 2002. ACM Press.
- [2] Stephen Brewster. Nonspeech auditory output. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, pages 220–239. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 2003.
- [3] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale. *Human-Computer Interaction (2nd edition)*. Prentice Hall Europe, Hertfordshire, 1998.
- [4] Alistair D. N. Edwards. The design of auditory interfaces for visually disabled users. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 83–88, New York, NY, USA, 1988. ACM Press.
- [5] Richard E. Ladner, Melody Y. Ivory, Rajesh Rao, Sheryl Burgstahler, Dan Comden, Sangyun Hahn, Matthew Renzelmann, Satria Krisnandi, Mahalakshmi Ramasamy, Beverly Slabosky, Andrew Martin, Amelia Lacenski, Stuart Olsen, and Dmitri Groce. Automating tactile graphics translation. In *Assets '05: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 150–157, New York, NY, USA, 2005. ACM Press.
- [6] Rolf-Gerd Matthesius, Kurt-Alphons Jochheim, and Gerhard S. Barolin. *ICIDH, International Classification of Impairments, Disabilities, and handicaps*. Huber, Bern, Switzerland, 1999.
- [7] David K. McGookin and Stephen A. Brewster. Space, the final frontearcon: The identification of concurrently presented earcons in a synthetic spatialised auditory environment. In *Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, 2004*.
- [8] Veronika Putz. Spatial auditory user interfaces. Master's thesis, Institute of Electronic Music, University of Music and Dramatic Arts Graz, Graz, Austria, October 2004.
- [9] Martin Rotard, Sven Knödler, and Thomas Ertl. A tactile web browser for the visually disabled. In *HYPertext '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 15–22, New York, NY, USA, 2005. ACM Press.
- [10] Niklas Röber and Maik Masuch. Interacting with sound - an interaction paradigm for virtual auditory works.

In *Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, 2004*.

- [11] Barbara Shinn-Cunningham and Antje Ihlefeld. Selective and divided attention: Extracting information from simultaneous sound sources. In *Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, 2004*.
- [12] Catherine Stevens and David Brennan and Simon Parker. Simultaneous manipulation of parameters of auditory icons to convey direction, size, and distance: Effects on recognition and interpretation. In *Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, 2004*.

Clustering the Google Distance with Eigenvectors and Semidefinite Programming

Jan Poland and Thomas Zeugmann
Division of Computer Science
Hokkaido University
N-14, W-9, Sapporo 060-0814, Japan
{jan,thomas}@ist.hokudai.ac.jp

Abstract

Web mining techniques are becoming increasingly popular and more accurate, as the information body of the World Wide Web grows and reflects a more and more comprehensive picture of the humans' view of the world. One simple web mining tool is called the *Google distance* and has been recently suggested by Cilibrasi and Vitányi. It is an information distance between two terms in natural language, and can be derived from the “similarity metric”, which is defined in the context of Kolmogorov complexity. The Google distance can be calculated from just counting how often the terms occur in the web (page counts), e.g. using the Google search engine. In this work, we compare two clustering methods for quickly and fully automatically decomposing a list of terms into semantically related groups: Spectral clustering and clustering by semidefinite programming.

1. Introduction

The Google distance has been suggested by Cilibrasi and Vitányi [2] as a semantical distance function on pairs of words or terms. For instance, for most of today's people, the terms “Claude Debussy” and “Béla Bartók” are much tighter related than “Béla Bartók” and “Michael Schumacher”.

The World Wide Web represents parts of the world we live in as a huge collection of documents, mostly written in natural language. By just counting the relative frequency of a term or a tuple of terms, we may obtain a *probability* of this term or tuple. From there, one may define conditional probabilities and, by taking logarithms, complexities. Li et al. [7] have proposed a distance function based on Kolmogorov complexity, which can be

used for other complexities, such as those derived from the WWW frequencies.

Spectral clustering is an increasingly popular method for analyzing and clustering data by using only the matrix of pairwise similarities. It was invented more than 30 years ago for partitioning graphs (see e.g. [11] for a brief history). Formally, spectral clustering can be related to approximating the normalized min-cut of the graph defined by the adjacency matrix of pairwise similarities [15]. Finding the exactly minimizing cut is an NP-hard problem.

The Google distances can be transformed to similarities by means of a suitable kernel. However, as such a transformation potentially introduces errors, since in particular the kernel has to be chosen appropriately and the clustering is quite sensitive to this choice, it seems natural to

work directly on the similarities. Then, the emerging graph-theoretical criterion is that of a *maximum cut*. Optimizing this cut is again NP-hard, but can be approximated with *semidefinite programming* (SDP).

The main aim of this work is to compare spectral clustering and clustering by SDP, both of which are much faster than the computationally expensive phylogenetic trees used by [2]. We will show how state-of-the-art techniques can be combined in order to achieve quite accurate clustering of natural language terms with surprisingly little effort.

There is a huge amount of related work, in computer science, in linguistics, as well as in other fields. Text mining with spectral methods has been for instance studied in [3]. A variety of statistical similarity measures for natural language terms has been listed in [13]. For literature on spectral clustering, see the References section and the references in the cited papers.

The paper is structured as follows. In the next section, we introduce the similarity metric, the Google distance, and spectral clustering as well as clustering by SDP, and we shall state our algorithms. Section 3 describes the experiments and their results. Finally, in Section 4 we discuss the results obtained and give conclusions.

2. Theory

2.1. Similarity Metric and Google Distance

We start with a brief introduction to Kolmogorov complexity (see [8] for a much deeper introduction). Let us fix a universal Turing machine (which one we fix is not relevant, since each universal machine can interpret each other by using a “compiler” program of constant length). For concreteness, we assume that its program tape is binary, such that all subsequent logarithms referring to program lengths are w.r.t. the base 2. The output

alphabet is ASCII or UTF-8, according to which character set we are actually using. (For simplicity, we shall always use the term ASCII in the following, which is to be replaced by UTF-8 if necessary.) Then, the (prefix) Kolmogorov complexity of a character string x is defined as

$$K(x) = \text{length of the shortest self-delimiting program generating } x.$$

By the requirement “self-delimiting” we ensure that the programs form a prefix-free set and therefore the Kraft inequality holds, i.e.,

$$\sum_x 2^{-K(x)} \leq 1,$$

where x ranges over all ASCII strings.

The Kolmogorov complexity is a well-defined quantity regardless of the choice of the universal Turing machine, up to an additive constant.

If x and y are ASCII strings and x^* and y^* are their shortest (binary) programs, respectively, we can define $K(y|x^*)$, which is the length of the shortest self-delimiting program generating y where x^* , the program for x , is given. $K(x|y^*)$ is computed analogously. Thus, we may follow [7] and define the *universal similarity metric* as

$$d(x, y) = \frac{\max \{K(y|x^*), K(x|y^*)\}}{\max \{K(x), K(y)\}} \quad (1)$$

We can interpret $d(x, y)$ as an approximation of the ratio by which the complexity of the more complex string decreases, if we already know how to generate the less complex string. The universal similarity metric is almost a metric according to the usual definition, as it satisfies the metric (in)equalities up to order $1/\max \{K(x), K(y)\}$.

Given a collection of documents like the World Wide Web, we define the probability of a term or a tuple of terms by counting relative frequencies. That is, for a tuple of terms $X = (x_1, x_2, \dots, x_n)$, where

each term x_i is an ASCII string, we set

$$p^{www}(X) = p^{www}(x_1, x_2, \dots, x_n) = \quad (2)$$

$$\frac{\# \text{ web pages containing all } x_1, x_2, \dots, x_n}{\# \text{ relevant web pages}}.$$

Conditional probabilities can be defined likewise as

$$p^{www}(Y|X) = p^{www}(Y \cup X)/p^{www}(X),$$

where X and Y are tuples of terms and \cup denotes the concatenation. Although the probabilities defined in this way do not satisfy the Kraft inequality, we may still define complexities

$$K^{www}(X) = -\log(p^{www}(X)) \text{ and}$$

$$K^{www}(Y|X) = K^{www}(Y \cup X) - K^{www}(X).$$

Then we use (1) in order to define the *web distance* of two ASCII strings x and y , following [2], as

$$d^{www}(x, y) = \frac{K^{www}(x \cup y) - \min\{K^{www}(x), K^{www}(y)\}}{\max\{K^{www}(x), K^{www}(y)\}} \quad (3)$$

Since we use Google to query the page counts of the pages, we also call d^{www} the Google distance. Since the Kraft inequality does not hold, the Google distance is quite far from being a metric, unlike the universal similarity metric above.

A remark concerning the “number of relevant web pages” in (2) is mandatory here. This could be basically the number of all pages indexed by Google. But this quantity is not appropriate for two reasons: First, since some months ago there seems to be no way to directly query this number. Hence, the implementation by [2] used a heuristic to estimate this value, which however yields inaccurate results. Second, not all web pages are really relevant for our search. For example, billions of Chinese web pages are irrelevant if we

are interested in the similarity of “cosine” and “triangle.” They would be relevant if we were searching for the corresponding Chinese terms. So we use a different way to fix the relevant database size. We add the search term “the” to each query, if we are dealing with English language. This is one of the most frequent words used in English and therefore gives a reasonable restriction of the database. The database size is then the number of occurrences of “the” in the Google index. Similarly, for our experiments with Japanese, we add the term \mathcal{O} (“no” in hiragana) to all queries.

2.2. Spectral Clustering

Consider the block diagonal matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Its top two eigenvectors, i.e., the eigenvectors associated with the largest two eigenvalues, are $[1 \ 1 \ 0]^T$ and $[0 \ 0 \ 1]^T$. That is, they separate the two perfect clusters represented by this similarity matrix. In general, there are conditions under which the top k eigenvectors of the similarity matrix or its Laplacian result in a good clustering, even if the similarity matrix is not perfectly block diagonal [9, 14]. In particular, it was observed in [4] that the transformed data points given by the k top eigenvectors tend to be aligned to lines in the k -dimensional Euclidean space, therefore the kLines algorithm is appropriate for clustering. In order to get a complete clustering algorithm, we therefore only need to fix a suitable kernel function in order to proceed from a distance matrix to a similarity matrix.

For the kernel, we use the Gaussian $k(x, y) = \exp(-\frac{1}{2}d(x, y)^2/\sigma^2)$. We may use a globally fixed kernel width σ , since the Google distance (3) is scale invariant. In the experiments, we compute the mean value of the entries o of the distance matrix D and then set $\sigma = \text{mean}(D)/\sqrt{2}$.

In this way, the kernel is most sensitive around $mean(D)$.

The final spectral clustering algorithm for a known number of clusters k is stated below. We shall discuss in the experimental section how to estimate k if the number of clusters is not known in advance.

Algorithm Spectral clustering of a word list

Input: word list $X = (x_1, x_2, \dots, x_n)$,
number of clusters k

Output: clustering $c \in \{1 \dots k\}^n$

1. for $x, y \in X$, compute Google relative frequencies $p^{www}(x)$, $p^{www}(x, y)$
2. for $x, y \in X$, compute complexities $K^{www}(x)$, $K^{www}(x, y)$
3. compute distance matrix

$$D = (d^{www}(x, y))_{x, y \in X}$$

4. compute $\sigma = mean(D)/\sqrt{2}$
5. compute similarity matrix

$$A = \left(\exp(-\frac{1}{2}d(x, y)^2/\sigma^2) \right)$$

6. compute Laplacian $L = S^{-\frac{1}{2}}AS^{-\frac{1}{2}}$,
where $S_{ii} = \sum_j A_{ij}$ and $S_{ij} = 0$ for $i \neq j$
7. compute top k eigenvectors $V \in \mathbb{R}^{n \times k}$
8. cluster V using kLines [4]

2.3. Semidefinite programming

Given a weighted graph $G = (V, D)$ with vertices $V = \{x_1, \dots, x_n\}$ and edge weights $D = \{d_{ij} \geq 0 \mid 1 \leq i, j \leq n\}$ which express pairwise distances, a k -way-cut is a partition of V into k disjoint subsets S_1, \dots, S_k . Here k is assumed to be given. We define the predicate $A(i, j) = 0$ if $\exists \ell [1 \leq \ell \leq k, 1 \leq i, j \leq n$ and $i, j \in S_\ell]$ and $A(i, j) = 1$, otherwise. The weight of the cut (S_1, \dots, S_k) is defined as

$$\sum_{i, j=1}^n d_{i, j} A(i, j) .$$

The max - k -cut problem is the task of finding the partition that maximizes the weight of the cut. It can be stated as follows: Let $a_1, \dots, a_k \in \mathcal{S}^{k-2}$ be the vertices of a regular simplex, where

$$\mathcal{S}^d = \{x \in \mathbb{R}^{d+1} \mid \|x\|_2 = 1\}$$

is the d -dimensional unit sphere. Then the inner product $a_i \cdot a_j = -\frac{1}{k-1}$ whenever $i \neq j$. Hence, finding the max- k -cut is equivalent to solving the following integer program:

$$\begin{aligned} \text{IP:} \quad & \text{maximize } \frac{k-1}{k} \sum_{i < j} d_{ij} (1 - y_i \cdot y_j) \\ & \text{subject to } y_j \in \{a_1, \dots, a_k\} \\ & \text{for all } 1 \leq j \leq n. \end{aligned}$$

Frieze and Jerrum [5] propose the following semidefinite program (SDP) in order to relax the integer program:

$$\begin{aligned} \text{SDP:} \quad & \text{maximize } \frac{k-1}{k} \sum_{i < j} d_{ij} (1 - v_i \cdot v_j) \\ & \text{subject to } v_j \in \mathcal{S}^{n-1} \\ & \text{for all } 1 \leq j \leq n \text{ and} \\ & v_i \cdot v_j \geq -\frac{1}{k-1} \text{ for all } i \neq j \\ & \text{(necessary if } k \geq 3). \end{aligned}$$

The constraints $v_i \cdot v_j \geq -\frac{1}{k-1}$ are necessary for $k \geq 3$ because otherwise the SDP would prefer solutions where $v_i \cdot v_j = -1$, resulting in a larger value of the objective. We shall see in the experimental part that this indeed would result in invalid approximations. The SDP finally can be reformulated as a convex program:

$$\text{CP:} \quad \text{minimize } \sum_{i < j} d_{ij} Y_{ij} \quad (4a)$$

$$\text{subject to } Y_{jj} = 1 \quad (4b)$$

$$\text{for all } 1 \leq j \leq n \text{ and} \quad (4c)$$

$$Y_{ij} \geq -\frac{1}{k-1} \text{ for all } i \neq j \quad (4d)$$

$$\text{(necessary if } k \geq 3) \quad (4e)$$

$$\text{and } Y \succeq 0. \quad (4f)$$

Here, $Y \in \mathbb{R}^{n \times n}$ is a matrix, and the last condition $Y \succeq 0$ means that Y is positive

semidefinite. Hence, Y will be a kernel matrix. Efficient solvers are available for this kind of optimization problems, such as CSDP [1] or SeDuMi [12]. In order to implement the constraints $Y_{ij} \geq -\frac{1}{k-1}$ with these solvers, actually positive slack variables Z_{ij} have to be introduced together with the equality constraints

$$Y_{ij} - Z_{ij} = -\frac{1}{k-1}.$$

Finally, in order obtain the partitioning from the vectors v_j or the matrix Y , [5] propose to sample k points z_1, \dots, z_k randomly on \mathcal{S}^{n-1} and assign each v_j to the group by the closest z_i . They show approximation guarantees generalizing those of Goemans and Williamson [6]. In practice however, the approximation guarantee does not necessarily yield a good clustering, and applying the k-means algorithm for clustering the v_j gives better results here. We use the kernel k-means (probably introduced for the first time by [10]) which directly works on the scalar products $Y_{ij} = v_i \cdot v_j$, without need of recovering the v_j . We thus arrive at the following algorithm:

Algorithm SDP clustering of a word list

Input: word list $X = (x_1, x_2, \dots, x_n)$,
number of clusters k

Output: clustering $c \in \{1 \dots k\}^n$

1. for $x, y \in X$, compute Google relative frequencies $p^{\text{www}}(x)$, $p^{\text{www}}(x, y)$
2. for $x, y \in X$, compute complexities $K^{\text{www}}(x)$, $K^{\text{www}}(x, y)$
3. compute distance matrix

$$D = (d^{\text{www}}(x, y))_{x, y \in X}$$

4. solve the SDP by using CP (cf. (4a) through (4f))
5. cluster the resulting matrix Y using kernel k-means [10]

3. Experiments

In this section, we present the experiments of our clustering algorithm applied to four lists of terms. First we show step by step how the algorithm acts on the first data set, which is the following list of 60 English words:

axiom, average, coefficient, probability, continuous, coordinate, cube, denominator, disjoint, domain, exponent, function, histogram, infinity, inverse, logarithm, permutation, polyhedra, quadratic, random, cancer, abnormal, abscess, bacillus, delirium, betablocker, vasomotor, hypothalamic, cardiovascular, chemotherapy, chromosomal, dermatitis, diagnosis, endocrine, epilepsy, oestrogen, ophthalmic, vaccination, traumatic, transplantation, nasdaq, investor, obligation, benefit, bond, account, clearing, currency, deposit, stock, market, option, bankruptcy, creditor, assets, liability, transactions, insolvent, accrual, unemployment

The first 20 words are commonly used in mathematics, the next 20 words have been taken from a medical glossary, and the final 20 words are financial terms. The matrix containing the complexities is depicted in Figure 1 (large complexities are white, small complexities black). Clearly, the single complexities on the diagonal are smaller than the pairwise complexities off-diagonal.

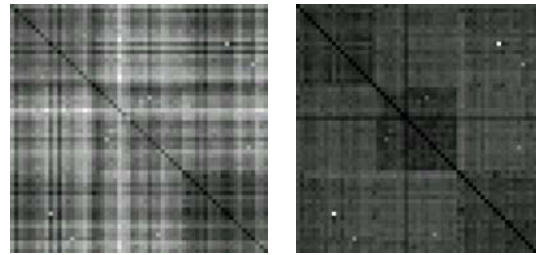


Fig. 1: Complexities

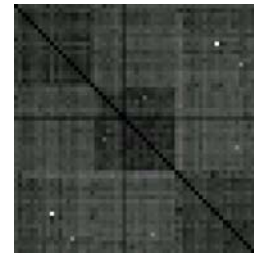


Fig. 2: Distances

In the distance matrix (Figure 2), the

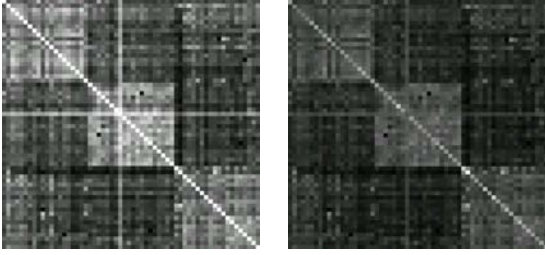


Fig. 3: Similarity Fig. 4: Laplacian

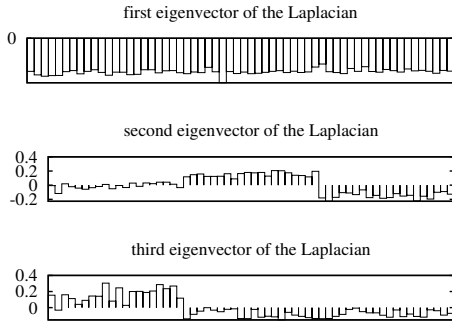


Fig. 5: Eigenvector plot

block structure is visible. After transformation to the Similarity matrix (Figure 3) and Laplacian (Figure 4), the block structure of the matrix becomes very clear. Figure 5 shows the top three eigenvectors of the Laplacian. The first eigenvector having only negative entries seems not useful at all for the clustering (but in fact it is useful for the kLines algorithm). The second eigenvector separates the medical terms (positive entries) from the union of mathematical and financial terms (negative entries). This indicates that the mathematical and financial clusters are closer related than each is related

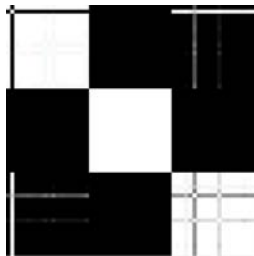


Fig. 6: Kernel matrix after SDP

to the medical terms, in a hierarchical clustering we would first split off the medical terms and then divide mathematical and financial terms. The spectral clustering correctly groups all terms except for “average”, which is assigned to the financial terms instead of the mathematical terms (this is also visible from Figure 5). As **average** also occurs often in finance, we cannot even count this as mis-clustering. We stress that the clustering algorithm is of course invariant to permutations of the data, i.e. yields the same results if the terms are given in a different order. It is just convenient for the presentation to work with an order corresponding to the correct grouping.

The same clustering result is obtained from the SDP clustering. The kernel matrix resulting from solving the SDP clearly displays the block structure, again with the exception of the term “**average**”.

In case that we do not know the number of clusters k in advance, there is a way to estimate this quite reliably from the eigenvalues of the Laplacian, if the data is not too noisy. Consider again the case of a perfect block diagonal matrix, i.e. all intra-cluster similarities are 1 and all other entries 0. Then the number of non-zero eigenvalues of this matrix is equal to the number of blocks/clusters. If the matrix is not perfectly block diagonal, we may still expect some dominant eigenvalues which are clearly larger than the others. Figure 7 top left shows this for our first example data set. The top three eigenvalues of the Laplacian are dominant, the fourth and all subsequent ones are clearly smaller. (Observe that the smallest eigenvalues are even negative: This indicates that the distances we used do not stem from a metric. Otherwise all eigenvalues should be nonnegative, since the Gaussian kernel is positive definite.)

We propose a simple method for detecting the gap between the dominant eigenvalues and the rest: Tentatively split

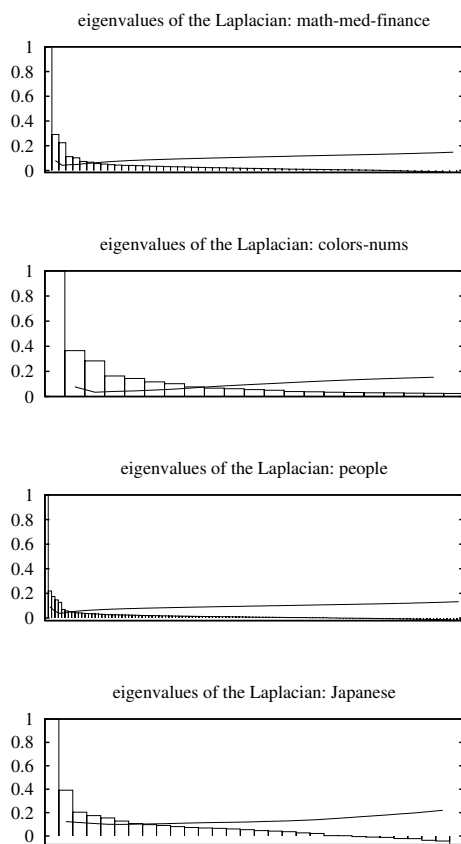


Fig. 7: Plots of the eigenvalues of the Laplacian (bars) and the s.s.e. score for determining the number of clusters (lines) for the four data sets

after the second eigenvalue, compute the means of the eigenvalues in the two groups “dominant” and “non-dominant” (ignore the top eigenvalue, which is always much larger), and calculate the sum square error (s.s.e.) of all eigenvalues w.r.t. their means. Compute this s.s.e. score also for the split after the third eigenvalue, the fourth eigenvalue and so forth. Choose the split with the lowest score. We have depicted the s.s.e. scores in Figure 7 by solid lines. For the math-med-finance data set, the minimum score is at the correct number of $k = 3$ clusters.

Note that this method works only for the spectral clustering, there is no obvious corresponding algorithm for the SDP clustering.

The next data set is the “colors-nums” data set from [2]:

purple, three, blue, red, eight,
transparent, black, white, small, six,
yellow, seven, fortytwo, five,
chartreuse, two, green, one, zero,
orange, four

Although the intended clustering has two groups, colors and numbers (where “small” is supposed to be a number and “transparent” a color), the eigenvalues of the Laplacian in Figure 7 bottom left indicate that there are three clusters. Indeed, in the final spectral clustering, “fortytwo” forms a singleton group, and “white” and “transparent” are misclustered as numbers. Clustering with SDP gives a slightly different result: Here, best results are obtained with $k = 2$ clusters, in which case only “fortytwo” is wrongly assigned to the colors.

The next data set,

Takemitsu, Stockhausen, Kagel, Xenakis,
Ligeti, Kurtag, Martinu, Berg, Britten,
Crumb, Penderecki, Bartok, Beethoven,
Mozart, Debussy, Hindemith, Ravel,
Schoenberg, Sibelius, Villa-Lobos, Cage,
Boulez, Kodaly, Prokofiev, Schubert,
Rembrandt, Rubens, Beckmann, Botero, Braque,
Chagall, Duchamp, Escher, Frankenthaler,
Giacometti, Hotere, Kirchner, Kandinsky,
Kollwitz, Klimt, Malevich, Modigliani,
Munch, Picasso, Rodin, Schlemmer, Tinguely,
Villafuerte, Vasarely, Warhol, Rowling,
Brown, Frey, Hosseini, McCullough, Friedman,
Warren, Paolini, Oz, Grisham, Osteen,
Gladwell, Trudeau, Levitt, Kidd, Haddon,
Brashares, Guiliano, Maguire, Sparks,
Roberts, Snicket, Lewis, Patterson, Kostova,
Pythagoras, Archimedes, Euclid, Thales,
Descartes, Pascal, Newton, Lagrange, Laplace,
Leibniz, Euler, Gauss, Hilbert, Galois,
Cauchy, Dedekind, Kantor, Poincare, Godel,
Ramanujan, Wiles, Riemann, Erdos, Thomas
Zeugmann, Jan Poland, Rolling, Stones,
Madonna, Elvis, Depeche, Mode, Pink, Floyd,
Elton, John, Beatles, Phil, Collins,

Toten, Hosen, McLachan, Prinzen, Aguilera,
Queen, Britney, Spears, Scorpions,
Metallica, Blackmore, Mercy

consists of five groups of each 25 (more or less) famous people: composers, artists, last year’s bestseller authors, mathematicians (including the authors of the present paper), and pop music performers. We deliberately did not specify the terms very well (except for our own much less popular names), in this way the algorithm could “decide” itself if “Oz” meant one of the authors Amos and Mehmet Oz or one of the pop music songs with Oz in the title. The eigenvalue plot in Figure 7 top right shows clearly five clusters. From the 125 names, 9 were not clustered into the intended groups using the spectral method. The highest number of incorrectly clustered names (4 mis-clusterings) occurred in the least popular group of the mathematicians (but our two names were correctly assigned). We also observed that the spectral clustering gets disproportionately harder when the number of clusters increases: Clustering only the first 50, 75, and 100 names gives 0, 2, and 5 clustering errors, respectively. We also tried clustering the same data set w.r.t. the Japanese web sites in the Google index, this gave 0, 1, 4, and 16 clustering errors for the first 50, 75, 100, and 125 names, respectively.

Clustering with SDP gives better results here: 0, 0, 1, and 4 clustering errors for the first 50, 75, 100, and 125 names, respectively.

The last data set consists of 20 Japanese terms from finance and 10 Japanese terms from computer science (taken from glossaries):

依頼, 為替, 営業, 円高, 株価, 環境, 金利, 景気, 雇用, 購入, 財政, 株価, 落札, 輸出, 税金, 売上高, 破綻, 流動性, 有価証券, 販路, 回路, 画像処理, 画像圧縮, 関数, 近似, 係数, 形式, 論理, 実験, 算術.

The eigenvalue plot in Figure 7 does not

clearly indicate the correct number of $k = 2$ clusters. However, when using $k = 2$, only the term “環境” (which means “environment”) is non-intendedly grouped with the computer science words by the spectral clustering. SDP clustering gives the same result here.

The computational resources required by our clustering algorithms are much lower than those needed by Cilibrasi and Vitányi’s algorithm [2]. Their clustering tries to optimize a quality function, a task which is NP hard. The approximation costs hours even for small lists of $n = 20$ words. On the other hand, our spectral clustering’s naive time complexity is cubic $O(n^3)$ in the number of words. This can be improved to $O(n^2k)$ by using Lanczos method for computing the top k eigenvectors.

On our largest “people” data set with $n = 125$, our spectral clustering needs about 0.11sec on a 3GHz Pentium4 processor with the ATLAS library.

Solving the SDP in order to approximate max-cut is more expensive. The respective complexity is $O(n^3 + m^3)$ (cf. [1]), where m is the number of constraints. If $k = 2$, then $m = n$ and the overall complexity is cubic. However, for $k \geq 3$, we need $m = O(n^2)$ constraints, resulting in an overall computational complexity of $O(n^6)$. On our largest “people” data set with $n = 125$, our SDP clustering needs about 2544sec.

4. Discussion and Conclusions

We have shown that it needs surprisingly little effort, just a few queries to a popular search engine together with some state-of-the-art methods in machine learning, to automatically separate lists of terms into clusters which make sense. We have focused on unsupervised learning in this paper, but for other tasks such as supervised learning, appropriate tools are available as well (e.g. SVM). Our methods are theoretically quite well founded, basing on the

theories of Kolmogorov complexity on the one hand and graph cut criteria and spectral clustering or semidefinite programming on the other hand. The SDP clustering is the more direct approach, as it needs less steps. Also, it yields slightly better results. However, it is computationally more expensive than spectral clustering.

References

- [1] B. Borchers and J. G. Young. Implementation of a primal-dual method for sdp on a shared memory parallel architecture. March 27, 2006.
- [2] R. Cilibrasi and P. M. B. Vitányi. Automatic meaning discovery using Google. Manuscript, CWI, Amsterdam, 2006.
- [3] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining(KDD)*, pages 269–274, 2001.
- [4] I. Fischer and J. Poland. New methods for spectral clustering. Technical Report IDSIA-12-04, IDSIA / USI-SUPSI, Manno, Switzerland, 2004.
- [5] A. Frieze and M. Jerrum. Improved algorithms for max k-cut and max bisection. *Algorithmica*, 18:67–81, 1997.
- [6] M. X. Goemans and D. P. Williamson. .879-approximation algorithms for MAX CUT and MAX 2SAT. In *STOC '94: Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 422–431. ACM Press, 1994.
- [7] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.
- [8] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.
- [9] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2001.
- [10] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
- [11] D. A. Spielman and S. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *IEEE Symposium on Foundations of Computer Science*, pages 96–105, 1996.
- [12] J. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(12):625–653, 1999.
- [13] E. L. Terra and C. L. A. Clarke. Frequency estimates for statistical word similarity measures. In *HLT-NAACL 2003: Main Proceedings*, pages 244–251, Edmonton, Alberta, Canada, 2003.
- [14] U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In *Advances in Neural Information Processing Systems (NIPS) 17*. MIT Press, 2005.
- [15] S. X. Yu and J. Shi. Multiclass spectral clustering. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 313–319. IEEE Computer Society, 2003.

Toward Soft Clustering in the Sequential Information Bottleneck Method

Tetsuya Yoshida

Graduate School of Information Science and Technology,
Hokkaido University

West 9, North 14, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

Abstract

This paper briefly explains the Information Bottleneck method which was originally proposed by Tishby [8], and describes our preliminary attempt to extend the Information Bottleneck (IB) method toward soft clustering. Until now four algorithms have been proposed in the framework of IB. Among them, we focus on an algorithm called sequential IB (sIB), and propose an approach for extending it toward soft clustering. We propose an approach for softening the partition in the original sIB and for cluster merger among the softened clusters. Some properties in our extension are reported, and remaining issues are discussed.

1. Introduction

By rapid progress of network and storage technologies, a huge amount of electronic data has been accumulated and available these days. Especially, as typified by Web pages, the amount of machine readable documents have been ever increasing. Since it is almost impossible for humans to manually classify or categorize the huge amount of documents over the Web, a lot of research efforts have been conducted on ext clustering or classification [2, 1, 6]

Recently, a new information theoretic principle, termed the *Information Bottleneck method* (IB), has been proposed and investigated [8, 3, 4, 7] It is based on the following idea: given the empirical joint distribution of two variables, one variable is compressed so that the compressed representation preserves the information about the other relevant variable as much as possible. It has been applied to many

application domains, especially for document clustering. One of the strength of this approach is that it is possible to find out some “structure” out of the given data itself without imposing extra assumptions or constraints.

This paper describes our preliminary attempt to extend the Information Bottleneck method toward soft clustering. Until now four algorithms are proposed in the IB method. Among them, we focus on an algorithm called sequential IB (sIB), and proposed an approach toward soft clustering. Some properties in our extension are reported and remaining issues are described.

This paper is organized as follows. Section explains the preliminaries in probabilistic approach in clustering. Section briefly explains the Information Bottleneck method based on [8, 3], and sets up the context of our research. Section describes our approach for toward soft clustering, and reports our current

status. Section gives brief concluding remarks.

2. Preliminaries

Definition 1 (KL divergence) *The Kullback-Leibler (KL) divergence between two probability distributions $p_1(x)$ and $p_2(x)$ is defined as*

$$\begin{aligned} & D_{KL}[p_1(x)||p_2(x)] \\ \triangleq & \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)} \\ = & \sum_x p_1(x) \log p_2(x) - H[p_1(x)] \end{aligned} \quad (1)$$

where $H[p_1(x)]$ represents the entropy of $p_1(x)$.

Definition 2 (JS divergence) *The Jensen-Shannon (JS) divergence between two probability distributions $p_1(x)$ and $p_2(x)$ is defined as*

$$\begin{aligned} & JS_{\Pi}[p_1(x), p_2(x)] \\ \triangleq & \pi_1 D_{KL}[p_1(x)||\bar{p}(x)] \\ & + \pi_2 D_{KL}[p_2(x)||\bar{p}(x)] \\ = & H[\bar{p}(x)] \\ & - (\pi_1 H[p_1(x)] + \pi_2 H[p_2(x)]) \end{aligned} \quad (2)$$

where $\Pi = \{\pi_i, \pi_j\}$, $0 < \pi_1, \pi_2 < 1$, $\pi_1 + \pi_2 = 1$, and $\bar{p}(x) = \pi_1 p_1(x) + \pi_2 p_2(x)$.

3. Information Bottleneck Method

3.1. IB Variational Principle

Tishby *et al.* proposed to cope with the precision complexity tradeoff without defining any distortion measure in advance [8]. They formulated the problem by introducing a *relevant* variable on which the mutual information should be

preserved as high as possible, while the give data points are compressed.

The goal is: minimizing the compression information while preserving the relevant information as high as possible. Here, the relevant information refers to the information about the relevant variable. The above problem is formulated as: given a joint statics $p(x, y)$ for a random variable X and the relevant variable Y , find a compressed representation T of the original representation X . This process is illustrated in Figure 1. To solve this problem, one observation is that T should be completely defined given X , irrelevant to Y , since T is a compressed representation of X . This observation is formulated as follows.

Definition 3 (IB Markovian relation)

$$T \leftrightarrow X \leftrightarrow Y \quad (3)$$

where X is a random variable, Y is the relevant variable, and T is a compressed representation of X .

Under IB Markovian relation, the following properties hold:

$$p(t|x, y) = p(t|x) \quad (4)$$

$$\begin{aligned} p(x, y, t) &= p(x, y)p(t|x, y) \\ &= p(x, y)p(t|x) \end{aligned} \quad (5)$$

To find the optimal compression of X into T using the method of Lagrange multipliers, Tishby *et. al* [8] suggested the following *IB variational principle*.

Definition 4 (IB variational principle)

$$\mathcal{L}[p(t|x)] \triangleq I(X; T) - \beta I(T; Y) \quad (6)$$

where $I(T; X)$, $I(T; Y)$ are the mutual information between the variables, β is a Lagrange multiplier.

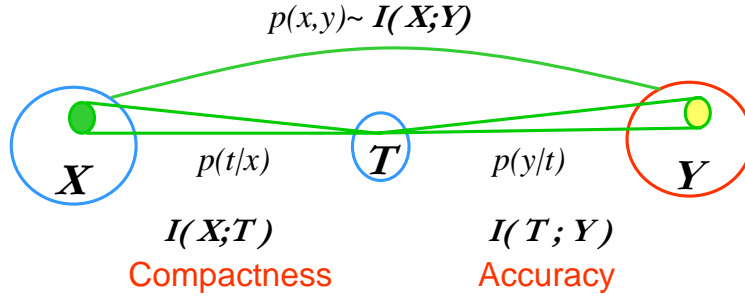


Figure 1: Information Bottleneck Framework.

Intuitively, $I(T; X)$ represents the compactness of the new representation T , while $I(T; Y)$ represents the accuracy of T for the relevant variable Y . The parameter β acts as a controlling parameter for the trade-off. The goal is to find a new representation T which minimize $\mathcal{L}[p(t|x)]$, denoted as $\mathcal{L}_{min}[p(t|x)]$.

It is proved that the optimal solution to the IB problem, *i.e.*, the conditional probability $p(t|x)$ which minimizes $\mathcal{L}[p(t|x)]$, can be determined by the following theorem [8, 3]

Theorem 5 (from [8, 3]) *Assume that $p(x, y)$, β are given and that IB Markovian relation holds. The conditional probability distribution $p(t|x)$ is a stationary point of $\mathcal{L}_{min}[p(t|x)]$ if and only if*

$$p(t|x) = \frac{p(t)}{Z(x, \beta)} e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \quad (7)$$

where $Z(x, \beta)$ is a normalization function.

Note that $D_{KL}[p(y|x)||p(y|t)]$, the Kullback-Leibler (KL) divergence between two probability distributions, emerges as the effective distortion measure from the IB variational principle.

3.2. IB Algorithms

To find out a new representation T under the IB variational principle, four

algorithms have been proposed to find out exact or approximated solutions of Eq.(6) [5, 4, 3]. There algorithms are termed as:

- iIB: an iterative optimization algorithm
- dIB: a deterministic annealing-like algorithm
- aIB: an agglomerative algorithm
- sIB: a sequential optimization algorithm

The fixed-point equations in Eq.(7) is applied in the iIB algorithm. It is reported that the sIB algorithm shows the “best” performance in terms of the time complexity and the quality of the constructed clusters, if the size of the compressed representation is rather small (*i.e.*, the number of clusters in T is small) [3]. This condition is enforced by setting the value of β in Eq.(8) sufficiently large. This results in constructing a new representation T which sustains large mutual information with the relevant variable Y . Several research efforts have been conducted by applying the sIB to document clustering¹. Thus, in the following section, we focus

¹Dr.Wang, Yaguchi, and Professor Tanaka have already utilized the sIB algorithm for document clustering in the Meme Laboratory.

on the sIB algorithm and consider its extension toward soft clustering. In the following, we explain the aIB and sIB algorithm as the preliminaries for Section .

3.3. Cluster Merger in IB

For aIB and sIB algorithms, the following dual form is considered for the optimization of $p(t|x)$:

$$\mathcal{L}_{max}[p(t|x)] \triangleq I(T; Y) - \beta^{-1}I(X; T) \quad (8)$$

Clearly, minimization of $\mathcal{L}[p(t|x)]$ in Eq.(6) corresponds to maximization of $\mathcal{L}_{max}[p(t|x)]$ in Eq.(8).

The agglomerative procedure is conducted by merging two values (clusters) t_i and t_j into a single value \bar{t} , i.e., the cluster merger $\{t_i, t_j\} \Rightarrow \bar{t}$.

Definition 6 (Merger Membership)

(from [3]) Let $\{t_i, t_j\} \Rightarrow \bar{t}$ be some merger in T . We define the cluster merger membership as the union of the events t_i and t_j as:

$$p(\bar{t}|x) = p(t_i|x) + p(t_j|x), \quad \forall x \in X \quad (9)$$

Proposition 7 (Cluster Merger in IB)

(from [3]) Let $\{t_i, t_j\} \Rightarrow \bar{t}$ be some merger in T defined through Eq.(6). If IB Markovian relation holds, then

$$p(\bar{t}) = p(t_i) + p(t_j) \quad (10)$$

$$p(y|\bar{t}) = \pi_i p(y|t_i) + \pi_j p(y|t_j) \quad (11)$$

where

$$\Pi = \{\pi_i, \pi_j\} = \left\{ \frac{p(t_i)}{p(\bar{t})}, \frac{p(t_j)}{p(\bar{t})} \right\} \quad (12)$$

To select the merger cluster, it is proved that the cluster which minimizes the following merger cost should be selected to maximize $\mathcal{L}_{max}[p(t|x)]$ in Eq.(8).

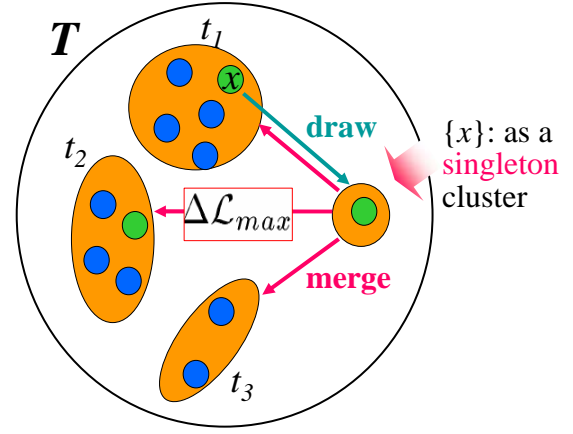


Figure 2: Overview of sIB algorithm.

Proposition 8 (Merger Cost) (from [3])

Let $t_i, t_j \in T$ be two clusters. Then, the merger cost is defined as:

$$\Delta\mathcal{L}_{max}(t_i, t_j) = p(\bar{t}) \cdot \bar{d}(t_i, t_j) \quad (13)$$

where

$$\begin{aligned} \bar{d}(t_j, t_x) &\triangleq JS_{\Pi}[p(y|t_i), p(y|t_j)] \\ &\quad - \beta^{-1}JS_{\Pi}[p(x|t_i), p(x|t_j)] \end{aligned} \quad (14)$$

Using Eq.(13), two clusters $t_i, t_j \in T$ which minimize $\Delta\mathcal{L}_{max}(t_i, t_j)$ are selected and merged iteratively.

3.4. sIB: a sequential optimization algorithm

In [4], Slonim *et al.* proposed a framework for casting a known agglomerative algorithm into a sequential K -means like algorithm, and proposed a sequential optimization algorithm called sIB. The sequential maintains a (flat) partition in T with exactly m clusters ($t_1 \dots t_m \in T$).

Given the initial partition, at each step some data $x \in X$ is drawn from its current cluster t (x was exclusively assigned to t), and represented as a new singleton cluster $\{x\}$. Then, the singleton cluster $\{x\}$ is merged with

or inserted into t_j such that $t_j^{new} = \operatorname{argmin}_{t_j \in T} \Delta \mathcal{L}_{max}(\{x\}, t_j)$, $t_j^{new} = t_j \sqcup \{x\}$. If $t_j \neq t$, this process constitutes a reassignment of x , which contributes to increasing the value of $\mathcal{L}_{max}[p(t|x)]$ in Eq.(8). The above process is illustrated in Figure 2.

4. An Extension toward Soft Clustering in sIB

In the original sIB algorithm, hard clustering is assumed, *i.e.*, each data $x \in X$ is assigned to only one cluster $t \in T$. Hard clustering is formulated as:

$$p(t|x) = \begin{cases} 1 & x \in t \\ 0 & x \notin t \end{cases} \quad (15)$$

However, the original IB variational principle in Definition 3 does not impose this (extra) constraint. Actually, the solution found by the iIB algorithm, which applies the standard strategy in variational methods, does not necessarily obey this constraint and conducts soft clustering.

4.1. An Approach toward Soft Clustering in sIB

To extend sIB algorithm toward soft clustering, we employ the following approach:

- a) “cluster division & merger” strategy
As in the original sIB, $\forall x \in X, \exists t \in T$ s.t. $x \in t$, divide the cluster t into $\{t', t_x\}$. Here, $t \Rightarrow t' \cup t_x$, t_x is treated as a singleton cluster but with probabilistic assignment, as explained below. Then, t_x is merged with some $t_j \in T$ as $\{t_x, t_j\} \Rightarrow \bar{t}_j$.
- b) sequential update for each $x \in X$ for simplicity
One drawback of this approach is that, it requires the $O(|X|)$ loop. However, as a initial trial, we employ sequential update for simplicity.

```

procedure  $sIB_{soft}(p(x, y), \beta, |T|)$ 
//  $p(x, y)$ : joint distribution
//  $\beta$ : trade - off parameter
//  $|T|$ : cardinality of clusters
// initialization
1: For every  $x \in X$ 
2: For every  $t \in T$ 
3: Set random value for  $p(t|x) \in [0, 1]$ 
   s.t.  $\sum_t p(t|x) = 1$ 
// Main Loop
4: While not Done
5: Done  $\leftarrow TRUE$ 
6: For every  $x \in X$ 
7: Select  $t \in T$ 
8: Divide  $t$  into  $\{t', t_x\}$ 
9: Select  $t_j \in T \setminus \{t\} \cup \{t'\} \cup \{t_x\}$ 
10: If  $t_j \neq t$ ,
11:   Done  $\leftarrow FALSE$ 
12: Merge  $t_x$  into  $t_j$ 
13: Return  $p(t|x)$ 

```

Figure 3: Pseudo-code of soft clustering in sIB.

Difference from the original hard clustering in sIB is that, a data $x \in t$ is not entirely drawn from the cluster t . Instead, only some portion of x is drawn and assigned to a new cluster t_x .

Our procedure of soft clustering in sIB is summarized as the procedure sIB_{soft} in Figure fig:sIB-psuedo-code. The extension from the original sIB algorithm is the lines 7,8,9.

To realize the above soft clustering, the following issues should be resolved.

- Selection of the cluster $t \in T$ for each $x \in X$ s.t. $x \in t$ (line 7)
- Soft partitioning of t into $\{t', t_x\}$. (line 8)
- Selection of the merger cluster t_j (line 9)

In the following sections, we describe our current approach for the above issues.

4.2. Soft Partitioning and Merger

Our approach for soft partitioning in sIB is formulated as follows:

Definition 9 (Soft Partitioning) For each $x \in X, \exists t \in T$ s.t. $x \in t$, let $t \Rightarrow \{t', t_x\}$ be a soft partitioning of the cluster t with a parameter γ ($0 < \gamma \leq 1$). Then, $\forall x' \in X, \forall t_j \in T$,

$$p(t_x|x') = \begin{cases} \gamma p(t|x) & x' = x \\ 0 & x' \neq x \end{cases} \quad (16)$$

$$p(t'|x') = \begin{cases} (1 - \gamma)p(t|x) & x' = x \\ p(t|x') & x' \neq x \end{cases} \quad (17)$$

For the rests, i.e., $\forall t_j \in T$ s.t. $t_j \neq t' \wedge t_j \neq t_x, \forall x' \neq x, p(t_j|x')$ unchanged.

For the clusters which are defined in Definition 9, we follow the cluster merger membership in Definition 6. Thus, in our soft clustering, if clusters $\{t_i, t_j\}$ are merged into \bar{t} , the cluster merger membership is set as

$$p(\bar{t}|x) = p(t_i|x) + p(t_j|x), \quad \forall x \in X$$

We need to The value of γ ($\gamma \in [0, 1]$)

4.3. Some Properties in Soft Clustering

In hard clustering, $JS_{\Pi}[p(x|t_i), p(x|t_j)]$ in Eq.(14) can be simplified as $JS_{\Pi}[p(x|t_i), p(x|t_j)] = H[\Pi]$ ([3], p.36). However, this does not hold for soft clustering. Fortunately, similar relation to Eq.(11) holds.

Proposition 10 Let $\{t_i, t_j\} \Rightarrow \bar{t}$ be some merger in T defined through Eq.(6). If IB Markovian relation holds, then

$$p(x|\bar{t}) = \pi_i p(x|t_i) + \pi_j p(x|t_j) \quad (18)$$

Note that Eq.(18) holds for both hard and soft clustering.

Proposition 11 Let $t \Rightarrow \{t_x, t'\}$ be some soft partitioning of a cluster $t \in T$ defined in Definition 9. If IB Markovian relation holds, then

$$p(t) = p(t') + p(t_x) \quad (19)$$

$$p(y|t_x) = p(y|x) \quad (20)$$

$$p(x'|t_x) = \begin{cases} 1 & x' = x \\ 0 & x' \neq x \end{cases}$$

Eq.(19) shows that Eq.(10) holds in both hard and soft clustering.

4.4. Cluster Merger Criterion

Observation 12 Let $t_i, t_j \in T$ be two clusters and consider the cluster merger $\{t_i, t_j\} \Rightarrow \bar{t}$. For soft clustering, if we set conditional probability in the merged cluster $p(\bar{t}|x)$ following Eq.(6), from Proposition 12, Eqs.(10) and (11) hold. Thus, Merger cost $\Delta\mathcal{L}_{max}(t_i, t_j)$ in Eq.(13) holds both for hard and soft clustering.

From Observation 12, for our definition of soft partitioning in Eqs.(16) and (17), we can utilize $\Delta\mathcal{L}_{max}(t_x, t_j)$ as the merger cost of $\{t_x, t_j\}$ into \bar{t}_j . Thus, we can set line 9 in Figure 3 as ‘‘Select $t_j = \operatorname{argmin}_{t \in T} \Delta\mathcal{L}_{max}(\{t_x\}, t)$ ’’.

Based on the above observation, we follow the criterion in Eq.(8) and select the cluster $t_j \in T$ which minimizes $\Delta\mathcal{L}_{max}(t_x, t_j)$ in Eq.(13). There are two cases for cluster merger:

- Cluster re-merge: t' is selected and merged as $t' \Rightarrow t' \cup t_x = t$ (as \bar{t})
 $\forall t_j \in T$ s.t. $t_j \neq t'$,
 $\Delta\mathcal{L}_{max}(\{t_x\}, t') \leq \Delta\mathcal{L}_{max}(\{t_x\}, t_j)$

- cluster merge: $\exists t_j \in T$ s.t. $t_j \neq t', t_j \Rightarrow t_j \cup t_x = \bar{t}_j$
 $\exists t_j \in T$ s.t. $t_j \neq t'$,
 $\Delta \mathcal{L}_{max}(\{t_x\}, t') > \Delta \mathcal{L}_{max}(\{t_x\}, t_j)$

For cluster re-merge,

$$p(y|t) = \pi_{t'}p(y|t') + \pi_{t_x}p(y|t_x) \quad (21)$$

$$\begin{aligned} \Pi &= \{\pi_{t'}, \pi_{t_x}\} \\ &= \left\{ \frac{p(t')}{p(t)}, \frac{p(t_x)}{p(t)} \right\} \\ &= \{1 - \gamma p(x|t), \gamma p(x|t)\} \end{aligned} \quad (22)$$

Here, $p(x|t)$ corresponds to the weight (contribution) of $x \in t$ for $p(t)$

$t_j = t'$ corresponds the re-merger to the original cluster t ($\{t_x, t'\} \Rightarrow \bar{t} = t$); $t_j \neq t'$ corresponds to a cluster update.

4.5. Cluster Selection

There are two ways to conduct cluster selection.

- No a priori selection:
 $\forall x \in X, \forall t \in T$ s.t. $x \in t$,
select the cluster t which minimize $\Delta \mathcal{L}_{max}(\{t_x\}, t_j)$.
Since this approach requires at least $O(XT)$ loops, as in aIB, the worst case time complexity is at least $O(|X|^2)$. However, if we can assume that $|T|$ is fixed and that $|T| \ll |X|$, this does not matter in practice.
- Heuristic selection:
Select the most “distant” cluster t for each x .

Our current plan is to utilize the latter approach and select a cluster $t \in T$ for each $x \in X$. Since the framework of IB conducts clustering based on the conditional distribution with respect to the relevant variable Y (see Eq.(7)), we plan to

utilize the measure for conditional distributions.

4.6. Remaining Issues in Soft Clustering

We have described our current status for the extension toward soft clustering in sIB. The followings are the remaining issues resolved:

- i). Selection of the most “distant” cluster t for x
- ii). Value of γ

For i.), as stated in the previous subsection, one heuristic approach for cluster selection is to select the most distant cluster w.r.t. the (conditional) probability distribution for the relevant variable.

Among many diversity measure for probability distributions, it would be natural to utilize either D_{KL} in Eq.(1) or JS_{Π} in Eq.(2), because these diversity measures emerged out of the framework of IB. At this moment we have not yet decided which diversity measure, D_{KL} in Eq.(1) or JS_{Π} in Eq.(2), should be utilized to measure the distance.

For ii.), ideally, it is desirable to show the analytic solution for the value of γ . If it is difficult or impossible, we plan to utilize uniform or adaptive value, *i.e.*,

- 1) uniform for $\forall x \in X$, or,
- 2) adaptive for each x .

Our current plan for determining the value of γ is as follows. We assume that the cluster t s.t. $x \in t$ is already selected (hopefully using the diversity measure as described above). Once the cluster t is fixed, for each $t_j \in T \setminus t \cup \{t'\} \cup \{t_x\}$, since $p(t_j)$, $p(y|t_j)$, and $p(x|t_j)$ are calculated beforehand, $\Delta \mathcal{L}_{max}(\{t_x\}, t_j)$ can be expressed as a function of γ . Thus, we can represent $\Delta \mathcal{L}_{max}(\{t_x\}, t_j)$

as $\Delta\mathcal{L}_{max}(\gamma)$ for each t_j . We plan to utilize the variational method and determine the stationary point of γ for each t_j , and utilize the value to compute $\Delta\mathcal{L}_{max}(\gamma)$.

5. Concluding Remarks

This paper briefly explained the Information Bottleneck method which was originally proposed by Tishby [8] and substantially extended in [3], and described our preliminary attempt to extend the Information Bottleneck (IB) method toward soft clustering. Until now four algorithms have been proposed in the IB method. Among them, we focus on an algorithm called sequential IB (sIB), and proposed an approach toward soft clustering. We proposed an approach for softening the partition in the original sIB and for cluster merger among the softened clusters.

Our current status is very preliminary and there are many remaining issues, e.g., the selection of the cluster to divide, determination of the value of γ . We plan to tackle these issues in near future. We also plan to implement the described approach and conducts experiments to evaluate our approach.

Acknowledgments This work is partially supported by JSPS Core-to-Core Program “Center for Research on Knowledge Media Technologies for the Advanced Federation, Utilization and Distribution of Knowledge Resources” and The 21st Century COE Program “Meme-Media Technology Approach to the R&D of Next-Generation Information Technologies” The author is grateful to Professor Yuzuru Tanaka for encouraging this path of research in this project.

References

- [1] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional Word Clusters vs. Words for Text Categorization. *Journal of Machine Learning Research*, 3:1183–1208, 2003.
- [2] F.C. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.
- [3] N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, School of Computer Science and Engineering, Hebrew University, 2002.
- [4] N. Slonim, N. Friedman, and N. Tishby. Unsupervised Document Classification using Sequential Information Maximization. In *Proceedings of the 25th International Conference on Research and Development in Information Retrieval (SIGIR)*, 2002.
- [5] N. Slonim and N. Tishby. Agglomerative Information Bottleneck. In *Advances in Neural Information Processing Systems (NIPS) 12*, pages 617–623, 1999.
- [6] N. Slonim and N. Tishby. The Power of Word Clusters for Text Classification. In *Proceedings of the 23rd European Colloquium on Information Retrieval Research*, 2001.
- [7] N. Tishby. Efficient Data Representations that Preserve Information. In *Proceedings of the 6th International Conference on Discovery Science (DS2003)*, page 45, 2003.

- [8] N. Tishby, F. Pereira, and W. Bialek. The Information Bottleneck Method. In *Proceedings of the 37th Allerton Conference on Communication and Computation*, 1999.

Appendix

Proof of Eq.(18)

$$\begin{aligned}
p(x, \bar{t}) &= \sum_y p(x, y) p(\bar{t}|x, y) \\
&= \sum_y p(x, y) p(\bar{t}|x) \\
&= \sum_y p(x, y) (p(t_i|x) + p(t_j|x)) \\
&= \sum_y p(x, y) p(t_i|x) \\
&\quad + \sum_y p(x, y) p(t_j|x) \\
&= \sum_y p(x, y) p(t_i|x, y) \\
&\quad + \sum_y p(x, y) p(t_j|x, y) \\
&= \sum_y p(x, y, t_i) + \sum_y p(x, y, t_j) \\
&= p(x, t_i) + p(x, t_j) \\
&= p(t_i) p(x|t_i) + p(t_j) p(x|t_j) \\
\therefore p(x|\bar{t}) &= \pi_i p(x|t_i) + \pi_j p(x|t_j)
\end{aligned}$$

where $\Pi = \{\pi_i, \pi_j\} = \left\{ \frac{p(t_i)}{p(\bar{t})}, \frac{p(t_j)}{p(\bar{t})} \right\}$

Proof of Eq.(19)

$$\begin{aligned}
p(t_x) &= \sum_{x'} p(t_x, x') \\
&= p(t_x, x) = p(t_x|x) p(x) \\
&= \gamma p(t|x) p(x) \\
&= \gamma p(t, x) \\
p(t') &= \sum_{x'} p(t', x') \\
&= \sum_{x'} p(x') p(t'|x') \\
&= \sum_{x' \neq x} p(x') p(t|x') \\
&\quad + p(x) ((1 - \gamma) p(t|x)) \\
&= \sum_{x'} p(x', t) - p(x) \gamma p(t|x) \\
&= p(t) - p(t_x) \\
\therefore p(t) &= p(t_x) + p(t')
\end{aligned}$$

Proof of Eq.(20).

$$\begin{aligned}
p(y|t_x) &= p(y, t_x) / p(t_x) \\
&= \frac{\sum_{x'} p(x', y) p(t_x|x', y)}{\sum_{x'} p(t_x|x') p(x')} \\
&= \frac{\sum_{x'} p(x', y) p(t_x|x')}{\sum_{x'} p(t_x|x) p(x')} \\
&= \frac{p(x, y) p(t_x|x)}{p(t_x|x) p(x)} \\
&= \frac{p(x, y)}{p(x)} \\
&= p(y|x)
\end{aligned}$$

Proof of Eq.(21).

$$\begin{aligned}
p(x|t_x) &= \frac{p(t_x|x) p(x)}{p(t_x)} \\
&= \frac{p(t_x, x)}{\sum_{x'} p(t_x, x')} \\
&= \frac{p(t_x, x)}{p(t_x, x)} \\
&= 1 \\
p(x'|t_x) &= 0 \quad (x' \neq x)
\end{aligned}$$

Multiple News Articles Summarization based on Event Reference Information with Hierarchical Structure Analysis

Masaharu Yoshioka Makoto Haraguchi

Graduate School of Information Science and Technology, Hokkaido University
N14 W9, Kita-ku, Sapporo-shi, Hokkaido, JAPAN
{yoshioka,mh}@ist.hokudai.ac.jp

Abstract

In this paper, we describe our method to generate summary from multiple news articles. Since most of news articles report several events and these events are refereed with following articles, we use this event reference information to calculate importance of a sentence in multiple news articles. We also propose a method to delete redundant description by using similarity of events. Finally we discuss its effectiveness based on the evaluation result.

1. Introduction

Nowadays, we can access a large amount of text data. As a result, even for a simple topic, it becomes difficult to read through all documents that are related to the topic. Therefore, demand for multiple document summarization is increasing and a task for multiple document summarization that uses news articles is proposed for tackling this issue in TSC-3[4].

Most significant characteristic of multiple document summarization compared with single one is that there are redundant information in a document set. For example, most of news articles reports events that occurred at particular date and these events were refereed in following articles. Based on the assumption that repetitions of the same event description represent the relationship among different articles, we have already proposed a method to summarize multiple news article by using event reference information [10].

By using TSC-3 data, we show the usage of event reference information is useful. However this method tends to select sentences that include many repetitive events. As a result, sentences that have similar repetitive

events among different news articles gains high score and sentences that are important only in one article are not selected as important sentences. In order to solve this problem, we propose a method to combine important sentence extraction method for each document and for whole document for reducing this effect in this paper.

2. Multiple News Articles Summarization based on Event Reference Information

2.1. Extraction of the Event Reference Information

As McKeown proposed in [7], there are several categories for multiple document summarization. Single-event summarization is a category that aims to summarize a set of articles includes ones that reports occurrence of events and ones that reports following events (e.g., real fact of the event and sequel of the event). In this type of articles sets, following articles refer to the events that were already described in previous articles and add another information that were related to previous events. Therefore, we assume identification of events in different articles and reference information among these events is

useful to make a summary.

Lexical cohesion method is one approach to deal with this reference information for summarizing a single document. However, in order to deal with reference information in different documents, we cannot use information such as distance between two sentences.

So we propose a method to extract event information from news articles and to identify event by using similarity measure between two events. In this paper, we define “*Event*” as follows.

Event is information that describes facts and related information on particular date.

Extraction of Events

Event is a unit to represent relationship among different articles and it should have information that is useful to identify same events. In order to extract rich event information from sentences in a document, it is better to analyze deep structure of sentences in a document; e.g., discourse analysis and anaphoric analysis. However, it is very difficult to use such deep structural information, we decide to use results of dependency analysis and we set a size of a unit simple. In addition, date information is useful for discrimination of similar events; e.g., press release in May is different from press release in April.

Based on this discussion, we select following slots to define an event.

Root is a word that dominates an event (verb that represents action or noun that represents subject or object)

Modifier is words that modify root word. Words are categorized into several groups, such as subject and object words for verbs and adjective and adnominal words for nouns.

Negative represents modality of expression.

Depth is a path length between Root of the event and root of the sentence in dependency analysis tree.

Date is a date that characterize the event. This slot is not a required slot to define an event.

ArticleDate is a date that the article was published.

Chunks represents list of word positions in a sentence.

In this method, we extract event information from a sentence by using following steps.

1. We apply Cabocha[5] to obtain dependency analysis tree.
2. We select verbs and nouns that have modification words as candidates of “Root” for events.
3. We check whether negative expression is included in root or not and set “Negative” based on this analysis.
4. We extract “Modifier” information from dependency analysis tree. At this time, we classify types of modifier by using POS tag and postpositional particle (postpositional particle with “か” and “は” are categorized into “Subject” and other postpositional particle are categorized into “postpositional-postpositional particle” (e.g., “postpositional-に”). Modifier information includes not only words that directly dependent on Root word but also modifiers for modifier words. Modifiers for a modifier word are categorized into the same category of the modifier words.
5. When we can extract date information from the sentence, we set this date as “Date” for events that has dependency with date words.
6. “Article Date” is obtained from article information.

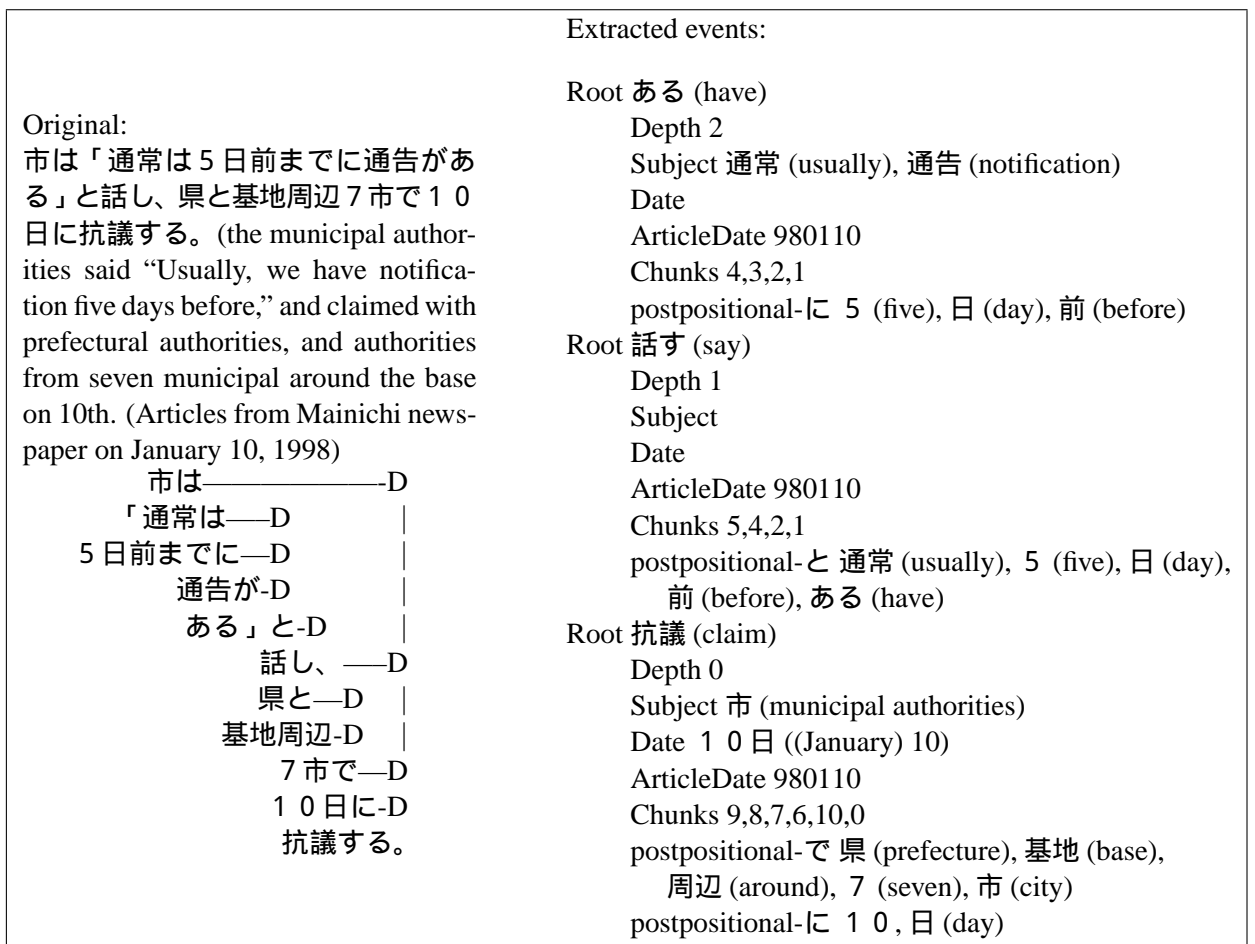


Figure 1: Example of Event Extraction from a sentence

7. “Depth” and “Chunks” are calculated by comparing event information with the dependency analysis tree.

Figure 1 shows a set of original sentence and extracted events.

Dealing with Event Reference Information

We have already proposed an algorithm to calculate importance of a sentence in a single document based on PageRank [1] algorithm [2].

PageRank algorithm is the one that can calculate importance of WWW pages by using link analysis. Basic concept of the algorithm is distribution of page importance through link structure; i.e., page that has many links collects importance from other page and links from page with higher importance has higher importance compared to the

links from ones with lower importance.

We formalize important sentence extraction algorithm by using following correspondence between web link structure and sentence structure.

A page in PageRank corresponds to a sentence.

A link in PageRank corresponds to sharing of same words in two different sentences (A and B). Since it is difficult to determine the direction of the link, we formalize that there are two links (A to B and B to A) for one sharing word.

In addition, all links in a page have same probability to distribute its importance in PageRank. However, in important sentence extraction, all words do not have same importance; e.g., sharing of large numbers of words has closer relationship than sharing of small numbers of words and sharing of rare

words has closer relationship than sharing of common words. Therefore, we calculate importance of link based on the role of shared word in a sentence and Inverted Document Frequency (IDF). This important measure affects a transition matrix of PageRank that is used to calculate distribution of importance.

In this formalization, we merge all documents as a single document and calculate importance by using same algorithm for a single document.

Therefore, we employ latter approach to calculate importance of all sentences.

In previous algorithm, links are generated when two different sentences shares same word(s). In order to handle event reference information, we modify link generation algorithm.

Since we can express same events in different ways, identification of same event is difficult to identification of same word. Therefore, we introduce similarity measure between two events by using following two criteria.

1. Similarity of words

First, we compare all words in “Root” and each category in “Modifiers.” For each category, we calculate existence ratio of same word(s) that is biased by IDF. Second, we calculate weighted average of existence ratio for all categories. “Root” and “Subject” in “Modifiers” has higher weight compared to the other ones. We set threshold value to check whether the event pair belongs to candidate similar event pairs or not.

2. Judgment of consistency of date

When the event has “Date” information, we verify consistency of date. When “Date” information lacks specific date such as year and/or month, we complement this information by using “ArticleDate” information. When one article has “Date” information and other one does not have this information, we compare them by using “Ar-

ticleDate” information. When inconsistency is found, the pair of events is removed from candidate similar event pairs.

We generate links between two different sentences that shares candidate similar event pair(s). We also calculate importance of link based on the importance of events in a sentence. Since we assume most important issues discussed in a sentence are located at the end of the sentence, we calculate importance of event based on the “Depth” information. Another possible measure to calculate this importance is a frequency based measure. However, since frequency of events is already considered to calculate importance as a number of links, we do not use this measure.

We also set direction of the link as previous one; i.e., we formalize that there are bidirectional two links for one similar event pair.

We applied this event identification method for test run data and we found some links between two related events are missing. One reason of this problem is that we can express same event by using different vocabulary. However, we found some words are shared for these event sets. Therefore, we also use sharing of same word(s) for handling these relationships.

In PageRank algorithm, importance of page is calculated as a convergent vector of following recurrence formula.

m_{ij} is an element of transition matrix M of i -th row and j -th column and represents transition probability from j -th sentence to i -th sentence based on link structure. Since $\{m_{1j}, m_{2j}, \dots, m_{nj}\}$ represents transition probability, $\sum_{i=1}^n m_{ij} = 1$. When there is a sentence k that has no relationship with other sentence, we set $m_{ik} = 0$.

$$\vec{r} = M\vec{r} \quad (1)$$

In order to handle both types of links (sharing events and sharing words), we make two matrices $M_e(me_{ij})$ and $M_w(mw_{ij})$ that corresponds to transition matrix made by shar-

ing events and one made by sharing words, respectively. In order to satisfy constraint $\sum_{i=1}^n m_{ij} = 1$, overall transition matrix $M(m_{ij})$ is calculated by using parameter β and following formula.

$$\begin{aligned}
m_{ij} &= \beta * me_{ij} + (1 - \beta) * mw_{ij} \\
&\text{when } \sum_{i=1}^n me_{ij} \neq 0, \sum_{i=1}^n mw_{ij} \neq 0 \\
m_{ij} &= me_{ij} \\
&\text{when } \sum_{i=1}^n mw_{ij} = 0 \\
m_{ij} &= mw_{ij} \\
&\text{when } \sum_{i=1}^n me_{ij} = 0
\end{aligned}$$

Since a sentence that has no relationship with other sentence is meaningless to include into an abstract, we remove rows and columns that corresponds to the sentence in the matrix M . Calculation of a convergent vector is conducted by using eigen vector calculation. Since convergent vector satisfies following formula, convergent vector is an eigen vector of matrix M with eigen value = 1.

$$\vec{r}_c = M\vec{r}_c \quad (2)$$

Usage of Sentence Position

Since, in news articles, important sentences may be described early part of each article, we use sentence position information for calculating importance of each sentence. In PageRank, an algorithm to set initial importance of each page is proposed as Topic-Sensitive PageRank [3]. This algorithm is proposed to calculate importance of each page based on the category that the page belongs to. In this algorithm, they modify recurrence formula of PageRank as follows. \vec{v} corresponds to initial importance vector and α is a parameter to control strength of the effect by the vector.

$$\vec{r}_c = (1 - \alpha) * M\vec{r}_c + \alpha * \vec{v} \quad (3)$$

In this paper, we use simple formula $1/\log(n+1)$ (n : sentence number in an article) for initial value of \vec{v} . \vec{v} is normalized with $\sum_{i=0}^m v_i = 1$ (m is number of all sentences and v_i is an initial importance value for i -th sentence).

By using the algorithm discussed above, since similar sentences have similar links, similar sentences have similar scores. As a result, there is a chance to select redundant sentences when we select sentences from higher score ones. Therefore, we need a mechanism to detect such redundant sentences and it is required to remove such redundant description[6].

In this research, we use similarity measure of two events to calculate redundancy of new description. Since we can describe same information by using different numbers of sentences, we do not compare sentences one by one. We decompose a sentence into a set of events and we check redundancy of a sentence by comparing with an extracted event set that is obtained from extracted sentence; i.e., we calculate weighted average of existence ratio of events in the sentence. Since we assume most important issues discussed in a sentence are located at the end of sentence, we set higher weight for an event with lower ‘‘Depth.’’

By using this redundancy check mechanism, sentence extraction algorithm is as follows.

1. Construction of an initial extracted event set
A sentence with highest importance is selected as an extracted sentence. An initial extracted event set is constructed from events in the selected sentence.
2. Redundancy check and addition of new description
Our system tries to add new description from a sentence with higher importance. The system checks redundancy of the sentence and add it when it does not exceed predefined redundancy level. The system also adds events in the selected sentence to the extracted

event set. This step reiterates to select a desired number of sentences.

2.2. Experiment and Discussion

We apply this system for the task of TSC-3 [4]. In TSC-3, there are two subtasks. One is sentence extraction and the other is abstraction. In this system, we do not use “set of questions about important information of the document sets” given by task organizer.

In order to evaluate the the effectiveness of using event sharing information, we conduct following three different experiments. Since LEAD method that selects important sentence from the beginning of the articles does not work well, we set $\alpha = 0.1$ for the evaluation.

Event only Transition matrix made only from event references is used for calculating importance ($\beta = 1.0$).

Word and Event Transition matrix that combines event and word references is used for calculating importance ($\beta = 0.3$).

Word only Transition matrix made only from word references is used for calculating importance ($\beta = 0.0$).

Table 1 shows a result of these experiments. From this result, calculation of importance by using event reference only has poorer performance than others. Short is a task to select about 5% important sentence and long is a task to select about 10%. Since multiple document summarization has a similar sentences in different document, a correct answer for each sentence extraction is defined as a set of corresponding sentences. In order to take into account the redundant sentence selection, following two measures are used in TSC-3[4]. Please consult [4] for detail.

Precision is the ratio of how many sentences in the system output are included in the set of the corresponding sentences.

Coverage is an evaluation metric for measuring how close the system output is to the abstract taking into account the redundancy found in the set of the output.

Table 1: Sentence Extraction Results with Different System Setting

		Short	Long
Event only	coverage	0.325	0.313
	precision	0.491	0.540
Word and Event	coverage	0.323	0.341
	precision	0.523	0.592
Word only	coverage	0.313	0.344
	precision	0.521	0.593

In every cases, value of “Precision” is larger than “Coverage”. It means that our system tends to select redundant sentences.

2.3. Discussion

Our method has better performance in “Long” compared with “Short.” Since our algorithm does not pay attention to the length of a sentence and longer sentence has more chance to have more links, a longer sentence tends to have higher importance. As a result, for the “Short” abstraction, such a longer sentence takes larger room and it becomes difficult to add another sentence. However for the “Long” abstraction, removal of redundant description may make new room to add another sentence. For the future work, it is necessary to have a mechanism that pays attention to the length of a sentence.

3. Hierarchical Structure Analysis about News Articles

3.1. Two Stage Evaluation for Important Sentence Extraction

From this experiment, we found that our algorithm tends to select longer sentences that have similar repetitive events among different news articles. As a result, it underestimate the importance of some sentences that are important only in one article but not discussed in other articles. These sentences are

important when we aim to include several sub-episode in the summary.

For solving this problem, we propose to use two stages evaluation for important sentence extraction; e.g., the first stage is an evaluation based on single article, and the second one is an evaluation based on the relationships among different articles. Followings are algorithm for this two stage evaluation

1. Calculate single article importance value of each sentence

Equation 3 for each document is used for calculate importance value of each sentence. a and b is a document number and a sentence number of the i -th sentence in a whole document respectively. M_a and v_a correspond to transitional matrix and initial importance vector based on the sentence position of the document a respectively.

$$\vec{r}_{ca} = (1 - \alpha) * M_a \vec{r}_{ca} + \alpha * \vec{v}_a \quad (4)$$

The importance value of the i -th sentence v_{si} is a value of b -the sentence value of r_{ca} .

2. Calculate overall importance value of each sentence

There are two approaches for combining single article importance value with overall evaluation proposed in previous research. One approach is just calculating summation of these two importance value. The other one is using single article importance value as a part of initial importance vector in equation 3. By using the former approach, it is better to take the importance of each article into account. On the other hand, the latter approach can propagate the single article importance to other articles by using transition matrix and calculate importance of each sentence as a part of a whole set of articles. So we use the latter approach in this paper.

In our system, we use following initial vector \vec{v}' instead of \vec{v} in previous section. Initial importance value based on

the sentence position for i -th sentence is as follows. In this equation, n is a sentence number in an article and m is a sentence number in a paragraph.

$$v_{pj} = 1/\log(n+1)+1/\log(m+1) \quad (5)$$

We also normalize \vec{v}'_s and \vec{v}'_p that satisfy $\sum_{i=0}^l v'_{si} = 1$ and $\sum_{i=0}^l v'_{pi} = 1$. By using these values i -th element of \vec{v}' is defined as follows.

$$v'_j = \gamma * sj + (1 - \gamma) * pj \quad (6)$$

At last, the overall importance value of each sentences are calculated by using following formula.

$$\vec{r}'_c = (1 - \alpha) * M \vec{r}'_c + \alpha * \vec{v}' \quad (7)$$

3.2. Small Changes

In previous formalization, transition matrix are constructed based on the number of the words shared in a sentence pair. As a result, longer sentence tends to overestimate its importance.

In order to reduce this effect, cosine similarity measure is used for calculating transition possibility. In this formalization, transition possibility based on the word reference is as follows. The i -th sentence is represented as a word vector \vec{s}_i whose value for each word is calculated by TF-IDF($tf/(1+\log(df+1))$) In this formula, tf is a value of term frequency of a corresponding word in a sentence. df is a value of a document frequency of a corresponding word in a newspaper article database).

$$mw_{ij} = \frac{s_i \cdot s_j}{|s_i||s_j|} \quad (8)$$

For the event we use number of events ($events_i, events_j$) and number of similar events shared with a sentence pair $similarEvents_{ij}$ are used for the formalization.

$$me_{ij} = \frac{similarEvents_{ij}}{events_i \cdot events_j} \quad (9)$$

3.3. Experiment

In order to evaluate the effect of the new algorithm, we also apply it to TSC-3 data. Since there are several parameters for tuning, we conduct several experiments with different parameter settings. Since “Coverage” is a value for representing the closeness the system output and the abstract, parameter settings for achieving highest score in “Coverage” are selected. Table 2 shows some results. First two results achieves highest score in “Coverage” in “Short” and “Long” respectively.

We confirm new algorithm improves the value of “Coverage” but degrades the value of “Precision” compared with previous result in Table 1. This is because previous method tends to select sentences that have repetitive event reference and redundant important sentences are selected as a result.

In addition, optimum α value for “Short” is different from that for “Long.” α is a parameter for controlling the effect of initial importance vector. Therefore, higher α value tends to select sub-episode sentences in each article and is good for “Long” case.

Table 3 shows the best performance system in terms of “Coverage” in the TSC-3 [4]. Our system exceeds the performance of the top system.

Table 2: Sentence Extraction Results with Different Parameter Setting

		Short	Long
$\alpha = 0.2, \beta = 0.9$ $\gamma = 0.4$	coverage	<u>0.374</u>	0.401
	precision	0.450	0.543
$\alpha = 0.4, \beta = 0.9$ $\gamma = 0.4$	coverage	0.341	<u>0.419</u>
	precision	<u>0.473</u>	<u>0.578</u>
$\alpha = 0.4, \beta = 0.9$ $\gamma = 0.0$	coverage	0.335	0.408
	precision	0.458	0.565
$\alpha = 0.4, \beta = 0.3$ $\gamma = 0.4$	coverage	0.348	0.373
	precision	0.463	0.533

4. Conclusion

In this research, we propose a method to extract important sentences and to generate

Table 3: Best Performance System in TSC-3[4]

		Short	Long
Best coverage (Short)[9]	coverage	0.372	0.363
	precision	0.591	0.587
Best coverage (Long)[8]	coverage	0.329	0.391
	precision	0.567	0.680

an abstract based on event reference information with hierarchical structure analysis. From the important sentence extraction experiment, we confirm our algorithm produce better results compared with the best performance system in TSC-3. However, further analysis is needed for characterizing our algorithm.

Acknowledgment We would like to thank organizer of TSC-3 for their effort to construct this test data. In this paper, we use news articles data of Mainichi newspaper and Yomiuri newspaper on year 1998 and 1999.

References

- [1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
- [2] Makoto Haraguchi, Masaki Yotsutani, and Masaharu Yoshioka. Towards an organization and access method of story databases. In *7th World Multi-conference on Systematics, Cybernetics and Informatics (SCI2003) Vol. V*, pages 213–216, 2003.
- [3] T. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.
- [4] Tsutomu Hirao, Manabu Okumura, Takahiro Fukusima, and Hidetsugu Nanba. Text summarization challenge

- 3 - text summarization evaluation at ntcir workshop 4 -. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-OV-TSC-HiraoT.pdf>.
- [5] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69, 2002.
- [6] Inderjeet Mani. *Automatic Summarization*. John Benjamin Publishing Company, 2001.
- [7] K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M. Yen Kan, B. Schiffman, and S. Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of Document Understanding Conference 2001*, 2001.
- [8] Tatsunori Mori, Masanori Nozawa, and Yoshiaki Asada. Multi-document summarization using a question-answering engine - yokohama national university at tsc3. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-MoriT.pdf>.
- [9] Koji Eguchi Yohei Seki and Noriko Kando. User-focused multi-document summarization with paragraph clustering and sentence-type filtering. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-SekiY-1.0.pdf>.
- [10] Masaharu Yoshioka and Makoto Haraguchi. Multiple news articles summarization based on event reference information. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-YoshiokaM.pdf>.

Using Metaphors for the Pragmatic Construction of Discourse Domains

Roland Kaschek¹, Heinrich Mayr², Klaus Kienzl³

¹ Department of Information Systems, Massey University
Palmerston North, New Zealand

R.H.Kaschek@massey.ac.nz

² IWAS, University of Klagenfurt

Klagenfurt, Austria

heinrich@ifit.uni-klu.ac.at

³ Höhere Bundeslehr- und Versuchsanstalt Villach

Villach, Austria

klaus.kienzl@intercon.at

Abstract

Conceptual modeling is an important task in systems development. It, however, is more often used as a tool rather than made the subject of research papers. Human thought is considered as essentially metaphorical. The paper therefore proposes to look at the metaphors that might be employed by conceptual modeling. We identify the metaphor INFORMATION IS A NETWORK as a key tool of conceptual modeling. We suppose that focusing on this metaphor has the potential for improving understanding of how conceptual modeling proceeds. The implications of this metaphor and a variant of it are discussed. This discussion includes the metaphor's impact on the theory of abstraction. The paper contributes in two ways to a theory of conceptual modeling. It firstly provides a unifying view of the pragmatic construction of discourse domains that is based on using the metaphor INFORMATION IS A NETWORK. It secondly proposes an approach to abstraction theory that fits this unifying view on conceptual modeling.

1. Introduction

In a wide sense of the term conceptual modeling (CM) can be used to mean conceptualizing, i.e. constructing a discourse domain in terms of concepts that belong to the domain. For example CM can be used for analyzing or describing computer applications or to prescribe their characteristics. In other realms of informatics such as reference modeling, enterprise modeling or similar CM is contributing essentially to the achievements. CM thus is understood to be of major importance in informatics. While there are major international conferences dedicated to CM, such as the Entity-Relationship conference, approaches towards a theory of CM seem to be rare, see

[29].

A theory of CM could aid in teaching CM, practically carry out CM, and impact the research on CM. We presuppose that it would be useful to cover automated activities by our theory of CM. Therefore we aim at a theory that avoids mentalism and is oriented at linguistics and semiotics. Obviously we have to take a look at papers from artificial intelligence that deal with related subjects. Abstraction must be expected to be of major importance for our study since informatics or computer science occasionally is considered as the discipline of abstraction.

CM has a precursor in philosophy, i.e. Carnap's logical construction of the world. We are going to briefly discuss Carnap's ap-

proach and identify the metaphor INFORMATION IS A NETWORK as the key methodological ingredient of his work. It turns out that this metaphor is omnipresent in conceptual modeling and that it helps developing a semantic abstraction theory for CM.

Software development is a process of knowledge creation and representation. One of the characteristics of this process is a successive formalization of the language in which that knowledge is obtained and represented. CM is that step in software processes that can be characterized by a two-fold formalization. Models can be given a formal semantics and the next step, i.e. design, makes a formal prescription of a computer application to build. Obviously both of these formalizations are important regarding the success of software development. In many cases requirements holders are not able or willing to understand the formal statements referred to right now. It is, however, important that they can learn to effectively and efficiently talk about the conceptual models. The paper is dedicated to simplifying that task. Our respective idea is that using metaphors can bridge the gap between the formal area and the application area. Note that we consider the need to speak metaphorically as symmetric in terms of stakeholders and developers, i.e. both of these talk about things they don't understand in terms of what they understand. The point is that these actors understand different things and that none of these is superior to the other.

We are modelers. Thus we believe in a certain pluralism of views that is similar to the one in [27]. We are, however, not relativists. We presuppose that among competing conceptualizations there can be defined and justified a partial order of preference. The respective justification requires certainly a detailed and specific consideration of the conceptualization's parameters in the sense of section 3.

One can consider our approach in this paper as an epistemological one. Then one would be interested in identifying and discussing its ontological counterpart. Such counter-

part could be based on the system concept. The decomposition of systems into subsystems that are related to each other by structural or communication relationships would then in this ontological approach resemble the use of the metaphor INFORMATION IS A NETWORK. In this paper we are not going to compare and assess both approaches. Rather we just want to develop our pragmatic approach to constructing domains of discourse.

Paper Outline. In the next section we briefly discuss Carnap's approach to a logical construction of the world. We then, in section 3, discuss the foundations of modeling in so far as we believe to understand them. In section 4 we briefly discuss the expressiveness of the model concept that we have chosen for this paper. We then discuss metaphor and change of representation in sections 5 and 6 respectively. After that we discuss how domains of discourse are constructed by CM. We conclude our paper with section 8 regarding abstraction that is followed by an outlook and our references.

2. Carnap's Logical Construction of the World

In this paper we focus on the theory of model creation. We ground our approach on Carnap's attempt to logically construct the world, see [4]. Carnap attempted to logically construct the world as a network, i.e. a graph. His idea was that (1) the nodes of the graph should represent the things in the world and that (2) the edges of the graph should represent the relationships between the things of the world. Carnap assumed that (3) for each two things T_1, T_2 in the world it would be possible to find relationships to other things in the world such that T_1 and T_2 could be distinguished from each other in terms of their pattern of relationship participation.

Carnap's objective was to formally characterize all entities in the world in one given language, i.e. the language of graph theory. Achieving that objective would turn all discourse regarding entities within the world into a mathematical discourse. The

aspect of that appealing to Carnap might have been that finally one of the competing views would have been singled out as the true one by means by mathematical reasoning. We identify two major challenges regarding that objective. Firstly, that which is counted as an entity in the world as well as that which is counted as a relationship between two things in the world is subjective, i.e. depends on the one who admits a thing or a relationship. Secondly, it is not likely that for each two entities e, e' in the world there is exactly one way of identifying sets of relationships r, r' to other entities in the world respectively that would allow to distinguish between e and e' . The consequence of each of the two objections is that each list of formal characterizations of entities in the world is incomplete or contains subjective cognitive aspects (that aim at justifying selections or decisions that were made) and thus deciding about assertions regarding entities in the world by means of only mathematical reasoning becomes impossible.

Carnap didn't achieve his goal. His method, however, did not become obsolete therefore. Rather it is quite successful. Using the terminology of metaphors that we are going to introduce below one can say that Carnap has used the metaphor INFORMATION IS A NETWORK. His method of constructing a discourse domain consists in creating information that specifies admissible and filters out inadmissible phenomena and structure. That metaphor enables humans to handle the abstract concept of 'information'. It suggests firstly using information as a filter applied to the phenomena of a discourse. Secondly it suggests representing that information as a graph the nodes of which represent information chunks and the edges of which represent relationships between information chunks. Carnap's method nowadays is very frequently used in CM. Independently of the thing discussed (whether for example structure, behavior, causality, modality or quality of systems is discussed) and the modeling language used (whether for example a graphical, textual, or formal modeling language is used) Carnap's method is applied.

CM follows Carnap in representing information as a network. It, however, deviates essentially from Carnap in several respects, i.e. CM

- uses a material language rather than a formal one, i.e. it provides intensional definitions for the information chunks in the network and the relationships between these. The respective definitions are typically gathered together in a thesaurus or data dictionary, see, for example [25];
- focuses on a discourse domain rather than the whole world;
- introduces typing of nodes and edges of the network used for representing information. This typing of nodes results in information chunks being considered as (1) entity type, relationship type etc. in the entity relationship model (see [5], [33]), or as (2) classes, associations etc. in a class diagram (see [26]), or as (3) state, state transition etc. in a state chart (see [26]), or similar. Aspects of this typing can be understood in terms of a further metaphor, i.e. INFORMATION IS A CONTAINER that we are going to discuss below;
- relaxes the concept of identifiability of things in terms of their network representation. In fact, the row-id of relational databases as well as the object-identity of object-oriented modeling approaches enables modeling discourse domains in which different phenomena exist that cannot be distinguished from each other by limiting to only their characteristics in this discourse domain.¹

In summing up we say that CM in our view aims at a pragmatic approach to constructing discourse domains rather than at a formal one.

¹The latter, however, not necessarily is a problem, as one always can define handles that enable distinguishing the mentioned entities from each other.

Note that the quality of information to function as a filter does not depend on the kind of information chunks and relationships thereof that are introduced with the metaphor INFORMATION IS A NETWORK. What counts is that the metaphor enables making distinctions that make a difference (see [2] for a definition of information that is based on this idea) with respect to recognizing and conceptualizing the phenomena in a discourse domain.

3. Modeling

CM achieves an understanding of a discourse domain such that (1) those phenomena can be distinguished from each other that are important with respect to a given purpose; (2) a state of affairs is introduced that aids in meeting the objective of modeling the discourse domain; and (3) the flux of state of affairs can be represented by a sequence of model changes.

We are going to discuss abstraction below. We use a preliminary understanding of abstraction as that what one achieves by abstracting. The everyday-language meaning of “to abstract” according to the Oxford English Dictionary Online is “(t)o withdraw, deduct, remove, or take away (something)” from something else. Before becoming more technical we briefly sketch the context of what we focus at. Contrary to practice in artificial intelligence (see, for example [36, p. 1296-8]) we do not consider it essential for an abstraction to preserve desirable properties or similar. We rather consider quality of abstraction as the thing to deal with when it comes to assessing the utility of representations. Regarding how an actor conceptualizes a discourse domain that essentially uses input from outside the actor’s own cognitive system we distinguish from each other three stages,

- **sense irritation**, i. e. signals trigger irritation patterns being present in a number of the actor’s input interfaces. Incompleteness and imperfection of the actor’s senses lead to a first kind of ab-

straction regarding reasoning about the physical world;

- **phenomenon constitution**, i.e. a number of phenomena is created employing a semantic model that was chosen in an ad-hoc manner. Note that the actor’s attention as well as the coarseness of the actor’s concepts suggests another kind of abstraction in reasoning about the physical world. Note also that we consider the semantic model chosen ad-hoc for conceptualizing sense irritation as in part depending on cultural and individual aspects such as education and available resources;
- **modeling**, i.e. the phenomenon constitution is reconsidered employing a semantic model that is purposefully chosen and that leads to presupposing a set of adjusted phenomena that are tailored towards the utility for solving a problem at hand. In contrast to the first two levels on the level of modeling we deal with relatively conscious choices of abstractions. Note that in coincidence with [12] we consider a model (i.e. the particular phenomenon constitution worked out with the chosen semantic model) as an ontology if it is considered as the (only) true conceptualization, i.e. the necessary phenomena constitution with respect to the discourse domain.

Our approach incorporates only a rather limited version of realism. We find ample support for our staged approach to conceptualization in [27]. We mention in particular I. Kant (p. 259), M. Schlick (p. 239), E. Cassirer and W. v. Helmholtz (p. 235), K. Bühler (p. 231), L. Boltzmann and H. Hertz (p. 157), and J. S. Mill (pp. 115 - 117).

In this context we consider a **semantic model** Σ as a pair (MC, RC) of **modeling concept** MC and **representation concept** RC , for more detail on the latter concepts see [19]. To summarize briefly, a modeling concept is a set of modeling notions

and abstraction concepts together with intensional definitions, usage rules, including concept compatibility constraints, and application hints. Employing these definitions, rules and hints is supposed to result in a **model**, i.e. a set of domain concepts and abstractions which are instantiations of modeling notions and abstraction concepts respectively. A representation concept is a set of conventions for representing models. Obviously, communication about models would be close to impossible or useless without the effective use of representation concepts. With respect to what follows below we focus on the stage of modeling, i.e. we focus on semiotic entities. Rather than the term semantic model the term **modeling language** is used as well.

With respect to modeling we see at least two different approaches. The modeling actor may be human or not. In the latter case the artificial intelligence community usually focuses on computerized systems. We first discuss what we consider approaches to semantic modeling (i.e. modeling processes as conducted by humans) and deal with change of representation (i.e. automated modeling processes) later.

We use the model theory of Stachowiak [30, 31, 32] as a starting point. According to his theory models are typically used as proxy objects. The **model relationship**, i.e. the predicate $R(O, M, A)$ means that actor A uses model M as a proxy for model original O for specified purposes only, period of time and context, and subject to using specified tools and techniques. Stachowiak admits models and originals that are not semiotic for a given actor. His theory of how model and original are related to each other, however, is based on semiotic objects that a modeler uses to represent the non semiotic ones.

Assuming that a modeler A has established a model relationship $R(O, M, A)$ Stachowiak assumes that O^* and M^* are semiotic representations of O and M respectively, i.e. they are sets of predicates with which A conceptualizes O and M respectively. We assume

that in case O or M is semiotic then $O^* = O$ and $M^* = M$ respectively. With these assumptions one can explicate the model relationship more precisely and explain the following properties that were introduced by Stachowiak.

- **mapping property**, i.e. A uses M^* as a proxy for O^* and has defined an isomorphism, i.e. a pair (τ, γ) of partial mappings $\tau : O^* \rightarrow M^*$ and $\gamma : M^* \rightarrow O^*$. We call these mappings **truncation** and **grounding** respectively.
- **truncation property**, i.e. $O^* \setminus \text{dom}(\tau) \neq \emptyset$;
- **pragmatic property**, i.e. using M^* as a proxy for O^* is only legitimate for specified actors, purposes of these actors, admissible tools and techniques of investigation, time of use, etc.

Stachowiak requests $\tau^{-1} \subseteq \gamma$ as a compatibility constraint between these two partial mappings. This implies that τ is injective where it is defined. The constraint implies $\tau \circ \gamma(p) = p$, for all predicates $p \in \text{im}(\tau)$. Therefore $\tau \circ \gamma \circ \tau = \tau$. We are going to use this equation as the condition required that relates truncations and groundings.

We point out that additionally to the truncation property a model in general also has an **extension property**, i.e. $M^* \setminus \text{im}(\tau) \neq \emptyset$. Stachowiak called the predicates in $M^* \setminus \text{im}(\tau)$ **abundant** and the ones in $O^* \setminus \text{dom}(\tau)$ the **preterited** ones. We call the former **surplus** and the latter **ignored** ones. The surplus predicates contribute essentially to the use that can be made from the model. For a respective example see [12]. One is tempted to define abstraction as that kind of model relationship $R(O, M, A)$ in which no surplus predicates occur. We choose a slightly different approach, as that temptation, however, might lead to difficulties with respect to the syntax of the chosen modeling language.

For a simplifying description of how modeling processes can be understood according

to Stachowiak (see [30, pp. 317]²) we assume that an actor A wants to solve a problem P_O that is encoded in terms of an object O . For some reason or another A cannot solve the problem and assumes a reformulation of P_O might help. A therefore replaces O and M by semiotic representations O^* and M^* of these respectively and establishes a model relationship $R(O^*, M^*, A)$. Then A obtains a reformulation P_{O^*} of P_O in terms of O^* . A observes that (s)he likely can turn a solution of P_{O^*} into a candidate solution of P_O . A furthermore uses the truncation τ to obtain a problem P_{M^*} in terms of M^* as an encoding of P_{O^*} . If A can work out a solution S_{M^*} of P_{M^*} then A uses the grounding γ for obtaining a hypothesis S_{O^*} as an encoding of S_{M^*} in terms of O^* . If A has chosen a successful approach the hypothesis S_{O^*} can be turned into a solution proposal S_O regarding P_O . If things work out well S_O is an admissible solution of P_O . If the latter is not the case then a number of feedback steps can be incorporated easily into the description given right now. Stachowiak notes that the process of solution generation may lead to increased knowledge that in turn may result in the models M or M^* , or the originals O or O^* being changed.

Semiotic models are similar to signs, metaphors and other concepts such as simile and analogy. We are going to discuss the first two of these below. Prior to that we mention that our epistemic pluralism suggests that epistemological theories that do not focus on modeling might have their strengths, as they represent different angles on acquiring knowledge about the world. Note furthermore that of course the terminology used by the various authors is not coordinated. In [22, ch. 7] for example it is mentioned that the word model sometimes is used as a synonym of analogy. In that source, moreover, analogy is said to be fundamental for how humans understand the world. That valuation of analogy is comparable to the one of Ch. Fourier (see [27, p.

²Note that we use a different terminology and notation.

67]) who considered analogy as a key cognitive tool for understanding that, which was not understood before.

With the term **sign relationship** we refer to the predicate $S(T, R, A)$, which means that actor A uses the **token** T to refer to the sign **referent** R . We do not elaborate more on semiotics or linguistics and refer the interested audience to [20]. We restrict the class of things from which a sign referent may be chosen. We assume that the referent is a **cultural unit**, i.e. something to which one in a given culture can meaningfully refer to. We have adopted this concept from [6]. Using the now already obvious metaphor CULTURAL UNIT IS A NETWORK can be expected to operationalize the concept of cultural unit. Putting things that way is a consequence of assuming that the sign referent can only serve as a referent based on a convention of a group of actors. The obvious structural similarity between the model relationship and sign relationship can be made explicit by assigning to the model the role of the token and to the original the role of the referent. The grounding γ can then be used for conducting the reference to the sign referent. A model can be used for transferring information, that was requested in terms of the original, from model to original. In contrast to this with the sign token only the existence of the referent is pointed out. We are going to explore this issue in more depth below.

For us semantic information is the quality of a string to function as a filter and to help eliminate states of affairs out of a set of eligible such states. Compare, for example [9, p. 435]. The amount of semantic information contained in a string s is proportional to the cardinality of the set of states of affairs eliminated by that constraint.

4. Expressiveness of Stachowiak's Model Concept

Our view of models as a reference concept similar to signs and related to metaphors suggests that models should be a tool of many scientific disciplines. At least in Logic

and Mathematics concepts referred to as models are actually important. For simplifying our terminology we call our concept of model a Philosophical one. Additionally we consider a Linguistic concept of Model. We are going to show that our Philosophical model concept in fact can be understood as a generalization of the Linguistic model concepts.

4.1. Structures

Modeling, as considered above, involves creating semiotic models, i.e. sets of predicates. Since predicates reflect someone accrediting a notion to an entity we consider semiotic models as verdictive utterances, i.e. an utterance by means of which an actor accredits to a number of subjects a predicate notion. For more detail regarding verdictive utterances see, e.g. [1]. In some cases of verdictive utterances the utterance creates a social fact. In these cases the utterance in general will only then be successful if the utterer is authorized, performs the utterance in the right context and in the required way. These social parameters of verdictive utterances are not in the focus of our paper. However, they cannot be ignored completely because in organizations where CM is applied a consensus needs to be fixed and a decision needs to be made that is binding in or for that organization. Here we abstract from these parameters. Considering CM in the narrower sense leads to the concept of judgment coming into the focus. Judgments are forms of verdictive utterances that are elementary with respect to the information conveyed. Judgments can be represented as predicates $U(S, P, C, A)$. Such a predicate means that the actor A accredits to all instances of the subject notion S the predicate notion P in a way specified by the copula notion C . For an elaborated theory of judgment see, e.g. [24].

It has been shown in [17] that algebraic as well as relational structures can be specified as sets of judgments. It thus is not a severe limitation to (with respect to modeling purposes) restrict oneself to verdictive

utterances. At least the more well-known semantic models of informatics such as entity-relationship models, class diagrams, state charts, Petri nets, or similar can be shown to be covered by that language. Recall that a structure R is called retract of a structure C if there are morphisms $\iota : R \rightarrow C$, and $\pi : C \rightarrow R$, such that $\pi \circ \iota = 1_R$. In such situation is C called a co-retract of R . Since the identity morphism 1_R is the identity on R the morphism ι is an injective mapping and so the pair (ι, π) can be considered as an isomorphism. Since $\pi \circ \iota = 1_R$ implies $\iota \circ \pi \circ \iota = \iota$ one can (in the sense of Stachowiak) consider a structure as a model of each of its retracts. A characterization of co-retracts of algebraic structures is given in [11].

4.2. Grammars

We are going to show that a formal grammar can be considered as a model of each of the words that belong to the language defined by the grammar. Recall for this (see for example the chapter on formal languages in [34]) that a formal grammar is defined as a quadruple $\Gamma = (S, N, T, R)$, where N, T are finite disjoint sets that are called **non-terminals** and **terminals** respectively. Further $S \in N$ is called the **start symbol** and $R = \{(\sigma, \tau) \mid \sigma, \tau \in (N \cup T)^*, \sigma = \sigma_1 \dots \sigma_m, \{\sigma_1, \dots, \sigma_m\} \cap N \neq \emptyset\}$ a set of **production rules**. The language $\Lambda(\Gamma)$ defined by Γ is the intersection of all sets Σ with the properties

- $S \in \Sigma$, and
- $s \in \Sigma, (\sigma, \tau) \in R, \alpha, \omega \in (N \cup T)^*, s = \alpha\sigma\omega$ imply that $\alpha\tau\omega \in \Sigma$.

If there is a rule $r = (\sigma, \tau) \in R$ in a grammar $\Gamma = (S, N, T, R)$ and $\alpha\sigma\omega, \alpha\tau\omega \in \Lambda(\Gamma)$ then one says that $\alpha\tau\omega$ is directly derived from $\alpha\sigma\omega$. In that case we write $r(\alpha\sigma\omega \rightarrow \alpha\tau\omega)$. Direct derivation is a relation on $\Lambda(\Gamma)$ its transitive closure is called derivation δ and for $\xi, \psi \in \Lambda(\Gamma)$ one says that ψ is derived from ξ if $(\xi, \psi) \in \delta$.

For a grammar $\Gamma = (S, N, T, R)$ let $\Gamma' = \{(1, S)\} \cup \{(2, n) \mid n \in N\} \cup \{(3, t) \mid$

$t \in T\} \cup \{(4, r) \mid r \in R\}$. Obviously $\mu : \{\Gamma \mid \Gamma \text{ is a grammar}\} \rightarrow \{\Gamma' \mid \Gamma \text{ is a grammar}\}, \Gamma \mapsto \Gamma'$, is a bijection. For each $\lambda \in \Lambda(\Gamma)$ choose a shortest derivation $s_1 s_2 \dots s_m$. Define $\Pi(\lambda) = \{(i, s_{i-1}, s_i, r) \mid 1 \leq i \leq m, s_0 = S, s_m = \lambda, r(s_{i-1} \rightarrow s_i), \text{ for all } i\}$. We define now $\tau : \Pi(\lambda) \rightarrow \Gamma'$, such that $(i, s_{i-1}, s_i, r) \mapsto (4, r)$. We furthermore define $\gamma : \Gamma' \rightarrow \Pi(\lambda)$, such that $(j, x) \mapsto (i, s_{i-1}, s_i, x)$, if $j = 4$ and i is the smallest index with $(i, s_{i-1}, s_i, x) \in \Pi(\lambda)$, and $(1, s_0, s_1, r)$ otherwise. Then it follows that $\tau\gamma\tau(i, s_{i-1}, s_i, r) = \tau\gamma(4, r) = \tau(k, s_{k-1}, s_k, r) = (4, r) = \tau(i, s_{i-1}, s_i, r)$. Therefore $\tau \circ \gamma \circ \tau = \tau$, which had to be shown.

5. Metaphors

One of the most basic phenomena of language use is the capability of speakers to use different words for the same purpose or thing.³ Metaphor is a particular way of achieving that effect. According to [18] a metaphor is a partial mapping out of a source domain in a target domain. These domains, if the metaphor is supposed to work, must be cultural units. Lakoff in [18] has used the metaphors LOVE IS A JOURNEY, and LIFE IS A JOURNEY to illustrate the concept of metaphor. The notational convention borrowed from Lakoff puts the metaphor name in small capitals and lists the source domain before the infix IS A and the target domain after that infix. Metaphors are used for transferring information from the target domain to the source domain. If one ever has understood a metaphor such as LOVE IS A JOURNEY then it offers no difficulty to understand a metaphorical expression such as *our relationship is at a dead end* even if one has never encountered it before. It is a striking advantage of Lakoff's theory of metaphor to enable understanding that fact that, by the way, couldn't be explained satisfactorily by the elder theory of metaphor (see for example Searle's paper in [23]). In the tradi-

³We thank Tilman Reuther for mentioning this observation to us.

tional theory of metaphor the so-called 'literal meaning' of a phrase was supposed to be the standard mode of language use. We do not go into respective detail here. Rather we refer to Rumelhart's paper in [23] for a discussion of some of the problems regarding 'literal meaning'.

Examples of further metaphors that are used with respect to software systems are not only the interface metaphors such as THE COMPUTER IS A DESKTOP or THE COMPUTER IS A COUPLE OF (DIRECT MANIPULATION) OBJECTS. Rather metaphors such as SOFTWARE IS A LIVING THING or SOFTWARE IS A BUILDING are used as well. The latter ones respectively enable humans to talk about software systems in terms of these systems aging or getting infected by a virus or similar and in terms of their architecture. It occurs such that the metaphor SOFTWARE IS AN ACTIVE THING is a precursor of the metaphor SOFTWARE IS A LIVING THING. It is presupposed in talking about programs in terms of the metaphorical expressions as "the program calls a subprogram", "the program is executing", or similar. Obviously, presupposing current computer technology, programs are not active. Rather the computer uses them for controlling its own behavior. In [13] the metaphorical expressions *assistant*, *agent*, *master* and *peer* are used for discussing roles that can be played by computers in human computer interaction. These metaphorical expressions fit the metaphor THE COMPUTER IS AN INTERLOCUTOR. The role of metaphors with respect to the evolution of knowledge that is relevant for computer use was discussed in [15].

The Oxford English Dictionary Online defines dynamics as "(t)hat branch of any science in which force or forces are considered." The term 'dynamic modeling' is quite frequently used in systems development when one wants to refer to understanding the changes of states of certain items such as the objects in an object model. Obviously the use of that term is a metaphorical one, as the events that occur and im-

pact the state of such objects neither are forces nor exert forces on the considered objects. Rather the computer is supposed to use the information that the event has occurred for adapting the objects' state adequately. One could use the metaphor INFORMATION IS A FORCE as the base of using the term 'dynamic modeling'. Comparable metaphors are in use in software process modeling. Finally, also talking about static modeling when referring to creating entity-relationship diagrams or class diagrams is a metaphoric language use, as static also is a term used in studying forces and their consequences.

The obvious similarity between metaphors and model relationships can be made explicit by assigning to the original and the model the role of the source domain and the target domain respectively. The actual mapping is then conducted by the truncation τ . The difference between models and metaphors is that metaphors compared to models achieve a significantly reduced information transfer from model to original because the only mappings that can be used as a grounding are subsets of τ^{-1} , i.e. the relation inverse of the truncation. Metaphors are important with respect to CM as the modern consensus of linguistics is that human thought is essentially metaphorical. For more detail see, e.g. [21].

6. Change of representation

According to [8, p. 1197] **change of representation** in artificial intelligence means that an actor changes a problem statement. The actor is usually presupposed to be a computerized one. The outline of modeling processes as described above obviously covers reformulation, since reformulation not necessarily has to explicitly refer to the non-semiotic entities in the role of original and model respectively that were mentioned above. [8, p. 1197] gives a nice example showing that in fact mentioning such entities not necessarily is required. In [36, p. 1297] it is then defined that "an abstraction is a change in representation, in a same for-

malism, that hides details and preserves desirable properties." Obviously the terms 'detail', and 'desirable property' need further clarification. It is our view that these terms only can be explained with respect to the purpose of the abstraction performed. Together with these terms also the term 'simplicity' is used in the literature for explaining aspects of abstraction. The latter term, however, as well is notoriously hard to define.

[36, p. 1300] states that it is important to choose abstractions carefully and that "... the utility of abstraction is related to the utility of the representation change that is associated with it." The drawbacks that might occur with respect to the utility of an abstraction are defined in terms of computational cost and problem complexity. It occurs as a reasonable assumption that change of representation is a particular kind of modeling approach that is designed for use by computerized systems. With respect to human modeling processes various reasons can be mentioned that justify modeling (such as ethical, legal, or practical reasons). Regarding change of representation, however, the dominating concern is the computational cost of problem solving. Abstraction is aimed at because it often, but by no means always (see [36, 4.(d)]), achieves its potential for cost reduction. With respect to the terminology used in CM one would want to refer to model quality to capture abstractness or simplicity of representations.

In reviewing abstraction theory, as used in artificial intelligence, [36, pp. 1298-9] uses the metaphors⁴

1. ABSTRACTION IS A MAPPING BETWEEN PREDICATE SETS
2. ABSTRACTION IS A MAPPING BETWEEN FORMAL SYSTEMS,
3. ABSTRACTION IS A SEMANTIC MAPPING OF INTERPRETATION MODELS, and

⁴Note that the source does not explicitly uses the terminology of metaphors.

4. ABSTRACTION IS A SEMANTIC MAPPING OF FORMULAE.

Using the first metaphor seems to suggest the problem that was already mentioned above when discussing abstraction in the context of Stachowiak's model theory. Therefore we do not accept that metaphor. As [36] mentions, the second metaphor seems not to be suited for aiding actors in finding new abstractions. It also seems not to fit to our general view of modeling processes, as it does not refer to the phenomena in a discourse domain. We thus reject that metaphor as well. The last two approaches comply with our view of modeling processes as involving phenomena that constitute an object level with respect to which on a meta level semantic information can be specified for restricting the state of affairs at the object level. We are going to use aspects of both of these metaphors for our proposal of abstraction in CM.

[36] proposes a new theory of abstraction. We reject that theory with respect to CM for reasons we explain shortly. That theory employs the metaphor ABSTRACTION IS A MAPPING BETWEEN PRESENTATION FRAMEWORKS. A presentation framework is a 4-tuple $D = (P, S, L, T)$ where P is a set of phenomena that are typed as either object, attribute, function, or relationship; S is a storage structure for storing the phenomena including their typing; L is a language for talking about the phenomena; and T is a theory on which reasoning about the world is based. A presentation framework is supposed to have a number of parameters. Two of these are of particular importance for our purposes. The first one is a world W that is represented by the framework. The second one is a process \mathcal{P} that generates the set P out of W . For the process \mathcal{P} [36] does not provide a specification. Only a reference to [28] is given. In that source, however, there only is given an example of a digital camera that is capable of providing bit patterns that function as phenomena. That is an important point as [36, p. 1303] says that the concept of 'world' by no means is re-

stricted to a physical one. For non-physical worlds, however, no hints at all are given as to how the phenomena in P could be determined. For another reason, that we are going to explain now, this point is very important as well. These two points make us reject that abstraction theory for CM.

Given a world W an abstraction is then a mapping \mathcal{A} from a presentation framework $D_g(W)$ onto a presentation framework $D_a(W)$ such that (1) $D_g(W) = (P_g, S_g, L_g, T_g)$, $D_a(W) = (P_a, S_a, L_a, T_a)$, (2) $\mathcal{A}(P_g) = P_a$, and (3) P_a is simpler than P_g . The predicate simpler is defined such that for the Kolmogorov complexity K the following condition holds: $K(\Gamma_g) \leq K(\Gamma_a)$.⁵ In this condition the sets Γ_g, Γ_a are so-called configurations of the perception processes $\mathcal{P}_g, \mathcal{P}_a$ respectively that generate P_g, P_a respectively. Not only is the Kolmogorov complexity known to be a non-computable function and therefore practically computing the values $K(\Gamma_g)$ and $K(\Gamma_a)$ might be not feasible. The configurations furthermore are assumed to be sets of bit strings and the Kolmogorov complexity of a configuration is the maximum of the Kolmogorov complexities of the bit strings in that configuration (see [28, p. 766]). For infinite configurations a maximum not necessarily exists and thus the complexity comparison might be meaningless. Finally, since there were no process quality aspects specified for the processes \mathcal{P}_g and \mathcal{P}_a it cannot be guaranteed that repeated or concurrent process execution is invariant against Kolmogorov complexity. We consequently currently see no way of using the abstraction theory under discussion for CM.

7. Constructing Discourse Domains

The main purpose of CM is conceptualizing in a suitable way a discourse domain at hand. The phenomena in the discourse domain can

⁵Recall that for a string S the complexity $K(S)$ is defined as the length of a shortest program that puts out S . A very brief introduction to Kolmogorov complexity can be found in The Wikipedia.

be distinguished from each other by a combination of structural, extensional, and intensional parts of the conceptualization. CM may be concerned with concepts (see, e.g. [14] for a definition and discussion) in two ways. First of all, if CM is used by an actor A to conceptualize a discourse domain D^* then A creates a model M^* of D^* , i.e. A creates a model relationship $R(D^*, M^*, A)$. For this paper D^*, M^* are a semiotic entity and thus involves concepts. Finally, the discourse domain D that has D^* as a semiotic representation in $R(D^*, M^*, A)$ may already be a conceptualization. That, for example, is the case in requirements elicitation and analysis.

A **discourse domain** is something that one can sensibly talk about, i.e. something that can be a matter of discourse. We assume that the minimum assumption for that to be possible is that one can identify in the discourse domain a collection of phenomena. In our view when conceptualizing a discourse domain a modeler aims at providing semantic information that rules out all non-sensible phenomena. Our thesis is that in this respect conceptual modelers use the metaphor INFORMATION IS A NETWORK. This metaphor suggests that the required information can be represented as a number of information chunks (nodes of the network) and that several of these chunks are related to each other (edges in the network). Applying that metaphor enables modelers to qualitatively classify discourse domain-phenomena as thing and thing relationship respectively. If the modeler wants a more sophisticated conceptualization then the modeler may chose a target domain for the mentioned metaphor that has node labels or edge labels. These labels can be used to define types of discourse domain-phenomena. The semantic information initially required most easily can be represented graphically such that nodes labeled equally will be represented by a given shape (such as a rectangle or a diamond etc.) and that the edges labeled equally will be represented as a particular edge type. Both edges and nodes will be attached an identifier of the phenomena

mapped onto node or edge.

A further metaphor is used in constructions of discourse domains, i.e. INFORMATION IS A CONTAINER. In this metaphor information is considered as something that can contain information in the sense that the latter can be put in or out. For example, when modeling with state charts (see, e.g. [25] or [26]) one uses two forms of nesting, i.e. of including state charts into a given state chart. The latter is then considered as being at a higher level of abstraction. These two ways represent generalization and aggregation of states respectively. Similarly, in class diagrams one uses boxes to represent classes. Inside of these boxes the attributes as well as the methods of objects of the respective class are listed. In class diagrams the metaphor INFORMATION IS A CONTAINER is used for representing data encapsulation. Please note that typically in ER diagrams that metaphor is not employed since data encapsulation is not a modeling notion of the ER model. ER modeling as well as Petri Net modeling use the metaphor INFORMATION IS A CONTAINER in a further way. An elementary model part such as an entity, a relationship, or a place, or a transition is defined as including a whole ER diagram or Petri Net respectively. The latter form of applying the metaphor INFORMATION IS A CONTAINER is called refinement.

In refinement the elementary model part is supposed to be at a higher level of abstraction than is the refining model. The purpose of refinement is exactly the introduction of these levels of abstraction. These levels make complex models more usable because for some use of them it is sufficient to focus on a given level of abstraction and ignore the rest of the model. Thus refinement has the potential of significantly reducing the complexity an actor has to deal with at once for solving a problem at hand. Obviously, applying the metaphor INFORMATION IS A CONTAINER can be made redundant by explicitly typing nodes and edges in a graph that represents information according to the metaphor INFORMATION IS

A NETWORK.

Providing evidence for a theory like the one we have outlined in this paper is not a simple task. Currently we have no clues of how to empirically confirm our theory. We can, however, mention three observations that could be considered as such confirmation. Firstly, in object-relational database management systems the subtype as well as the sub-table relationship is denoted as "under" (see [35]). This seems surprising since established different terminology is available. The chosen terminology seems to presuppose the metaphor INFORMATION IS A SPATIAL NETWORK. Secondly, there is a long tradition of work regarding the question whether relationship types in the ER-model can have attributes or not. There is quite some work of Ron Weber and collaborators devoted to that question. Weber has argued in two ways, i.e. empirically and theoretically. In his empirical respective work he has tried to show that modelers find it more difficult to understand relationships in the ER-model that have attributes rather than relationships without such attributes. In his theoretical respective work Weber has argued that ontological reasoning suggests that relationships must not have attributes. It is however a severe weakness in his latter reasoning that he only uses Bunge's ontology and does not provide a rationale for choosing an ontology. While from the view of an ontology's purpose that might not be particularly worrying it is however known that a number of ontologies exist and that for example Chisholm's ontology has been applied to information systems related questions. The obvious weakness of not providing a justification of the choice made for the ontology seems to suggest a predisposition in place against attributes of relationships. The conventional use of the metaphor INFORMATION IS A NETWORK has the potential to explain the existence of such a predisposition. According to that metaphor information is only represented by two-dimensional items (such as rectangle or circle which represent entity type and value type respectively) rather than one-

dimensional ones. One-dimensional items, i.e. lines cannot have attributes because that would involve a line connecting a line with a two-dimensional item, i.e. an information chunk.

Thirdly, our theory complies with a lot of modeling practice that we are aware of. We mention in particular the use of the metaphor INFORMATION IS A SPATIAL NETWORK that we are going to discuss in detail in section 8.1. We mention here only the interesting point that this metaphor enables making pragmatic distinctions that are usually not supported by formal model semantics.

8. Abstraction

Abstraction comes in at least two forms, i.e. intra model abstraction (as levels of abstraction within a model) and inter model abstractions (as different views on a given discourse domain). In contrast to that we suggest understanding the intra model abstraction as consequence of replacing the metaphor INFORMATION IS A NETWORK by the metaphor INFORMATION IS A SPATIAL NETWORK. Intra model abstraction is mainly used for improving model qualities such as readability, memorizability, or maintainability. This form of abstraction mainly is concerned with the detail of specification. We suggest understanding inter model abstraction as omitting information represented by one model in comparison to another model. For providing a framework that tells us what kind of information can be omitted from a model we refer to the metaphor INFORMATION IS A SPATIAL NETWORK. We thus suggest that the information that can be omitted concerns ruling out a discourse domain-phenomenon from the model, or typing it inadequately, or by relating it to undesirable phenomena in an undesirable way.

It appears to exist a separation of interest in abstraction with respect to data engineering and artificial intelligence. In [36, p. 1294-6] five examples of inter model abstraction are discussed while intra model abstraction is ignored. In [3, section 2.1] only the tra-

ditional examples of intra model abstraction are considered, i.e. classification, generalization and aggregation. In [7] a further intra model abstraction is mentioned, i.e. de-contextualization. [16] is among the papers that have discussed further concepts of intra model abstraction, i.e. cooperations. It is an obvious idea to employ intra model abstraction for achieving inter model abstraction. From a view of systematic model development one feels tempted to challenge that inter model abstraction is not discussed in data engineering. It seems that data engineering aims at finding the right abstractions only within a model while artificial intelligence aims at finding the right level of abstraction of the models, i.e. one tries to find minimum complexity models of a given discourse domain. From a holistic point of view both of the disciplines seem to have chosen an incomplete approach to abstraction.

8.1. Intra model abstraction

Diagrams as the ones obtained from employing the metaphor INFORMATION IS A NETWORK may be confusing as no obvious principle to group nodes and edges can be applied. For introducing hierarchy into the diagram the basic metaphor INFORMATION IS A NETWORK is complemented by the metaphors INFORMATION IS A SPATIAL NETWORK and INFORMATION IS A CONTAINER. This allows in the diagram to make use of the well-known pairs of spatial concepts such as (*up, down*), (*left, right*), (*front, rear*), and (*in, out*). These spatial concepts can be applied to the semantic information as represented by the network because of the typing of nodes and edges. Some of the nodes are considered as primary and others as secondary. The primary ones are considered such that they represent the information regarding the discourse domain and the secondary ones provide the detail or data for that. Of two related primary nodes the one would then be considered as less abstract that has a more detailed description. The more abstract node would in the diagram then be placed above the less abstract one. In doing so the revised metaphor en-

ables using Plato's idea of heaven of ideas by associating top with heaven, and bottom with earth.

A sophisticated concept of intra model abstraction was discussed in [10]. In that paper a cohesion C is understood to be a pair $C = (A, P_A)$ where A is a set of roles or perspectives and P_A is a mapping $P_A : \mathcal{P}(\bigcup_{a \in A} \varepsilon(a)) \rightarrow \{true, false\}$ that defines which entities in the union of the extensions $\varepsilon(a)$ of roles $a \in A$ are related to each other by C .⁶ Intra model abstraction was understood as provided by abstraction concepts such as aggregation, classification, and generalization. An abstraction concept was in that source understood as a mapping $\alpha : (A, P_A) \rightarrow (B, P_B)$ where (A, P_A) and (B, P_B) are cohesion and $B \subseteq A$, as well as $P_B = P_A|_B$, i.e. P_B is what one gets by restricting P_A to the roles in B . The metaphor INFORMATION IS A NETWORK together with the idea of typing nodes and edges suggests the concept of cohesion, as one node (i.e. information chunk) can be typed such that it represents a cohesion and the edges that connect it to other nodes can be understood as the function of the latter nodes in that cohesion. This concept of abstraction can be described by the metaphor ABSTRACTION IS A SEMANTIC EXCEPTION FROM COHESION. The latter is similar to the metaphor ABSTRACTION IS A SEMANTIC MAPPING OF FORMULAE (see [36]) because cohesion in a model can be understood as a formula.

The metaphor INFORMATION IS A SPATIAL NETWORK can be extended such that it exploits not only the vertical spatial dimension (i.e. up vs down). Rather the depth space dimension can be included as well (i.e. front vs. back). In that extended version the metaphor enables relating perceived size of a concept representation as proximity to the model user. That which appears to be larger is closer to the model user and thus more relevant for him or her. The spatial dimension of breadth (i.e. left vs. right) can be ex-

⁶For a set S the set of subsets of S is denoted as $\mathcal{P}(S)$.

ploited too. For example, in modeling with state charts it is a very common to arrange them such that object creation occurs in the top left corner of a diagram and that the sequence of object states is arranged in lines from left to right and from top to bottom of the diagram.⁷ A respective metaphor that could justify that implicit convention is INFORMATION IS A TEXT.

Proximity of information chunks in a model (and thus recognized by the model user) can be understood as meaning relatedness. Similarly the metaphor INFORMATION IS A CONTAINER can be used to address encapsulation and protection.

8.2. Inter model abstraction

We understand semantic information as something that can be used for eliminating or reducing uncertainty. That is what makes us interested in having information about many entities of the one or another discourse domain. This understanding is coherent with the well-known definition of information as the distinction or difference that makes a difference ([2]). Aiming at applying this view of information to our problem of defining inter model abstraction makes us focus at the ways in which phenomena are included into models. The details of the used modeling language come to mind. Since we are interested in a pragmatic theory of abstraction we do not, however, focus on technical or formal detail. We rather again make use of the metaphor INFORMATION IS A NETWORK and our explanation of it provided above.

Consider an entity O the history of which is captured as the sequence $\{O_i\}_{i \in I}$. Let $i \in I$ and O^* be a semiotic representation of O_i . Consider model relationships $R(O^*, M, A)$, and $R(O^*, M', A)$ with M and M' being semiotic entities. Consider the predicates $c_i(o, X)$, $t_i(o, X)$, $r_i(o, X)$. Let them mean that phenomenon o is being included in model X of O^* , that it is typed the right way in X and that it is

⁷We thank Bernhard Rumpe for mentioning this observation to us.

related the right way to other phenomena included into model X of O^* . We can then define the predicate $\alpha(M', M, P, O^*)$. It means that model M' is more abstract than model M with respect to purpose P of actor A and original O in state i if $|\{o \in O_i | c_i(o, M') = true \text{ or } t_i(o, M') = true \text{ or } r_i(o, M') = true\}| \leq |\{o \in O_i | c_i(o, M) = true \text{ or } t_i(o, M) = true \text{ or } r_i(o, M) = true\}|$.

Obviously more research is required for working out an operational approach to achieving more abstract models that meet given requirements such as given queries can be efficiently processed. Our idea in that respect is to investigate design primitives such as the one in [3] with respect to their impact on inter model abstraction.

9. Outlook

It would be of some interest to have a more complete list of the metaphors that are used in CM.

As we have introduced two kinds of abstraction it is our aim in future work for each item L in a given list of modeling languages to provide a list of abstraction concepts that enables for a given model M to create all the models that are more abstract than M . The top-down primitives proposed in [3] are not complete, i.e. do not enable constructing all entity-relationship diagrams. Providing a sufficiently expressive set of abstraction concepts must thus considered to be a non-trivial task.

The idea proposed in [28] to use the Kolmogorov complexity K as a simplicity measure for bit strings that represent a state of a world W appears as promising. It would be interesting to work out formulae for the complexity $K(p : M)$ for each model M and each design primitive p taken from a set P of design primitives. It could be possible to use K as a measure of model complexity and to find strategies for complexity reduction that preserves information contents, i.e. specifies the same discourse domain model.

References

- [1] John Longshaw Austin. Zur Theorie der Sprechakte. Philip Reclam jun., Stuttgart, 1979.
- [2] Gregory Bateson. Mind and nature: a necessary unity. Dutton, New York.
- [3] Carlo Batini, Stefano Ceri, and Shamkant Navathe. Conceptual database design. The Benjamin / Cummings Publishing Company; Inc., Redwood City, California, 1992.
- [4] Rudolf Carnap. The logical construction of the world, (In German). Felix Meiner Verlag, Hamburg, 2nd ed. 1961.
- [5] Peter Chen. The entity relationship model: toward a unified view of data. ACM Transactions on Database Systems 1(1976), pp. 9 - 37.
- [6] Umberto Eco. Introduction to semiotics, (In German). Wilhelm Fink Verlag & Co. KG, 1994.
- [7] Pierre Luigi Ferrari. Abstraction in mathematics. In Lorenza Saitta, editor, The abstraction paths: from experience to concept. Philosophical transactions of The Royal Society, Biological Sciences, vol. 358, no. 1435. July 2003, pp. 1225 - 1230.
- [8] Robert C. Holte and Berthe Y. Choueiry. Abstraction and reformulation in artificial intelligence. In Lorenza Saitta, editor, The abstraction paths: from experience to concept. Philosophical transactions of The Royal Society, Biological Sciences, vol. 358, no. 1435. July 2003, pp. 1197 - 1204.
- [9] Philip Johnson-Laird. A taxonomy of thinking. In R. J. Sternberg and E. Smith, editors, The psychology of human thought, Cambridge University Press, 1988, pp. 429 - 457.
- [10] Roland Kaschek. A little theory of abstraction. In Bernhard Rumpe, Wolfgang Hesse, editors, Modellierung 2004: Proceedings zur Tagung. GI Lecture Notes in Informatics, P-45, 2004.
- [11] Roland Kaschek. A characterization of co-retracts of functional structures. In W. Ghler and G. Preuss, editors, Categorical Structures and their Applications. World Scientific Publishers, New Jersey et al., 2004.
- [12] Roland Kaschek. Modeling ontology use for information systems. In Klaus-Dieter Althoff, Andreas Dengel, Ralph Bergmann, Markus Nick, Thomas Roth-Berghofer Th., editors, Professional Knowledge Management. LNCS 3782 Springer Verlag, 2005.
- [13] Roland Kaschek. Preface. In Roland Kaschek, editor, Intelligent assistant systems: concepts, techniques, technologies. To appear IDEA Group Inc., 2006.
- [14] Wilhelm Kamlah and Paul Lorenzen. Logical propaedeutics: preschool of sensible speech, (In German). Verlag J. B. Metzler, Stuttgart, Weimar, 1996.
- [15] Roland Kaschek and Alexei Tretiakov. Knowledge evolution: the metaphor case. In Proceedings of CSIT 2006, Amman 5 - 7 April 2006.
- [16] Roland Kaschek and Claudia Kohl and Heinrich Mayr: Cooperations- an abstraction concept suitable for business process reengineering. In J. Györkök and M. Krisper and H. Mayr, editors, Re-Technologies for Information Systems, ReTIS'95 Conference Proceedings, Oldenbourg Verlag, Wien, München, 1995.
- [17] Roland Kaschek and Klaus P. Jantke and Tibor-Istvan Nebel. Towards understanding meme media knowledge evolution. In Klaus P. Jantke and Aran Lunzer and Nicolas Spyrtatos and Yuzuru Tanaka, editors, Federation over the Web. LNAI 3847 Springer Verlag, 2006.

- [18] George Lakoff. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, 4th. printing of second edition of 1992. Cambridge University Press, 1998.
- [19] Peter C. Lockemann and Heinrich C. Mayr. Computer supported information systems, (In German). Springer Verlag Berlin, Heidelberg, 1978.
- [20] Angelika Linke, Markus Nussbaumer, and Paul R. Portmann; *Transscript linguistics*, (In German). Max Niemeyer Verlag, Tübingen, 4th. unchanged edition 2001.
- [21] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago Press, Chicago, 1980.
- [22] Werner Metzger and Martin Schuster. *Learning to learn*, (In German). Springer, Berlin, Heidelberg, 7th. edition, 2006.
- [23] Andrew Ortony (editor). *Metaphor and Thought*, 4th. printing of second edition of 1992. Cambridge University Press, 1998.
- [24] Alexander Pfänder. *Logic*, (In German). Verlag von Max Niemeyer, Halle a. d. Saale, 1921.
- [25] James Rumbaugh et al. *Object-oriented modeling and design*. Prentice-Hall, Englewood Cliffs, NJ, 1991.
- [26] James Rumbaugh and Michael Blaha. *Object-oriented modeling and design*, 2nd edition. Prentice-Hall, Englewood Cliffs, NJ, 2004.
- [27] Hans Jörg Sandkühler. *Nature and knowledge cultures: Sorbonne lectures on pluralism and epistemology*, (In German). Verlag J. B. Metzler, Stuttgart, Weimar, 2002.
- [28] Lorenza Saitta and Jean-Daniel Zucker. A model of abstraction in visual perception. *Applied Artificial Intelligence* 15,8(2001), pp. 761 - 776.
- [29] Reinhard Schuette and Thomas Rothow. *The Guidelines of Modeling - an approach to enhance the quality in information models*. Proceedings of the 1998 Entity-Relationship conference.
- [30] Herbert Stachowiak. *General model theory*, (In German). Springer, 1973.
- [31] Herbert Stachowiak. *Knowledge levels towards systematic neo-pragmatism and general model theory*, (In German). In Herbert Stachowiak, editor, *Modelle: Constructions of reality*, (In German), pages 87 - 146. Wilhelm Fink Verlag, München, 1983.
- [32] Herbert Stachowiak. *Model*, (In German). In Helmut Seiffert and Gerard Radnitzky, editors, *Handlexikon zur Wissenschaftstheorie*, pages 219 - 222. Deutscher Taschenbuch Verlag GmbH&Co. KG, München, 1992.
- [33] Bernhard Thalheim. *Entity-relationship modeling*. Springer-Verlag, Berlin, Heidelberg, 2000.
- [34] Allan Tucker, editor. *Computer science handbook*, Chapman & Hall / CRC and ACM, 2004.
- [35] Can Türker and Gunther Saake. *Object-relational databases* (In German). dpunkt, 2006
- [36] Jean-Daniel Zucker. A grounded theory of abstraction in artificial intelligence. In Lorenza Saitta (ed.) *The abstraction paths: from experience to concept*. *Philosophical transactions of The Royal Society, Biological Sciences*, vol. 358, no. 1435. July 2003, pp. 1203 - 1309.

Data Mining in Biological Data for BiOkIS

Gunar Fiedler, Bernhard Thalheim
Department of Computer Science, Kiel University,
Olshausenstr. 40, 24098 Kiel, Germany
Email: {fiedler,thalheim}@is.informatik.uni-kiel.de

Dirk Fleischer
Institute for Polar Ecology, Kiel University,
Wischhofstr. 13, 24148 Kiel, Germany
Email: dfleischer@ipoe.uni-kiel.de

Heye Rumohr
Leibniz-Institute of Marine Sciences, Benthic Ecology,
Düsternbrooker Weg 20, 24105 Kiel, Germany
Email: hrumohr@ifm-geomar.de

Abstract

Modern ecological research attempts to measure and explain global dependencies and changes and therefore needs to access and evaluate scientific data just in time. At the same time, the uniqueness of biological data has to be taken into account by ensuring a safe long term data storage with high availability as a base for new biological studies. The bio-ecological information system BiOkIS attempts to provide this platform. It offers a safe data store for biological data and promotes data exchange between researchers or within distributed research groups. In addition to a simple data archive BiOkIS will provide data evaluation methods to participating researches as an inducement for making research data available to others. This paper will introduce the problems of modern biological data management and sketches the possibilities of offering archive and mining facilities to non database experts.

1. Introduction

The efforts of large scale ecological research are nowadays handicaped by the lack of an information infrastructure that supports a free and uncomplicated data transfer between researches. The marine ecological research in Kiel with its long tradition was involved in long term observations for more than 20 years and participated in ecological projects with partners distributed over several countries. The experiences from these projects showed

that the existing complex and restrictive data exchange protocols between research partners only lead to an overall confusion and faulty data transfers. Although collaboration is explicitly promoted, the exchange rate of data between researches is not substantially increased.

From a technical point of view biological data can be separated in different categories, for example data obtained from biological surveys (survey data) and data obtained from biological experiments.

Survey data is usually of a descriptive character. To obtain biological survey data an area of interest is defined and samples are taken. These samples are associated with their geographic references (latitude and longitude values). The sample's parameters of interest, e.g. number of individuals of a certain species, or the individual's size, are determined by applying specific techniques. Where appropriate, the data is compared with results from former surveys. Projects that describe regional communities of species or changes within a region collect a huge amount of data over the years, e.g. 2500 records for 5 stations per year.

If hypotheses are made based on survey data, there is always the need for an experiment that verifies the result. A single biological experiment produces approximately 1000 records. Due to modern procedures of taking samples this size is even increased.

Additional to these traditional invasive methods of taking samples out of a habitat, photographs and video sequences are used for visual scans without removing individuals from the habitats of interest and for tracking species hardly represented by quantitative methods.

Geographically referenced data can be used e.g. for creation of distribution maps for species or other (higher) taxa. Using obtained biomass data it is possible to estimate the number of individuals of certain species in different regions.

Table 1 gives an impression of the size of available biological data. It shows the amount of survey data and experimental data available at the Research Group for Marine Benthic Ecology at the Leibniz Institute of Marine Science and the Kiel University. The data is geographically referenced and described by meta data. Figure 1 shows a map of stations where survey data is obtained. Addition-

ally, the research group processed numerous historical data sources from the last 100 years, e.g. diploma theses and PhD theses. This data is now annotated, documented, and available in an electronic form.

This stock of data seems to be a solid ground for formulating hypotheses on biological questions. But unfortunately, the management of biological data has to face several problems on the technical as well as on the organizational level that prevent an efficient evaluation. The technical problems are the typical ones known e.g. from distributed systems and data integration scenarios ([5]):

Heterogeneity of data: data provided by different researchers or research groups differ from each other. Due to the organization of biological research, data sets are designed based on different points of view. Depending on the utilization, the personal goals and expectations, each researcher chooses its own structures. There is a common agreement about procedures and the overall set of concepts that are represented in data sets, but it is very costly to fully map raw data from different sources.

Discretization of data or **conversion** of continuous data to discrete data leads to different interpretations of data. Discretization may be based on time, space, or other abstractions which may vary among different research groups.

The **scope of data representation** is often concentrated on the scope of the user. This leads to the representation of macro data in the data sets which are comprehensions of micro data. Additionally, **data abstractions** are often used instead of basic data. Attribute names are usually not documented, abbreviations are quite commonly used. The data sets are full of symbolic values and artificial identifiers without any meaning outside this par-

Type	Estimated Number of Records
survey data (quantitative)	ca. 250,000 records
survey data (qualitative)	ca. 100,000 records
experimental data	ca. 150,000 records
photographs	ca. 6500
video sequences	ca. 530 hours

Table 1: data collected by researches from Kiel

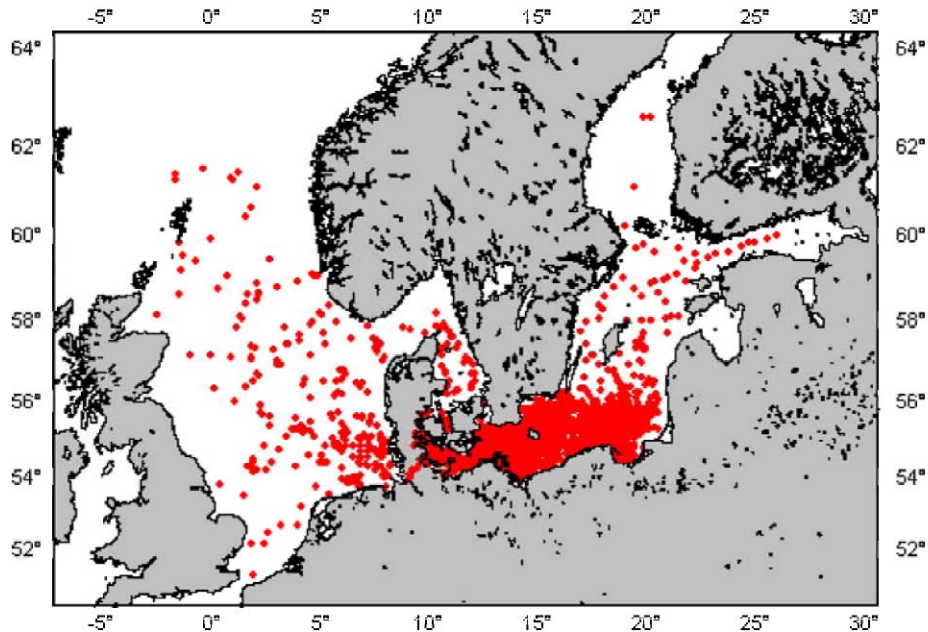


Figure 1: survey stations

ticular data set and without any documentation. Type information is often missing, e.g. because the data set is represented as a plain text file. This leads to additional problems concerning the quality of data at the instance level: values are missing, attributes are coded differently (e.g. using degrees with decimals for geographic references vs. using degrees with minutes of arc and decimals) or have wrong values.

The organization of biological research also leads to specific problems which prevent an effective usage of data. Most of the work is done during student's work or other time limited projects, e.g. while stu-

dents write their diploma theses or PhD theses. All these students are pressed for limited time only, so local and personal optimizations of obtaining, storing, and evaluating data naturally occur. Biological researchers are no computer scientists and so, they are usually not familiar with aspects of data modeling, data exchange, and long term data storage. Especially the description of data by meta data and backup procedures are often not present in mind. After the thesis is finished the researcher usually leaves the research group, home directories are deleted, and raw data is lost. Only comprehensions

like reports and publications remain. But even if raw data can be preserved, the knowledge about the data structure, the concepts behind the data, the description of the procedures that were applied to obtain the data is lost. That's why biological data sets may end up uninterpretable or have to be downscaled after a few years or even months.

Publication of research results usually doesn't contribute to the public outreach, because publications only contain comprehensions of data, raw data remains at the author. The usual peer review procedure delays publications together with the reuse of data sets for several months.

Additional to these arguments there are some further aspects that prevent an efficient reuse of biological data. Failed experiments are usually not documented or even published. But especially these failures may be important for the research community because learning from other failures may speed up the development of functional solutions. Distributed research projects often lack of an efficient data transfer between local groups due to technical difficulties and complex data transfer protocols. The interested public is excluded from participating in scientific results due to journal publication, so research groups have to make additional investments for public relationship management and popular science.

2. Existing Information Systems

The problem of data loss in biological research is known for a couple of years. There exist some information systems that try to address this topic. The world's data archive for environmental, marine, and geological data PANGAEA (www.pangaea.de) ensures long term archiving of geographically referenced data. Researchers are called to store their data in PANGAEA. It was in-

vented to store geological data from core drillings. This causes disadvantages: the existing data model has to be used to store biological data which is structurally and semantically totally different. PANGAEA is bringing data online right after publication, but unfortunately, diploma theses or student research projects are not considered as real publications. Therefore, they will seldomly become available. The data transfer protocols of PANGAEA are complex to use for many researchers. So many problems addressed above remain unsolved.

There is a number of highly specialized information systems for research data as well as popular science. For example, FishBase (www.fishbase.org) is a database containing facts about fishes. It can be used by taxonomists, fishermen, teachers, or pupils and provides facts about species as well as multimedia content like photographs. FishBase as an example shows the possibilities of providing scientific information over a Web based information system. It addresses the needs of the public due to its great stock of available photographs.

ReefBase (www.reefbase.org) is the first online information system concerning coral reefs. It provides information and services for professionals e.g. in the areas of management, science, and pollution control. ReefBase provides a great variety of geographically referenced and annotated photographs. Visualization is supported using a mapping service that produces interactive maps. This enables a playful handling of information.

The Ocean Biogeographic Information System OBIS (www.iobis.org) provides downloadable data sets from a network of institutions from 45 nations. Unfortunately, these data sets are restricted to observations (presence of individuals, not absence). Distribution maps that reveal

distribution pattern for a great number of species all around the world can be generated online.

3. General Goals

The problems mentioned above can be faced by using an information system with a central data storage that is capable of providing extensive descriptions and documentation of arbitrary data sets from the known domain. In detail, the following points are currently missing:

Controlled data input: data transfer into the system should be user-oriented and not backed by complex exchange protocols or data formats. Data import should take place as soon as possible, in the ideal case directly when data is obtained. The user should be guided through the import procedure, integrity checks should be applied both on syntactical as well as on the semantical level. In the case of detected errors, the user should instantly correct the data. Common data formats that are used by researches like CSV or spreadsheet formats have to be supported.

At each step the user has to be forced to annotate the data. Who obtained the data? Which procedures were made? Where the survey / experiment took place? What is the meaning of the record's attributes? How is the data encoded? Is this data set based on special assumptions? Most errors will be detected on the syntactical level, but it is also possible to apply a basic set of tests whether the data set is plausible according to the common knowledge of the research community for this domain.

Association between data sets: very important is the linking between depending data sets. Which experiments are based on which surveys? Which evaluations were already made for a certain data set? Which publications exist? Which research group used a particular data set?

Which data sets are geographically or taxonomically related? Is there any multimedia material supporting this data set? Many of these associations can be automatically obtained based on the data's usage or meta data.

Searchable data sets: Who is working on a similar topic? Are there data sets that I need for my hypotheses? Did the experiment I am currently planning already failed in a similar context?

Controlled output of biological data: Two points are important: first, the support for common data formats used in the research community and transformations between these formats. Second, there is a need for a mechanism that controls access to data sets such that every provider of data sets is notified when data is downloaded or displayed.

Virtual working groups: Data sets grow over time. To publish a data set is only useful if major steps of the survey or experiment are finished. Additionally, researchers are usually not interested in an uncontrolled distribution of the raw data; especially, if own publications are still in the queue. On the other hand it is adequate to share unfinished and locally incomplete data sets among project colleagues or trusted partners. After a pre-defined period of time where the data is reserved for private use it is made available to the public, but still under download control.

Cooperation with other systems: Annotated and documented data sets can be automatically transferred to cooperating information systems like PANGAEA or OBIS so researchers save time for data publication. For that reason, open standards and formats have to be supported like the Darwin Core exchange format ([6]) developed by the Taxonomic Databases Working Group ([8]).

A system that fulfills these requirements can be used as a library for sophisticated data evaluation. In principle, there are two possibilities for using this data: In the simple case data sets are searched, downloaded, and processed locally by third-party statistical or data mining software. As a surplus value to providers of data sets the information system can provide a set of evaluation methods, so data providers do not need to care about technical questions, e.g. obtaining and installing complex software just to process common evaluation techniques. In biological research, a couple of methods are established, e.g.:

- Bray-Curtis similarity matrices (similarity between samples according to found species and individuals, [2])
- Simpson Index (diversity index based on the number of species, [13])
- Margalef Index
- Shannon-Weaver Index (diversity index based on the number of species and individuals, [12])
- Pielou Evenness (analysis of the distribution of individuals, [10])
- taxonomic distinctness ([3])
- AMBI, BQI (diversity indexes for classifying ecological quality, [1, 11])
- statistical significance tests

Beside these statistical evaluation methods there are a couple of standard visualization techniques that can be applied, e.g. for creating project summaries or other kinds of reports. Some techniques may be:

Visualization of geographically referenced data using **annotated maps**: Figure 2 shows a screenshot of Google Earth ([7]) with a data set that is visualized according to the geographical references of the records. Because of the facilities of navigating in Google Earth together with

its support of linking placemarks with Web resources it is easily possible to produce distribution maps in different styles or to display measures or multimedia elements according to their coordinates.

Survey data can be visualized by diagrams showing the number of individuals or species at a certain place over a period of time. This enables researchers to choose data sets for comparative studies or to make a long term observation of a geographical region of interest. Figure 3 show an example. The same discussion can be made for experimental data.

Figure 4 shows possibilities for data mining approaches. For example, by using Bray-Curtis similarity matrices it is possible to identify the species of importance for a data set. This may lead to a better understanding of the biological system.

4. Provided Services within a bio-ecological Information System

Looking at different types of users a bio-ecological information system will provide different services. Although the system will mainly be used by scientists, there are also interesting services for the interested public or sponsors of research projects.

The group of researchers can be divided in two groups: data providers and data users. Data providers publish annotated data sets by using the system. Data users are researchers that will use data sets stored in the system for further research.

Data providers are offered a long term archiving of their published data sets, so the researcher needs not to deal with backup procedures. If it is wanted, the data sets are distributed, e.g. to the world data center. As a surplus value, each data provider can use predefined data evaluation methods. Predefined data visualiza-

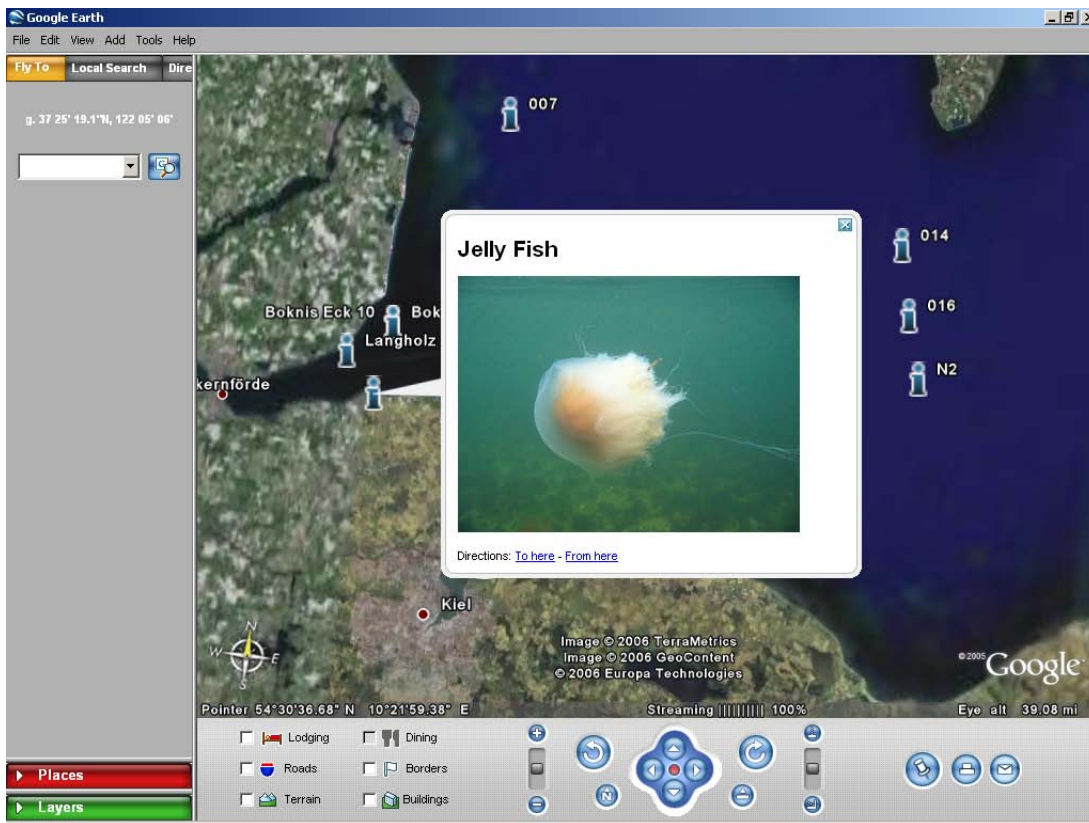


Figure 2: Data Visualization with Google Earth

tion can easily be used for public outreach. Each data provider is notified if some user accesses the provided data sets, so data providers always have an overview about the impact factor of their research. Researchers working in distributed working groups can define private areas for data that is currently under processing, so data transfer between local groups is simplified.

The online overview of downloaded or displayed data enables research sponsors to participate in data ascertainment and the project's progress. Within the system a statistical evaluation of queried data grouped by projects is possible. This evaluation continues after the project's lifetime and reveals the impact of the project within the research community or the public. Evaluation results can be linked to the sponsor's homepage to improve the sponsor's own public outreach.

Researchers acting as data users are offered an archive with semantically annotated, documented, and linked data sets that are searchable and browseable according to numerous parameters. Reports, visualizations, and publications related to data sets are accessible.

Because biology is a popular science the interested public should be explicitly included in any thoughts about a bio-ecological information system. Especially multimedial content that is produced during non-invasive surveys can be easily reused for public relations management as long as they are annotated. But there are also other interesting services possible like visual classification wizards that allow classification of individuals according to the taxonomy of species by iteratively presenting photographs of possible species.

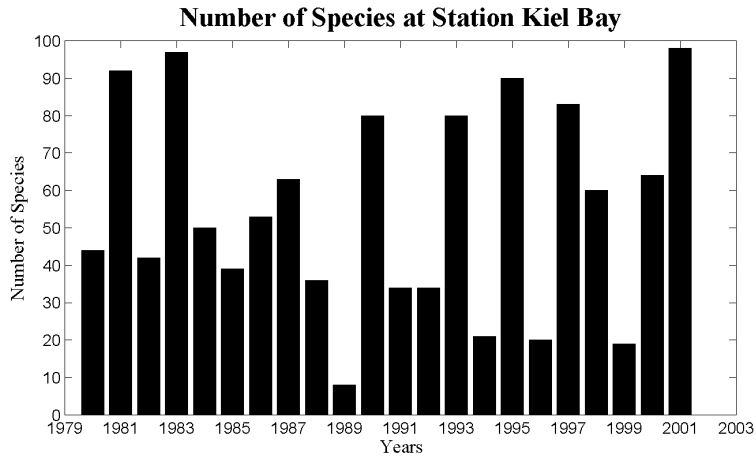


Figure 3: A diagram showing the number of individuals at a certain station over a longer period of time. Individuals of all species are summarized.

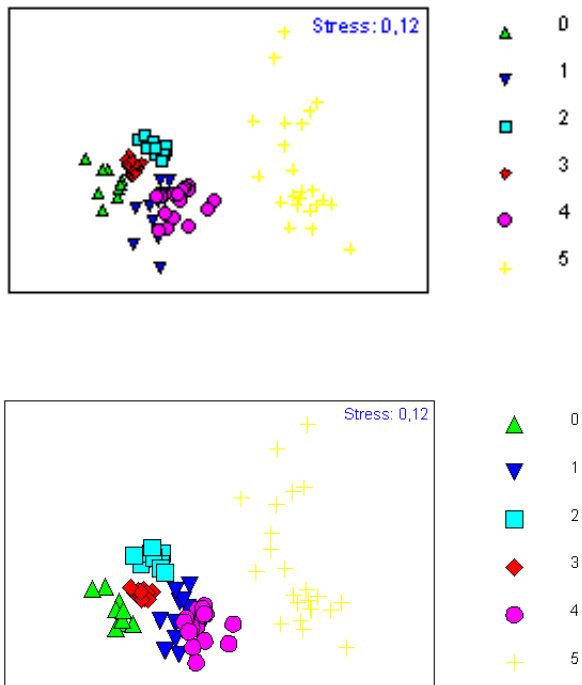


Figure 4: MDS plot of a time series of the seafloor of the Kiel Bay, based on ca. 150 species. The second picture shows the plot for 16 species. The most important species were extracted from the data set.

5. A General Architecture

From a technical point of view an information system with the described functionality has a number of points of contact with different disciplines in computer science. **Data warehouses** integrate data from different operative systems and offer numerous data evaluation procedures. In difference to classical data warehouses data in a bio-ecological information system will be structured more complex. **Content management systems** offer functionality to store such data and present it to users in specific ways.

To ensure data quality within the information systems, numerous domain specific integrity checks have to be applied. The (at least partial) explicit representation of these rules within the information systems will promote the long term understandability and that's why the long term usability of data sets. For this point many work was done in the area of **deductive database systems** as well as **active database systems**. Other technologies as well as languages for describing data can be found in the area of the **Semantic Web**.

Therefore, a general architecture for these kinds of information systems, called content warehouses can be derived. For a detailed description, see [4]. Typical system components are:

The central data store is responsible for representing the raw data. Due to the variety of data structures it has to support schema components that are flexible enough to integrate different points of view on the data but are structured enough to allow an efficient handling of mass data. For that reason, star- and snowflake schemas (see [16]) are introduced that define mandatory kernel types and optional types. The modeling approach defined in [5] introduces partially

collaborating schema fragments (called 'sunflower' schemas) that allow the co-existence of data that is shared between contexts and 'private', context dependent data.

The semantics of the data has to be available in an interpretable form. That's why the central data store is enhanced by a reasoning engine for managing terminological knowledge about the data as well as complex integrity constraints. This component strongly interacts with a component for the management of user profiles and portfolios to derive access rights and obligations based on the state of data sets and defined working groups.

The system's interface to the user provides plugable evaluation and visualization modules as well as import and export filters for common exchange formats. The interaction generator uses information from the user management to deliver information according to the user's needs, wishes, and permissions in a suitable style as explained in [15].

6. Conclusion

In this paper we discussed the challenges of using data obtained during biological research as a base for data mining. The central problems of data availability and data quality have to be faced. By integrating the functionality of a data archive system with data evaluation, data presentation, and workgroup functionality in an information system it is possible to invite more and more researchers to participate in creating a high-quality stock of biological data as a ground for new scientific work.

Acknowledgements

We like to thank our students Aylin Aksaç, Anna Jaworska, Helge Rabsch, and Meimei Xu for implementing the Google Earth visualization tool.

References

- [1] A. Borja, J. Franco, and V. Perez. A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos within European Estuarine and Coastal Environments. *Marine Pollution Bulletin*, 40:1100–1114, 2000.
- [2] J.R. Bray and J.T. Curtis. An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:325–349, 1957.
- [3] K.R. Clark and R.M. Warwick. A taxonomical distinctness index and its statistical properties. *Journal of Applied Ecology*, 35:523–531, 1998.
- [4] G. Fiedler, A. Czerniak, D. Fleischer, H. Rumohr, M. Spindler, and B. Thalheim. Content Warehouses. Preprint 0605, Department of Computer Science, Kiel University, March 2006.
- [5] G. Fiedler, Th. Raak, and B. Thalheim. Database Collaboration Instead of Integration. In Sven Hartmann and Markus Stumptner, editors, *APCCM*, volume 43 of *Conferences in Research and Practice in Information Technology*. Australian Computer Society Inc., 2005.
- [6] Darwin Core Exchange Format. <http://darwincore.calacademy.org/>.
- [7] Google, Inc. Google Earth (<http://earth.google.com>).
- [8] Taxonomic Databases Working Group. <http://www.tdwg.org/>, 2006.
- [9] T. Kruscha, B. Briel, G. Fiedler, K. Jannaschk, Th. Raak, and B. Thalheim. Integratives HMI-Warehouse für einen durchgängigen HMI-Entwicklungsprozess. In VDI, editor, *Elektronik im Kraftfahrzeug 2005. 12. Internationaler Kongress Electronic Systems for Vehicles*, number 1907 in VDI-Berichte. VDI, VDI-Verlag, 2005.
- [10] E.C. Pielou. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144, 1966.
- [11] R. Rosenberg, M. Blomqvist, H. Nilsson, H. Cederwall, and A. Dimming. Marine quality assessment by use of benthic species abundance distribution: a proposed new protocol within the European Union Water Framework Directive. *Marine Pollution Bulletin*, 49:728–739, 2004.
- [12] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1963.
- [13] E.H. Simpson. Measure of diversity. *Nature*, 163:pp 688, 1949.
- [14] B. Thalheim. *Entity-relationship modeling – Foundations of database technology*. Springer, Berlin, 2000. See also <http://www.informatik.tu-cottbus.de/~thalheim/HERM.htm>.
- [15] B. Thalheim. Co-Design of Structuring, Functionality, Distribution, and Interactivity of Large Information Systems. Technical Report 15/03, Brandenburg University of Technology at Cottbus, 2003.
- [16] B. Thalheim. Component development and construction for database design. *Data Knowl. Eng.*, 54(1):77–95, 2005.

Personalized information filtering for mobile applications

Christian Reuschling and Sandra Zilles
DFKI GmbH
Erwin-Schrödinger-Str. 57
67663 Kaiserslautern, Germany
{christian.reuschling,sandra.zilles}@dfki.de

Abstract

Recent technological development has enhanced research in the field of pervasive computing for mobile applications. In particular, topics like ad-hoc community building and personalized contextual product offers are of relevance for cellular radio providers. In this context, we propose a generic approach to personalized information filtering.

This concerns first an appropriate representation scheme for the application domain, the user, as well as the information to be filtered. Here we propose to model the domain in an ontology with special weight attributes in RDFS, such that personal interests or resources (e. g., product descriptions) can be represented as RDF instances of this ontology.

Using case based reasoning techniques, we second propose an implementation of a similarity measure between such instances. On the one hand, given a special domain model, this similarity measure allows for filtering a list of resources according to a person's interests in a way immediately suitable for the intended applications; on the other hand, this similarity measure is defined generally enough to allow for the comparison of RDF instances in general, with different specialized similarity measures depending on the intended semantics of similarity.

Third, in addition, the problem of how to maintain representations of user interests and resource descriptions in a dynamic domain is addressed briefly.

1. Introduction

Recent technological development, such as concerning UMTS and smartphones, has enhanced research in the field of pervasive computing for mobile applications. In particular, topics like socializing, ad-hoc community building, entertainment on demand, and personalized contextual product offers are of relevance for cellular radio providers and for the corresponding third party providers. What many products cellular radio providers aim at have in common is the fact that they implement

- a model for representation of the user,
- a model for representation of a user's context¹,

¹i. e., additional constraints describing the current situation of a user's fluently changing environment

- a model for representation of available information or resources (e. g., products, entertainment offers, other users²),
- a scenario for filtering and recommendation of information or resources,
- a process for content acquisition (including maintenance of the information represented).

Here the main difference compared to classical, non-mobile applications is the focus on the user's context, in particular, e. g., her current location, the current daytime, current and/or previous tasks, active applications and services, etc., cf. [15, 9, 14]. Roughly

²Note that users themselves can be resources, for instance, in applications focusing on socializing scenarios.

speaking, the user’s context is defined by the current status of her fluently changing environment, see Section 2.3 for more details.

However, the actual implementations of these models, scenarios, and processes differ in terms of methods used. The variety of the approaches of course results from different applications and thus different intended semantics of user/resource representations and different intended recommendation features. State of the art approaches to modeling personal taste for recommendation tasks are for instance discussed in [11], cf. also the references therein. Ideas especially aiming at mobile applications often concern socializing based on, for instance, music recommenders, cf. [2, 3].

The aim of this paper is to propose a unified framework allowing for a generic implementation of the required models, scenarios, and processes – independent from the actual application domain. This framework is based on previous work concerning

- Semantic Web technologies (here used for the representation model); see [6, 12] and
- case based reasoning methods (here used for the recommendation model); see [4, 5], as well as [1] for a fundamental overview.

The fundamental approaches combined are not new; however, the combination of Semantic Web technology with case-based reasoning methods in a unified framework as proposed below hopefully provides a fundament for fruitful application-oriented research.

Figure 1 sketches the underlying scenario in the scope of the applications our approach addresses. Note that the term ‘case base’ here represents a database containing all profiles – the latter being regarded as cases in a case-based reasoning approach. We will return to this point of view in Section 3. Basically, the idea is that a user will receive personalized, filtered information via a mobile device (e. g., cell phone). The recom-

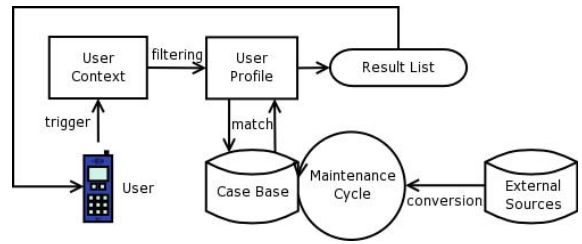


Figure 1: The general scenario.

mendation process is to first use context information for filtering the case base, and then select appropriate ‘cases’ (i. e., profiles) depending on a similarity match with the given user profile. The system has to be maintained in a clearly defined workflow, for instance by (but not restricted to), using external databases.

The paper is organized as follows: we first motivate and describe our chosen model for representing user and resource profiles, i. e., descriptions of their interests or features, (Section 2), then, in the main section (Section 3) we discuss our generic recommender approach, followed by a brief overview over possible approaches to content acquisition (Section 4). In the concluding section, we address some possible issues for future work in this context.

2. Representation model: the RDF ‘boost factor’ framework

2.1. Basic requirements

Let us first motivate our choice of the model for representing user interests (i. e., user profiles) and resource properties (i. e., resource profiles).

First of all, note that, particularly for mobile applications, a user should never be described by her profile alone, but also by her current context. Subtracting context from a user description, it is straightforward to use only a single model both for user profiles and for resource profiles. Requirements for a corresponding representation scheme should be:

1. The representation scheme must be general enough to express different

kinds of real-world concepts and their relations. In particular it must provide options for

- (a) representing taxonomies of concepts (in order to allow for proper recommendations of specialized to generalized profiles and vice versa),
 - (b) representing the amount of user interest in each of the real-world concepts (in order to allow for more specially personalized recommendations),
 - (c) transferring user or resource profiles from (publicly) available databases into the predefined representation scheme (in order to allow for content acquisition in a recommender system),
 - (d) modifying a given domain model as well as the profiles (in order to allow for reflecting the changes in a dynamic application domain).
2. The representation scheme must be specialized enough to be suitable for efficient content acquisition and recommendation. In particular, it must allow for
- (a) an *efficient* extraction and transformation of external data,
 - (b) an *efficient* computation of similarities between profiles (in order to determine recommendations),
 - (c) *efficient* profile generation and update tools.

In particular, the need for feasible content acquisition methods suggests using standards in the basic representation scheme. Moreover, basing on the requirement for representing taxonomies (profiles express some special interests or features of a collection of – maybe inter-related – concepts in the application domain), we chose an ontology-based approach for representation. That means, concepts of the domain are

modeled in an ontology, using a representation scheme expressive enough for representing taxonomies, whereas user and resource profiles are instances of this ontology (respecting some particular features, as will be explained in Subsection 2.2).

Currently, there are several standards for ontology representation schemes, e. g., RDFS, OWL (OWL Full and restrictions OWL DL and OWL Lite), etc., see also [17, 16]. On the one hand, a language like OWL Full is very expressive, but on the other hand, this expressiveness entails high costs in the tasks resulting from our requirements. Therefore a first reasonable step towards a generic framework for recommender systems is to choose RDFS/RDF for modeling the domain ontology and the profiles (as instances).

Note that an additional benefit from using RDF might be that RDF is the language of the Semantic Web, where many users express their personal interests within a net of linked RDF resources. This may entail further options for extracting contents for a recommender system.

Moreover, the quite simple structure of RDFS may have usability benefits; for real-world applications it would be rather inconvenient to model in OWL Full.

2.2. From general RDF to interest profiles

So we propose to model the application domain in an RDFS ontology; profiles are instances of the latter. However, some detailed requirements have to be taken into account, in order to use RDF specifically for representing interest profiles. These detailed requirements concern

1. the attribute types considered,
2. attributes for user rating in interest profiles,
3. degrees of relationships between concepts in the ontology.

Attribute types

Up to a certain degree of granularity, a simple metadata scheme may be sufficient for describing all relevant information about a person or a resource. However, one immediately recognizes limitations of such ontology-based metadata schemes. For instance, music is very hard to describe using metadata only, state-of-the-art music mining and retrieval techniques prove much more information in an audio file itself than can be expressed with simple metadata. Similar statements hold for text documents.

Thus including attributes of type mp3 or string in the ontology allows for integrating audio files or any kind of textual reviews/summaries (of music, literature, any kind of products, etc.) into a profile. Hence profiles can be enriched by additional information.

By the way, this also has a positive effect in easing profile generation, as will be discussed in Section 4.

Attributes for user rating

As already indicated, it should be possible to instantiate a concept in an ontology such that a person can express liking or disliking the class of objects associated to that concept. The motivation behind is that – under an open-world assumption – in general we cannot conclude a person dislikes a concept, if it is not instantiated in the person’s profile.

Our simple approach here is to demand that each concept in the ontology can have a special ‘weighting’ attribute, a float attribute with range $[-1, +1]$ reserved for expressing the so-called ‘boost factor’. In a person’s interest profile:

- a positive boost factor for some concept c indicates that the person rates c positively; the absolute value is interpreted as a weight in this rating,
- a negative boost factor for some concept c indicates that the person rates c

negatively; the absolute value is interpreted as a weight in this rating,

- a boost factor of 0 for some concept c indicates that the person is indifferent concerning c .

In a resource profile, the same kind of boost factor is reserved.

Note that the open-world assumption is reflected in our approach as follows: If, in a profile, a concept c is not instantiated, the intended semantics would not be equal to the case when c is instantiated with a boost factor of 0. The open-world assumption forbids us to interpret the rating of c as ‘indifferent’ (for persons) or ‘non-existing’ (for resources).

Degrees of concept relationships

Assume a taxonomy of concepts is modeled in an ontology. Then the intended semantics probably cannot always be reflected in the taxonomic structure only, as the following example shows:

Assume the ontology contains concepts c , c_1 , and c_2 , where c_1 and c_2 are subconcepts of c , see Figure 2.

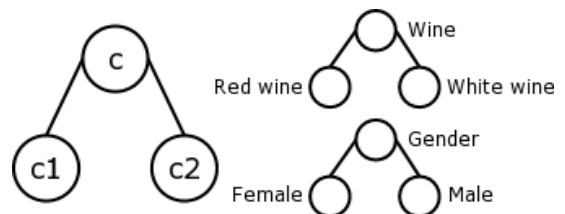


Figure 2: A simple example for taxonomies.

Then c_1 and c_2 are ‘brother’ concepts and could be interpreted as related. But the degree of their relation is not clear without any additional specification which might depend on the intended application. Focusing on recommender systems, a strong relation would mean that profiles instantiating c_1 could be recommended to profiles instantiating c_2 . Obviously, the intended recommender functionality depends on some kind

of *degree* of relationship you would like to assign to c_1 and c_2 . If the intended semantics would say that ‘red wine’ is closely related to ‘white wine’, but ‘female’ is not (closely) related to ‘male’ (which should definitely have an impact on how to compute recommendations), this requires some additional representation within the ontology. In particular, the concept c should be annotated with some value expressing the degree of relationship of its child concepts.

Thus, in the proposed generic model, each concept in the ontology has a special numerical attribute (e. g. with range $[0, 1]$) used for representing the degree of relationship of its child concepts. Extreme values then are interpreted as the maximum or minimum possible degree of relationship, i. e. if the range is $[0, 1]$, then the value 0 for c means the sons of the concept c are not related at all; the value 1 means that the sons of c are just representations of only one subconcept of c .

2.3. Context in addition to a user profile

A generic representation model for user context is not so easy to define. In general, there is no universal definition of the term context, cf. [7, 8], though this topic has gained a lot of interest in the scientific community, see also [9, 14]. In Section 1, we explained the user’s context as the current state of her fluently changing environment – but this is far from being a clear definition. Obviously, in mobile applications, location and time should be part of a user’s context, but it is unclear, how these should in general be evaluated by a recommender system. A generic approach would have to allow for reflecting any particular application-specific definition of context and its semantics.

Up to now, we have not yet elaborated a generic context model for mobile applications, though in several realizations of our proposed approach (recommender systems for leisure time, for shopping scenarios, and for media recommendations) the user’s context has been integrated.

Currently, we define context from a more

functionality-driven point of view: instead of asking first what context is, it is reasonable to ask first what context should be used for in the application. In our current model, we define the user’s context by all constraints that are used for pre-filtering information (in a database containing all profiles) before computing recommendations via similarities or for weighting the overall similarity values between profiles, which are determined for computing recommendations.

Note that it may be worth analyzing to what extent a user’s context could also be modeled as part of her interest profile, using the RDFS framework as explained above.

3. Recommender model: a generic similarity measure

From now on, suppose an ontology and a list of profiles have been defined according to our special RDFS/RDF representation approach. Additionally, assume some context information is attached to the profiles,

Given a profile p , how should recommendations for p be computed?

Let us first assume the existence of algorithms for computing

1. a similarity $\text{sim}(p, p')$ between the profile p and any other profile p' (according to some similarity measure),
2. a context relevance $\text{con}(p, p')$ of any other profile p' to profile p .

Then the recommendation process could be sketched as follows:

parameters: thresholds $\theta_s, \theta_c, \theta$, integer k

input: list L of all profiles,
query profile $p \in L$

output: list of profiles in L
(recommendations for p),
with similarity values in $[0, 1]$

1. (* optional *) Let $L' \subseteq L$ be the list of all profiles $p' \in L$, for which $\text{con}(p, p') < \theta_c$; $L := L \setminus L'$;

2. (* options: a, b, c, d *)

- (a) For all profiles $p' \in L$ with $\text{sim}(p, p') \geq \theta_s$, return $(p', \text{sim}(p, p'))$.
- (b) For all profiles $p' \in L$ with $\text{sim}(p, p') \cdot \text{con}(p, p') \geq \theta$, return $(p', \text{sim}(p, p') \cdot \text{con}(p, p'))$.
- (c) Return $(p_1, \text{sim}(p, p_1)), \dots, (p_k, \text{sim}(p, p_k))$, for profiles $p_1, \dots, p_k \in L$, such that $\text{sim}(p, p_i) \geq \text{sim}(p, p')$ for all $i \in \{1, \dots, k\}$ and all profiles $p' \in L$.
- (d) Return $(p_1, \text{sim}(p, p_1) \cdot \text{con}(p, p_1)), \dots, (p_k, \text{sim}(p, p_k) \cdot \text{con}(p, p_k))$, for profiles $p_1, \dots, p_k \in L$, such that $\text{sim}(p, p_i) \cdot \text{con}(p, p_i) \geq \text{sim}(p, p') \cdot \text{con}(p, p')$ for all $i \in \{1, \dots, k\}$ and all profiles $p' \in L$.

In other words, one usually recommends either all resources with a minimum similarity or just k resources with highest similarity. The context relevance may be used for pre-filtering and/or as a weight for the overall similarity.

So a generic framework requires a definition of a suitable similarity measure, and thus a generic algorithm for matching RDF files. We propose a matching algorithm which should be a fixed component in any concrete application task, such that, with each new application, it remains only to define a new ontology and a new context model (the latter including definitions of how to compute context relevance values). That means, the matching algorithm proposed is generic; its parameters are the ontology and the algorithm for computing the context relevance. Basically, the desired similarity measure is implemented in this matching algorithm in the sense that the latter compares two profiles and returns a similarity value in $[0, 1]$.

Requirements for such a similarity measure should concern local similarities, which must be defined for all relevant attributes, as well as the global (overall) similarity.

Local similarity

Concerning local similarities, there is a need for different type-specific similarity measures, e. g., for audio files of the types occurring in the ontology, for html documents, for general string values, and for numerical values. Here different similarities must be defined for different intended semantics of certain attributes:

Two strings s, s' can have a similarity defined by

1. 0, if $s \neq s'$ and 1, if $s = s'$; or
2. $1 - l(s, s')$, where $l(s, s')$ is the Levenshtein distance of s and s' (normalized in the interval $[0, 1]$), cf. [10]; or
3. ...

Two numerical values r, r' can have a similarity defined by

1. $1 - d$, where d is their absolute distance (normalized in the interval $[0, 1]$); or
2. 0, if $r \neq r'$ and 1, if $r = r'$; or
3. 0, if $r < r'$ and 1, if $r \geq r'$; or
4. 0, if $r > r'$ and 1, if $r \leq r'$; or
5. ...

Similar variants are conceivable for different types of attributes.

Note that each variant of a local similarity measure goes along with different intended semantics of the corresponding attribute. Consequently, this has to be expressed by formally different types of string attributes or different types of numerical attributes in the ontology.

The need for these different semantics is obvious: If, for instance, a numerical attribute represents a production year of a movie a user is interested in, then the local similarity for this attribute should be defined using their absolute distance, since the user may presumably like movies from that time, but not only from that year. The intended local

similarities would of course be different, if the attribute should represent some maximal allowed price for recommended products.

Global similarity

Note that, as for the case of local similarities, we assume $\text{sim}(p, p') \in [0, 1]$ for all profiles p, p' .

The overall similarity of two profiles should take into account the local similarities concerning different attribute values, as well as the degree of relationships between the instantiated concepts. Moreover, the user's interests expressed in the boost factors must define 'weights' in computing a similarity. Note that the boost factor is a special kind of attribute, the value of which is not included into the computation of similarities in the usual way. The boost factor merely determines the amount of how strongly the local similarity will contribute to the global similarity. Thus it works like a weight, though it is not really a weight, due to the fact that the boost factors can have arbitrary values in $[0, 1]$, without requiring that they sum up to 1, as would be usual for weights.

Here different requirements concerning a similarity measure are conceivable:³

1. Reflexivity

One might or might not require that $\text{sim}(p, p) = 1$ for all profiles p .

2. Symmetry

One might or might not require that $\text{sim}(p, p') = \text{sim}(p', p)$ for all profiles p, p' .

3. Triangle inequality

One should in general not require that $\text{sim}(p, p') \geq \text{sim}(p, p'') + \text{sim}(p'', p') - 1$ for all profiles p, p', p'' . The reason is, roughly speaking, that a user's boost factors can determine how much

special local similarities influence the global similarity value. Thus, even if all local similarities fulfill the triangle inequality, this does not hold for the global similarity.

So, in what follows, we should concentrate on reflexivity and symmetry constraints.

For instance, in a shopping scenario, if a user has rated her favorite movie with title t , a resource (e. g., a DVD) with the same title t would maybe not be the top recommendation, assuming the user already possesses such a product. Hence you would require a low similarity of a profile to itself, i. e., the corresponding distance measure would be non-reflexive. On the other hand, retrieval scenarios are conceivable, where each profile has maximum similarity to itself, i. e., where the global distance measure is reflexive. Similar scenarios are conceivable disallowing or requiring symmetry.

Now, if our global similarity measure is supposed to be generic, this means that whether or not reflexivity, symmetry, or the triangle inequality hold, must depend on the local similarities exclusively. Only in this case we can guarantee that the ontology model alone (and in particular the choice of special attributes for which local similarity measures are defined) determines the properties of the similarity measure.

Consequently, the main requirements concerning the global similarity measure can be summarized as follows:

1. The global similarity of two profiles must depend on their local similarities (in particular also concerning the boost factors) and the relationship degrees defined in the ontology, only.
2. Given a local similarity or a relationship degree as a parameter, the global similarity must be monotonically increasing in this parameter.
3. The global similarity function is reflexive, if and only if all given local similarity functions are reflexive.

³The requirements reflexivity and triangle inequality for a distance measure d are usually $d(p, p') = 0$ and $d(p, p') \leq d(p, p'') + d(p'', p')$. We obtain the given formulations regarding $\text{sim} = 1 - d$ for a $[0, 1]$ -valued distance measure d .

4. The global similarity function is symmetric, if and only if all given local similarity functions are symmetric.

The easiest way to achieve this is to define a global similarity by a weighted sum of all local similarities.

For this purpose we follow the case-based reasoning method proposed in [5]. Here the case base is just the profile database; each profile corresponds to a case. The query profile p is then understood as a new case, such that for each case p' (existing profile) in the case base, a similarity value for p and p' has to be computed.

The method explained in [5] can be sketched as follows:⁴

input: query profile p ,
case profile p'

output: similarity value $\text{sim}(p, p') \in [0, 1]$

1. Determine the set A of all attributes instantiated both in p and in p' (except for boost factor attributes).
2. For all attributes $a \in A$ compute a (local) similarity value $\text{lsim}(p, p', a)$ as follows:
 - If a is a complex attribute, then let $\text{lsim}(p, p', a) = \text{sim}(p_a, p'_a)$, where p_a and p'_a are the parts of the instances p and p' concerning a (* recursion *).
 - Else let $\text{lsim}(p, p', a)$ be the local similarity of p and p' concerning a .
3. For all attributes $a \in A$ compute a boosted local similarity value $\text{blsim}(p, p', a) = \frac{1}{2} \cdot (1 + \text{lsim}(p, p', a)(1 - |\text{bf}(p, a) - \text{bf}(p', a)|))$, where $\text{bf}(q, a)$ denotes the boost factor

⁴Here we deal only with a simplified case, assuming that no attribute in the ontology is of a list type. The algorithm can be extended to matching profiles in which lists occur as attribute types, but we omit the relevant details.

assigned to the attribute a in some profile q .⁵

4. Return the global similarity $\text{sim}(p, p')$ as the weighted mean of all values $\text{blsim}(p, p', a)$ for $a \in A$, where the weights are given by the degree of relationship between the concepts instantiated by p and p' .

4. Content acquisition model: generating and updating ontologies and profiles

Concerning the content in a recommender system in the proposed framework, both the ontology and the profiles have to be discussed. However, we only briefly sketch some well-known basic approaches here.

4.1. Ontology generation and update

How can an initial ontology be obtained and how can it be updated? In most cases, presumably, an initial ontology must be designed by an administrator – as usual. However, there are in fact approaches allowing for automatically computing suggestions for ontology updates, which can be used in the context of semi-automated ontology maintenance.

Non-automated workflow

Non-automated ontology modification here simply means manual editing. This only requires the provision of an interface allowing for ontology upload and download, with connection to an ontology editing tool.⁶

Semi-automated workflow

Though machine learning tools can be used for inferring new relations for the ontology

⁵The factor $\frac{1}{2}$ is needed for normalizing onto the interval $[0, 1]$.

⁶Note that some ontology editing tools will not store the ontology in the format required in our proposal. Here parsers must be used for transformation.

or for detecting superfluous/missing concepts or attributes, ontology changes should definitely always be executed only upon the decision of an administrator.

Suggestions for an ontology update can be obtained by automatically analyzing all profiles in the database or logging user actions – for instance, using association rule mining or clustering techniques, see [18]. Such techniques are well-known from market basket analysis.

4.2. Profile generation and update

How can profiles be generated and updated? Again we only sketch some basic ideas.

Non-automated workflow

Of course, it is possible to provide user interfaces for manual editing of profiles. However, this may cause problems concerning the usability as well as because of the often observed phenomenon of users incapable of describing their own interests in terms of concepts. Still manual editing might be reasonable for creating resource profiles or modifying them.

Semi-automated workflow

In semi-automated workflows, a human administrator can be assisted by machine learning systems inferring new knowledge from user and/or resource data.

Here we distinguish between ‘local’ approaches using single user data and ‘global’ ones using a larger part of the system content for learning.

1. Local approach

The interest profile of a user can for instance be obtained/updated by an automatic analysis from relevance feedback, see [13]. If the user rates resources, which have profiles, positively (or negatively), they can be merged into the user’s profile with boost factors as given in the resource profile (or boost

factors as given in the resource profile, but multiplied by -1). If the user rates resources without profiles, e. g., audio files or text documents, suitable mining methods may be used for extracting new profile features out of these.

This local approach works for acquisition of resource profiles as well, since, in general, features of resources can also be described by relevance feedback.

2. Global approach

User or resource profiles can be updated by an automatic global analysis of all profiles in the database (or of all user profiles of certain clusters/communities). If association rule mining or clustering methods, cf. [18], yield new relations between concepts, these can be used for predicting unknown values in single profiles.

Fully automated workflow

The most obvious approach for a fully automated workflow, as already mentioned before, is profile extraction from external databases – at least resource profiles can be automatically extracted from available databases. For instance, movie profiles could be obtained from the internet movie database (<http://www.imdb.com>).

Still note that some administrative tasks may be required concerning a comparison of the vocabulary (the metadata values themselves) used in existing profiles and those in the external resource descriptions.

5. Conclusions

We have proposed and implemented a generic framework for personalized information filtering, basing on a special RDFS/RDF representation model and a uniform algorithm computing a similarity measure for RDF instances.

The idea is that, for each new application, only a new ontology (and, if required, new local similarity measures) has to be defined.

The global similarity measure can be computed using a fixed uniform tool following our proposal.

However, there are still several aspects in which further details in our proposal have to be elaborated.

Efficiency/performance issues

Up to now we have not analyzed the efficiency of the proposed matching algorithm formally. A uniform approach however requires reliable statements concerning the conditions under which this algorithm works well. This should be one point of focus in future work.

Framework documentation

Even if a framework proves uniform for a class of applications, this does not immediately imply that it is uniformly usable in practice. Here a representation language is required which clearly explains the effect of all kinds of parameters in the ontology upon the resulting global similarity measure. In particular, if a designer of some new application has an intended similarity measure in mind, then it should be clear how to define an ontology appropriately. This is a further aspect which must definitely be addressed. Moreover, this aspect also includes the question of how the generic tools in our framework can be best embedded into a concrete application.

Context representation

As already indicated in Section 2.3, a suitable framework for context representation has to be defined; in particular when aiming at mobile applications this is of great importance.

Content acquisition and maintenance issues

Furthermore, the methods sketched in Section 4 have to be conceptualized in a way such that our generic framework is complemented by a toolkit easing content acquisition and update in concrete applications. We are currently developing a toolkit for merge functions allowing for profile update basing on relevance feedback. Adapting standard market basket analysis approaches to our framework will be one of the next steps.

Possible extensions

Because of its simple structure and the resulting efficiency benefits, we had decided to use RDFS/RDF for ontology modeling. However, one future task might be to extend our framework by designing and implementing a matching algorithm which defines a similarity measure for OWL instances.

References

- [1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7:39–59, 1994.
- [2] A. Bassoli and S. Baumann. Blue-tunA: music sharing through mobile phones. In *Proceedings of the Third International Workshop on Mobile Music Technology*, 2006.
- [3] S. Baumann. Smartmobs and music: Ad-hoc socializing by portable music profiles. In *Inspirational Ideas at International Computer Music Conference (ICMC 2005)*, 2005.
- [4] R. Bergmann. On the use of taxonomies for representing case features and local similarity measures. In L. Gierl and M. Lenz, editors, *Proceedings of the 6th German Workshop on CBR*. Universität Rostock, 1998. IMIB Series Vol. 7.

- [5] R. Bergmann and A. Stahl. Similarity measures for object-oriented case representations. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, 4th European Workshop, EWCBR-98, Dublin, Ireland, September 1998, Proceedings*, volume 1488 of *Lecture Notes in Computer Science*. Springer, 1998.
- [6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 89, 2001.
- [7] Patrick Brezillon. Context in Artificial Intelligence: I. A survey of the literature. *Computer and AI*, 18(4):321–340, 1999.
- [8] Patrick Brezillon. Context in Artificial Intelligence: II. Key elements of contexts. *Computer and AI*, 18(5):425–446, 1999.
- [9] A. Dey, B. Kokinov, D. Leake, and R. Turner, editors. *Modeling and Using Context, 5th International and Interdisciplinary Conference, CONTEXT 2005, Paris, France, July 5-8, 2005, Proceedings*, volume 3554 of *Lecture Notes in Computer Science*. Springer, 2005.
- [10] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848, 1965. English translation in *Soviet Physics Doklady* 10, 707–710, 1966.
- [11] H. Liu, P. Maes, and G. Davenport. Unraveling the taste fabric of social networks. *International Journal on Semantic Web and Information Systems*, 2:42–71, 2006.
- [12] F. Manola and E. Miller (Eds.). *RDF Primer*. W3C recommendation, W3C, <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>, 2004.
- [13] J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, 1971.
- [14] T. Roth-Berghofer, S. Schulz, and D. Leake, editors. *Modeling and Retrieval of Context, Second International Workshop, MRC 2005, Edinburgh, UK, July 31 - August 1, 2005, Revised Selected Papers*, volume 3946 of *Lecture Notes in Computer Science*. Springer, 2006.
- [15] Sven Schwarz. A context model for personal knowledge management. In Roth-Berghofer et al. [14].
- [16] W3C. OWL web ontology language, overview. Technical report, 2004. D. McGuinness and F. van Harmelen (Eds.), W3C Recommendation, <http://www.w3.org/TR/owl-features/>.
- [17] W3C. RDF vocabulary description language 1.0: RDF Schema. Technical report, 2004. D. Brickley, R. Guha, and B. McBride (Eds.), W3C Recommendation, <http://www.w3.org/TR/rdf-schema/>.
- [18] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

Efficient Algorithms for Mining Maximal Flexible Patterns in Texts and Sequences

Hiroki Arimura and Takeaki Uno

Graduate School of Information Science and Technology, Hokkaido University
Kita 14 Nishi 9, Sapporo 060-0814, Japan

National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract

In this paper, we study the enumeration algorithm for the maximal pattern discovery problem for the class \mathcal{ERP} of flexible sequence patterns, also known as erasing regular patterns, in a given sequence, with applications to text mining. Our notion of maximality is based on the position occurrences and weaker than the traditional notion of maximality based on the document occurrences. We present a polynomial space and polynomial delay algorithm for enumerating all maximal patterns without duplicates for the class of \mathcal{ERP} of flexible sequence patterns based on the framework of reverse search. As a corollary, the enumeration problem for maximal flexible patterns is shown to be solvable in output polynomial time. We also discuss the utility of maximal pattern discovery in document classification and a heuristic algorithm for discovering document-maximal flexible patterns in a set of input strings.

Keywords: Data mining, Enumeration algorithms, Sequence databases, Maximal pattern, Motif Discovery

1. Introduction

By the rapid growth of the amount of human-readable electronic data on networks and storages, there are potential demands for the efficient computational methods to extract useful information and knowledge from massive amount of electronic data scattered over the network. Some prominent examples of such knowledge discovery tasks are: Automatic classification of natural language texts and web pages, characteristic and descriptive pattern discovery, prediction of trends from market data, detection of malicious activities from audit data, and clustering of documents. Pattern discovery is one of the most basic technology to find a class of patterns appearing in a data set satisfying given constraints, and plays an important role in many knowledge discovery problems.

In this paper, we consider the *maximal pattern discovery problem* [9, 10, 13, 14] in a set of sequences. A pattern is *maximal* if there is no properly *more specific* pattern w.r.t. some generalization ordering over the class of patterns that has the same occurrence, or equivalently, has the same frequency, in a given set of input sequences. Since the set of all maximal patterns are typically much smaller than the set of all patterns appearing in a data set while the former contains the complete information of the latter, maximal pattern discovery has merits in efficiency and comprehensiveness. On the other hands, the computational complexity of maximal pattern discovery is higher than that of frequent pattern discovery. For example, there are few patterns classes for which the maximal pattern discovery problem have the polynomial output time complexity.

The class of patterns we consider is the class \mathcal{ERP} of *erasing regular patterns* of Shinohara [12], which is also called *flexible patterns* in bioinformatics area [9]. The class \mathcal{ERP} is a super class of the class of subsequence patterns for which polynomial output maximal pattern enumeration algorithm is known [14]. However, there is no output-polynomial algorithm for the maximal pattern problem for \mathcal{ERP} . A potential problem for the class \mathcal{ERP} of flexible patterns is, unlike classes of itemsets and rigid patterns [3, 9], there are no unique maximal pattern in each equivalence class of patterns. To overcome this problem, we introduce a weaker notion of maximality, called *position-maximality*, for \mathcal{ERP} , where two patterns P and Q are regarded as equivalent if they have the same sets of left-positions in an input string. The position-maximality is implicitly used in maximal pattern mining for subsequence patterns [14].

As a main result of this paper, under this definition of maximality, we present a polynomial-space and polynomial-delay algorithm for enumerating all maximal patterns appearing in a given string without duplicates in terms of position-maximality. This result generalizes the output-polynomial complexity of [14] for subsequence patterns to the class \mathcal{ERP} . The polynomial-space and delay property indicates that it can be used as a light-weight and high-throughput algorithm for pattern discovery. Finally, we presented a heuristics algorithm for enumerating document-maximal patterns in a collection of strings for the class \mathcal{ERP} with non-trivial pruning strategies. The motivation of this study is application of the maximal pattern discovery to the optimal pattern discovery problem in machine learning and knowledge discovery with application to text mining.

2. Preliminaries

Let A be an alphabet of symbols. We denote by A^* the set of all finite strings over A and define $\Sigma^+ = \Sigma^* - \{\varepsilon\}$. Let $s = a_1 \cdots a_n \in A^*$ be a string over A of length n . We denote

$|s| = n$ the length of s and by ε the empty string. For any indices $1 \leq i \leq j \leq n$, we denote by $s[i] = a_i$ the i -th letter of s , and by $s[i..j]$ the substring $s[i..j] = a_i \cdots a_j$ that starts with i and ends with j . For a set $S \subseteq A^*$, we denote by $|S|$ the cardinality of S and by $\|S\| = \sum_{s \in S} |s|$ be the total size of the strings in S .

For strings $u, v, w \in A^*$, we say that u , v , and w , respectively, are a *prefix*, a *substring*, and a *suffix* of a string $s = uvw$. Then, the substring v occurs in s at position $p = |u| + 1$. Equivalently, if $s = s[1] \cdots s[n]$, then v occurs in s at position i iff $v = s[i] \cdots s[i + |v| - 1]$. The *left position* and the *right position* of v corresponding to this occurrence of in t are i and $i + |v| - 1$, respectively. The *reversal* of a string $x = a_1 \cdots a_m$ is defined by $x^R = a_m \cdots a_1$.

2.1. Text and Patterns

In this subsection, we introduce the class \mathcal{ERP} of erasing regular patterns of Shinohara [12], also known as flexible patterns [9]. Let $\Sigma = \{a, b, c, \dots\}$ be a finite alphabet of *constant symbols*. We assume a special symbol $*$ $\notin \Sigma$ called a *variable* (a string wildcard or variable-length don't cares, VLDC), which represents arbitrary long possibly-empty finite string in Σ^* . Then, an *erasing regular pattern* (or *pattern*, for short) over Σ is a string $P \in (\Sigma \cup \{*\})^*$ consisting of constant symbols and variables. An erasing regular pattern P is said to be *in canonical form* if it is written as $P = w_0 * w_1 * \cdots * w_m$ for some integer $m \geq 0$ and some non-empty strings $w_0, w_1, \dots, w_m \in \Sigma^+$. Each constant string w_i is called a *segment* of P . An erasing regular pattern is also called as a *VLDC pattern* or a *flexible pattern*. We denote by \mathcal{ERP} the class of erasing regular patterns over Σ in canonical form. We note that $\Sigma^* \subseteq \mathcal{ERP}$.

Let Σ be a fixed alphabet of constants. An *input string* is a constant string $T = a_1 \cdots a_n$ ($n \geq 0$) over Σ . Let $P = w_0 * w_1 * \cdots * w_m \in (\Sigma \cup \{*\})^*$ be a pattern with $m \geq 0$ variables in canonical form. A *substitution* for P is any m -tuple $\theta = (u_1, \dots, u_m) \in$

$((\Sigma \cup \{*\})^*)^m$ of strings. Then, we define the application of θ to P , denoted by $P\theta$, as the string $P\theta = w_0u_1w_1u_2 \cdots u_mw_m \in (\Sigma \cup \{*\})^*$, where the i -th occurrence of variable $*$ is replaced with the i -th string u_i for every $i = 1, \dots, m$. The string $P\theta$ is said to be an *instance* of P by substitution θ .

For strings $P, Q \in (\Sigma \cup \{*\})^*$, P occurs in Q at position $1 \leq i \leq n$ if there exists some instance I of P that is a substring of Q starting at p , that is, $Q[i..i + |P\theta| - 1] = P\theta$ for some substitution θ for P . Particularly, the positions i and $i + |P\theta| - 1$, respectively, correspond to the left and the right ends of the occurrence of instance $P\theta$, and are called the *left* and the *right* positions of P in T (See Fig. 1). We denote by $LO(P, Q)$ and $RO(P, Q)$ the set of all left-positions and the set of all right-positions of pattern P in text Q , respectively. We refer to $LO(P, Q)$ and $RO(P, Q)$ as the *left-location list* and the *right-location list* of P in T , respectively.

Lemma 1 *Let P be a pattern and T be a text. For any position $1 \leq p \leq n$, $p \in RO_T(P)$ if and only if $n - p + 1 \in LO_{TR}(P^R)$.*

By the symmetry between the left and the right location lists in Lemma 1, it is sufficient to consider only $LO(P, T)$ than $RO(P, T)$. Thus, we only consider the *location list* $O(P, T) = LO(P, T)$ in what follows.

Lemma 2 *The location list $LO(P, T)$ is computable in $O(mn)$ time, where $m = |P|$ and $n = |T|$.*

We define a binary relation \sqsubseteq over \mathcal{P} as follows. For patterns $P, Q \in \mathcal{ERP}$, P is *more specific than* Q , denoted by $P \sqsubseteq Q$, iff P occurs in T at some position $1 \leq p \leq n - 1$. If $P \sqsubseteq Q$ and $Q \not\sqsubseteq P$ hold, then we say that P is *properly more specific than* Q and denote $P \sqsubset Q$. We note that if $P \sqsubseteq Q$ and $Q \sqsubset P$ hold then $P = Q$ holds.

Lemma 3 *Let T be any text. Then,*

$(\mathcal{ERP}, \sqsubseteq)$ is a partial ordering with the smallest element ε .

In machine learning area, the learning problem for the class formal languages defined by \mathcal{ERP} has been studied extensively [12]. For a erasing regular pattern $P \in \mathcal{ERP}$, the *language defined by* P is the set $Lang_\Sigma(P) = \{s \in \Sigma^* : P \sqsubseteq s\} \subseteq \Sigma^*$. It is easy to see that for any regular pattern $P \in (\Sigma \cup \{*\})^*$, there exists some $Q \in \mathcal{ERP}$ in canonical form such that $Lang_\Sigma(P) = Lang_\Sigma(Q)$.

2.2. Maximal patterns

Let T be a fixed text of length $n \geq 0$. The *frequency* of a pattern P in T is $|O_T(P)|$. A *minimum support threshold* is a nonnegative integer $0 \leq \sigma \leq n$. A pattern P is σ -*frequent* in T if it has the frequency no less than σ in T , i.e., $|O_T(P)| \geq \sigma$.

Definition 1 A pattern P is *maximal* in T if there is no proper specialization Q of P that has the same location list, i.e., $P \sqsubset Q$ and $O_T(P) = O_T(Q)$.

We see that a pattern $P \in \mathcal{ERP}$ is maximal iff P is a maximal element w.r.t. \sqsubseteq in the equivalence class $[P]_{\equiv_T} = \{Q \in \mathcal{ERP} : P \equiv_T Q\}$ under the equivalence relation \equiv_T defined by $P \equiv_T Q \Leftrightarrow O_T(P) = O_T(Q)$.

Lemma 4 *The maximal patterns in each equivalence class $[P]_{\equiv_T}$ is not unique in general.*

We denote by $\mathcal{F}_\sigma, \mathcal{M}$, and $\mathcal{M}_\theta = \mathcal{F}_\sigma \cap \mathcal{M}$ the classes of the σ -frequent patterns, the maximal patterns, and the maximal σ -frequent patterns in T . It is easy to see that the number of frequent flexible patterns in an input text S is exponential in the total length n of T in the worst case.

Lemma 5 *There is an infinite sequence $(T_i)_{i \geq 0}$ of texts such that the number of maximal flexible patterns in T_i is exponential in $n = |T_i|$, i.e., $|\mathcal{M}_\sigma| = 2^{\Omega(n)}$.*

Now, we state our data mining problem considered in this paper as follows.

Maximal Flexible Pattern Enumeration Problem:

Input: An alphabet Σ , a text T of length n , and a minimum support threshold σ .

Task: To generate all maximal frequent flexible patterns in $\mathcal{M}_\sigma \subseteq \mathcal{ERP}$ in T without duplicates.

It is easy to see that $\mathcal{M}_\theta = \mathcal{F}_\sigma \cap \mathcal{M}$. Therefore, we will first present an efficient enumeration algorithm for $\mathcal{M} = \mathcal{M}_1$, and then extend it for \mathcal{M}_σ for every $\sigma \geq 1$.

3. Motivated Applications of Maximal Pattern Discovery

In this section, we discuss potential applications of maximal pattern discovery considered in this paper to predictive mining and classification.

3.1. Predictive Mining and Classification

First, we introduce a practical model of machine learning from noisy environment, known as robust training or agnostic learning according to, e.g., [5, 6]. Suppose we are given a finite collection $\mathcal{S} = \{x_i : i = 1, \dots, m\} \subseteq \Sigma^*$ of strings, called a sample, and a binary labeling function $F : \mathcal{S} \rightarrow \{0, 1\}$, called an *objective function*. Each string $s \in \mathcal{S}$ is called a *document* and the value of the function $F(s) \in \{0, 1\}$ indicates whether the document is, e.g., interesting or not. Let \mathcal{P} be a class of *classification rules* or *patterns*, where each pattern $P \in \mathcal{P}$ represents a binary function $P : \mathcal{S} \rightarrow \{0, 1\}$. In our case, for any document $s \in \mathcal{S}$, $P(s) = 1$ if P matches s and $P(s) = 0$ otherwise. For a predicate $\pi(x)$, $[\pi(x)] \in \{0, 1\}$ is the indicator function that returns 1 or 0 depending on the truth value of $\pi(x)$. Now, we state our pattern discovery problem [6].

Empirical Error Minimization Problem:

Input: A sample \mathcal{S} and an objective function $F : \mathcal{S} \rightarrow \{0, 1\}$.

Task: To find an optimal pattern $P \in \mathcal{P}$ that

minimizes within \mathcal{P} the empirical error

$$\text{ERR}_{\mathcal{S},F}(P) = \sum_{x \in \mathcal{S}} [P(x) \neq F(x)].$$

In learning theory, it is known that any algorithm that efficiently solves the above empirical error minimization problem can approximate a target concept within a given concept class under arbitrary unknown probability distributions, and thus can work with noisy environments [11].

3.2. Optimal Pattern Discovery

The above framework can be extended for more general classes of score functions [4]. An *impurity function* is any real-valued function $\psi : [0, 1] \rightarrow \text{Real}$ such that (i) it takes the maximum value $\psi(1/2)$, (ii) the minimum value $\psi(0) = \psi(1) = 0$, and (iii) $\psi(x)$ is convex, i.e., $\psi(\frac{1}{2}(x + y)) \leq \frac{1}{2}(\psi(x) + \psi(y))$ for every $x, y \in [0, 1]$.

- The information entropy: $\psi_1(x) = -x \log x - (1 - x) \log(1 - x)$.
- The Gini index: $\psi_2(x) = 2x(1 - x)$.

Given objective function F , and pattern P , the *contingency table* is a 4-tuple (M_1, M_0, N_1, N_0) , where M_1 and M_0 are the numbers of the matched positive and negative examples, and N_1 and N_0 are the numbers of positive and negative examples in \mathcal{S} . Now, we describe the optimal pattern discovery problem, which is parameterized by an impurity function ψ , as follows (See, e.g., Devroy *et al.* [4]).

ψ -Optimal Pattern Discovery Problem:

Input: A sample \mathcal{S} and an objective function $F : \mathcal{S} \rightarrow \{0, 1\}$.

Task: To find an optimal pattern $P \in \mathcal{P}$ that minimizes within \mathcal{P} the cost

$$G_{\mathcal{S},F}^\psi(P) = N_1 \cdot \psi\left(\frac{M_1}{N_1}\right) + N_0 \cdot \psi\left(\frac{M_0}{N_0}\right),$$

where (M_1, M_0, N_1, N_0) is the contingency table defined by \mathcal{S} , F , and pattern P .

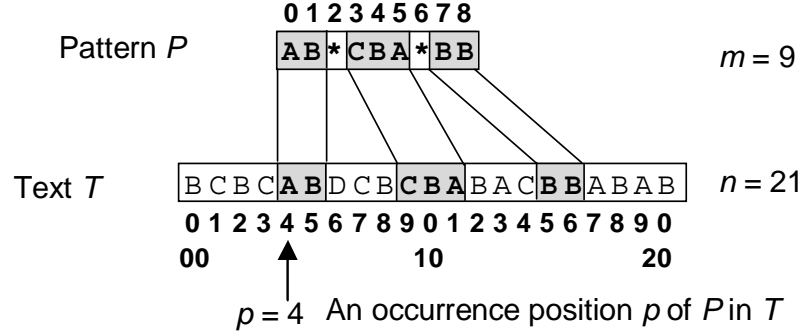


Figure 1: A pattern and its left and right positions in a text.

3.3. A Heuristic Algorithm for Optimal Pattern Discovery

In spite of the practical importance of optimal pattern problem, the above optimization problems are known to be computationally hard even for most pattern classes, e.g., half-spaces and conjunctions [5]. In particular for the class \mathcal{ERP} of flexible patterns, Miyano, Shinohara, and Shinohara [8] showed that the consistency problem for \mathcal{ERP} is NP-complete. Shimozono, Arimura, and Arikawa [11] showed that the empirical error minimization problem for \mathcal{ERP} is even hard to approximation within arbitrary small ratio.¹ Therefore, a class of straightforward generate-and-test algorithms, which enumerates frequent patterns with an adequate threshold, are used to solve the optimized pattern discovery problem in practice.

In Fig. 2, we show a heuristic algorithm `FINDOPTIMAL` for discovering top- K optimal patterns in terms of the score function $G_{S,F}^\psi(P)$ based on a generate-and-test strategy using maximal patterns in \mathcal{M} instead of frequent patterns in \mathcal{F} . If we set $K = 1$ then the algorithm solves the above ψ -optimal pattern problem. Then, the following theorem gives a justification of such an approach.

¹Although the original approximation hardness result in [11] has been shown for the class with proximity constraints, we can obtain the same result for \mathcal{ERP} without proximity constraints from its proof.

Theorem 6 Let $\mathcal{S} \subseteq \Sigma^*$ and $F : \mathcal{S} \rightarrow \{0, 1\}$. Let $\mathcal{P} = \mathcal{ERP}$ and $\mathcal{M} \subseteq \mathcal{P}$ be the class of maximal patterns in \mathcal{S} . Even if we restrict the class of patterns to \mathcal{M} in \mathcal{S} , then this does not lose the optimality of the answers for the empirical error minimization problem and the ψ -optimal pattern discovery problem. That is, $\min\{\text{ERR}_{\mathcal{S},F}(P) : P \in \mathcal{M}\} = \min\{\text{ERR}_{\mathcal{S},F}(P) : P \in \mathcal{P}\}$ and $\min\{G_{\mathcal{S},F}^\psi(P) : P \in \mathcal{M}\} = \min\{G_{\mathcal{S},F}^\psi(P) : P \in \mathcal{P}\}$ hold.

Proof: For any pattern P , the cost $G_{\mathcal{S},F}(P)$ is uniquely determined by the contingency table (M_1, M_0, N_1, N_0) corresponding to a sample \mathcal{S} , an objective function F , and pattern P . For any patterns P, Q , if $LO(P, \mathcal{S}) = LO(Q, \mathcal{S})$ then P and Q realizes the same classification function $P : \mathcal{S} \rightarrow \{0, 1\}$, and thus gives the same contingency table. Hence, the follows. \square

From Theorem 6, we know that the algorithm `FINDOPTMAX` correctly finds top- K optimal patterns in \mathcal{S} minimizing the cost $G_{\mathcal{S},F}(P)$. The efficiency of this heuristics algorithm heavily depends on enumeration of maximal patterns of \mathcal{M} at Line 2. Therefore, our goal in the remainder of this paper is to develop a memory and time efficient enumeration algorithm for maximal patterns in \mathcal{M} for the class \mathcal{ERP} . In what follows, we refer to as the delay of an enumeration algorithm \mathcal{A} the maximum computation time

Algorithm FINDOPTMAX(\mathcal{S}, F)

input: A sample \mathcal{S} , an objective function $F : \mathcal{S} \rightarrow \{0, 1\}$, and an integer $K \geq 1$.

output: The top- K optimal patterns P_1, \dots, P_K minimizing the cost $G_{\mathcal{S}, F}(P)$.

- 1 Let $\mathcal{Q} := \emptyset$ be an empty priority queue with the cost as the key.
 - 2 **foreach** maximal pattern $P \in \mathcal{M}$ in \mathcal{S} **do begin:**
 - 3 Let $LO(P, \mathcal{S})$ be the location list of P ;
 - 4 Compute the contingency table $\tau = (M_1, M_0, N_1, N_0)$ from $LO(P, \mathcal{S})$;
 - 5 $\mathcal{Q} := \mathcal{Q} \cup \{(P, G_{\mathcal{S}, F}^\psi(P))\}$;
 - 6 **end**
 - 7 **output** the top K patterns P in \mathcal{Q} in terms of $G_{\mathcal{S}, F}(P)$;
-

Figure 2: An algorithm for the optimal pattern discovery via maximal pattern enumeration.

of \mathcal{A} between consecutive outputs.

4. Tree-shaped Search Route for Position-Maximal Flexible Patterns

In this section, we will present our algorithm for finding position-maximal patterns in a given input string for the class \mathcal{ERP} of flexible patterns. First, we introduce a tree-shaped search route \mathcal{T} for the space of all maximal patterns in \mathcal{M} . Then, in the next section, we give a memory efficient algorithm for enumerating all maximal patterns based on the depth-first search over \mathcal{T} . Our strategy is explained as follows: First we define a binary relation between maximal patterns, called the *parent function*, which indicates a reverse edge from a child to its parent. Next, we reverse the direction of the edges to obtain a spanning tree for \mathcal{M} .

We start with technical lemmas. Let $P, Q \in \mathcal{ERP}$ be patterns over Σ . Recall that Q is a specialization of P , denoted by $P \sqsubseteq Q$, iff $Q = \alpha(P\theta)\beta$ for some $\alpha, \beta \in (\Sigma \cup \{*\})^*$ and for some substitution θ for P . We distinguish two cases whether $\alpha = \varepsilon$.

Definition 2 A pattern Q is said to be a *prefix specialization* of another pattern P if P occurs in the initial part of Q , i.e., $Q = (P\theta)\beta$ for some string $\beta \in (\Sigma \cup \{*\})^*$ and for some substitution θ for P . If there is no such β and θ , Q is said to be a *non-prefix specialization* of P .

The following two lemmas are essential for flexible patterns.

Lemma 7 Let $T \in \Sigma^*$ be an input string and $P, Q \in \mathcal{ERP}$ be flexible patterns. Suppose that $P \sqsubseteq Q$. Then, if Q is a prefix specialization of P then $O(P, T) \supseteq O(Q, T)$.

Proof: Let T be an input string of length n . Let $p \in O(P, T) = LO(P, T)$ be any left-position of P in T . Then, it follows from the definition that if $p \in LO(P, T)$ then some substring H of T starting at position p is an instance of P . On the other hand, since Q is a prefix specialization of P , some prefix H' of Q is an instance of P . Therefore, some prefix H'' of H , and thus a substring of T , is an instance of P . Since H' is a substring starting at p in T , the lemma is proved. \square

Lemma 8 Let $T \in \Sigma^*$ be an input string and $P, Q \in \mathcal{ERP}$ be flexible patterns. Suppose that $P \sqsubseteq Q$. Then, if Q is a non-prefix specialization of P then $O(P, T) \not\subseteq O(Q, T)$.

Proof: Let $p_{\max} = \max LO(P, T)$ be the largest left-position of P in T . Since $LO(P, T) \neq \emptyset$ and T has finite length, there always exists such a largest left-position p_{\max} in T . Now we assume to contradict that $O(P, T) \subseteq O(Q, T)$. Then, p_{\max} is also a position of Q in T . Since Q is a non-prefix specialization of P , P occurs in Q at some position $\delta > 1$. Thus, we know that

$q = p_{\max} + \delta - 1$ is a position of P in T . If $\delta > 1$ then q is strictly larger than p_{\max} . This contradicts the assumption that p_{\max} is the largest position of P in T , and thus we conclude that $O(P, T) \not\subseteq O(Q, T)$. Hence, the result is proved. \square

Corollary 9 *Let $T \in \Sigma^*$ be an input string and $P, Q \in \mathcal{ERP}$ be flexible patterns. Suppose that $P \sqsubseteq Q$. Then, if Q is a non-prefix specialization of P then $O(P, T) \neq O(Q, T)$.*

We define the parent-child relationship between two flexible patterns in \mathcal{ERP} as follows.

Definition 3 Let $Q = w_0 * w_1 * \dots * w_m$ be a flexible pattern over Σ , where $m \geq 0$ and $w_1, \dots, w_m \in \Sigma^+$. Then, we define the *parent* of Q , denoted by $\mathcal{P}(Q)$, is the pattern satisfying the the followings (i) or (ii):

- (i) If $|w_0| \geq 2$, that is, $w_0 = au_0$ for some letter $a \in \Sigma$ and a non-empty string $u_0 \in \Sigma^+$, then then $\mathcal{P}(Q) = u_0 * w_1 * \dots * w_m$.
- (ii) If $|w_0| = 1$, that is, $w_0 = a$ for some letter $a \in \Sigma$ then $\mathcal{P}(Q) = w_1 * \dots * w_m$.

In summary, the parent $\mathcal{P}(Q)$ is the flexible pattern obtained from Q by removing the first letter, and then remove the first variable $*$ at the starting position if it exists. The removal of the initial $*$ ensures the canonicity of the resulting pattern.

Lemma 10 *For any non-empty flexible pattern $Q \in \mathcal{ERP}$ in canonical form, its parent $\mathcal{P}(Q)$ is always defined, unique, and a flexible pattern in canonical form, i.e., a member of \mathcal{ERP} .*

Let $n \geq 0$ be a positive integer. For a nonnegative integer $0 \leq k \leq n$ and a set $X \subseteq \{1, \dots, n\}$, we define $X + k = \{x + k : x \in X\}$. For sets $X, Y \subseteq \{1, \dots, n\}$, we define $X \bowtie_{\leq} Y = \{p \in X : p \leq q \text{ for some } q \in Y\}$. By definition, both of

$X + k$ and $X \bowtie_{\leq} Y$ are subsets of X . Using these operators, we can describe the location lists of a composite pattern of the form wP or $w * P$, where $w \in \Sigma^+$ and $P \in \mathcal{ERP}$ as follows.

Lemma 11 *Let $T \in \Sigma^*$ be any input string, $w \in \Sigma^+$ be any non-empty constant string, and $P \in \mathcal{ERP}$ be any flexible pattern. Then, the following (a) and (b) hold:*

- (a) $LO(wP) = LO(w, T) \cap (LO(P, T) - |w|)$.
- (b) $LO(w * P) = LO(w, T) \bowtie_{\leq} (LO(P, T) - |w|)$.

Let $T \in \Sigma^*$ be any input string of length $n \geq 0$. A maximal pattern P is a *root pattern* in T if $O(P, T) = \{1, \dots, |T|\}$. Now, we show the main result of this section.

Theorem 12 (reverse search property of \mathcal{M})

Let $Q \in \mathcal{M}$ be a maximal pattern in T that is not a root pattern. Then, $\mathcal{P}(Q)$ is also a maximal pattern in T . That is, if $Q \in \mathcal{M}$ then $\mathcal{P}(Q) \in \mathcal{M}$ holds. Furthermore, $|\mathcal{P}(Q)| < |Q|$ holds.

Proof: Let Q be a maximal pattern that is not a root pattern, and let $P = \mathcal{P}(Q)$ be the parent of Q . In what follows, for any pattern R , we write $LO(R)$ for $LO(R, T)$ by omitting T for simplicity. Suppose to contradict that P is not maximal in T . Then, there exists some proper specialization P' of P , i.e., $P \sqsubset P'$, such that $LO(P) = LO(P')$.

If $P \sqsubset P'$ then P occurs in P' at some position, say, $1 \leq p \leq |P'|$. There are two cases below.

(i) The case where $p \neq 1$: Then, P' is a non-prefix specialization of P . It immediately follows from Lemma 9 that $LO(P) \neq LO(P')$. This is a contradiction.

(ii) The case where $p = 1$: Then, P' is a prefix specialization of P . By the definition of the parent, there are the following cases for P and Q .

(ii.a) The case where $Q = aP$ for some letter $a \in \Sigma$: Let $Q' = aP' \in \mathcal{ERP}$.

Since P' is a proper prefix specialization of P , Q' is also a proper specialization of Q , i.e., $Q \sqsubset Q'$. In this case, we have $LO(Q') = LO(aP') = LO(a) \cap (LO(P') - |a|)$ by Property (a) of Lemma 11. Since $LO(P') = LO(P)$ by the assumption, it is obvious that $LO(a) \cap (LO(P') - |a|) = LO(a) \cap (LO(P) - |a|)$. Again by applying Property (a) of Lemma 11 to the right hand side, we have $LO(a) \cap (LO(P) - |a|) = LO(aP)$. Thus, we have $Q \sqsubset Q'$ and $LO(Q') = LO(Q)$, which says that Q is not maximal in T . However, this contradicts the assumption.

(ii.b) The case where $Q = a*P$ for some letter $a \in \Sigma$: Let $Q' = a*P' \in \mathcal{ERP}$. Since P' is a proper specialization of P (in this case, P' is not necessarily prefix-specialization), we have $Q \sqsubset Q'$. Furthermore, since $LO(P') = LO(P)$ by assumption, we can also show that $LO(Q') = LO(Q)$ by applying Property (b) of Lemma 11 as in the proof for case (ii.a). Therefore, we see that Q is not maximal in T , and thus, the contradiction is derived.

By combining cases (i), (ii.a), and (ii.b) above, we conclude by contradiction that P is maximal in T . Furthermore, it is clear from the construction that P is strictly shorter than Q in length. Hence, the result is proved. \square

Definition 4 A search graph for \mathcal{M} w.r.t. \mathcal{P} is a directed graph $\mathcal{T} = (\mathcal{M}, \mathcal{P}, \mathcal{I})$ with roots, where \mathcal{M} is the set of nodes, i.e., the set of all maximal flexible patterns in T , \mathcal{P} is the set of reverse edge such that $(P, Q) \in \mathcal{P}$ iff $P = \mathcal{P}(Q)$ holds, and $\mathcal{I} \subseteq \mathcal{M}$ is the set of root patterns in T .

Since each non-root node has the unique parent in \mathcal{T} from Theorem 12, the search graph \mathcal{T} is actually a directed tree with reverse edges. Therefore, we have the following corollary.

Corollary 13 Let T be any input string. Then, $\mathcal{T} = (\mathcal{M}, \mathcal{E}, \mathcal{I})$ is a spanning forest for \mathcal{M} with the root set \mathcal{I} .

5. An Algorithm for Position-Maximal Flexible Pattern Enumeration

In Fig. 3, we show a polynomial-space and polynomial-delay enumeration algorithm POSMAXFLEXMOTIF for maximal flexible patterns. Given an input string T of length n , this algorithm enumerates all position-maximal patterns in T without duplicates in polynomial time per maximal pattern using polynomial space in the input size n using depth-first search over the search tree \mathcal{T} over \mathcal{M} based on Corollary 13,

Recall that the search tree \mathcal{T} for \mathcal{M} has reverse edges only, that is, each edge of \mathcal{T} is directed from a child to its parent. Therefore, the first step is to compute all children, given a parent pattern $P \in \mathcal{M}$. This can be done as follows.

Lemma 14 For any maximal flexible patterns $P, Q \in \mathcal{M}$, $P = \mathcal{P}(Q)$ if and only if there exists some constant letter $a \in \Sigma$ such that either (i) $Q = aP$ or (ii) $Q = a*P$ holds.

Furthermore, since $\mathcal{P}(Q)$ is defined also for non-maximal flexible patterns Q , we know that any flexible pattern can be obtained from finite applications of the operations in above Lemma 14 to the empty pattern ε . Then, Theorem 12 gives a sound pruning strategy that once an enumerated pattern P gets non-maximal then we can immediately prune all the descendants of P .

Secondly, we discuss how to efficiently test the maximality of a given pattern P . The refinement operator for \mathcal{ERP} was introduced by Shinohara [12]. The following version is due to [2].

Definition 5 A basic refinement of a pattern P is any pattern Q obtained from P by applying one of the following operations (r1) and (r2):

- (r1) Q is obtained by replacing some segment $w \in \Sigma^+$ in P with either aw , wa , $a*w$, $w*a$ for some $a \in \Sigma$.

Algorithm POSMAXFLEXMOTIF($\Sigma, \mathcal{S}, \sigma$)

input: An alphabet Σ , an input string \mathcal{S} ,
minimum frequency threshold $0 \leq \sigma \leq |\mathcal{S}|$;

output: All maximal patterns in \mathcal{M} ;

- 1 Let \perp be the maximal pattern in T which equivalent to ε (the root pattern).
- 2 ENUMMAXIMAL(ε, σ);

Procedure ENUMMAXIMAL($P, LO(P), \sigma$)

input: A maximal pattern P and its left-location list $LO(P)$.

output: All maximal patterns that are descendants of P .

- 1 Compute $LO(P)$;
 - 2 **if** $|LO(P)| = 0$ **then return**;
 - 3 **if** P is not maximal in \mathcal{S} **then return**;
 - 4 Output P ;
 - 5 **foreach** $a \in \Sigma$ **do begin**:
 - 6 ENUMMAXIMAL($aP, LO(aP), \sigma$);
 - 7 ENUMMAXIMAL($a * P, LO(a * P), \sigma$);
 - 7 **end**
-

Figure 3: An algorithm POSMAXFLEXMOTIF for enumerating all maximal flexible patterns in an input sequence.

(r2) Q is obtained by replacing a pair of consecutive segments $v * w \in \Sigma^+ \{*\} \Sigma^+$ in P with vw .

For a pattern P , we define $\rho(P) \subseteq \mathcal{ERP}$ to be the set of all basic refinements of P .

Lemma 15 *A flexible pattern P is maximal in T if and only if there is no basic refinement $Q \in \rho(P)$ such that $LO(P, T) = LO(Q, T)$.*

Corollary 16 *The maximality of a flexible pattern P in an input string T is decidable in $O(|\Sigma|m^2n)$ time, where $m = |P|$ and $n = |T|$.*

On the shape of \mathcal{T} , we have the following lemma.

Lemma 17 *Let $P \in \mathcal{M}$ be any maximal pattern in \mathcal{T} and $m = |P|$. Then,*

- (i) *The depth of P in \mathcal{T} (the length of the unique path from the root to P) is at most m .*

(ii) *The branching of P in \mathcal{T} (the number of the children for P) is at most $O(|\Sigma|m)$.*

By combining the above lemmas, we have the main result of this paper. This says that our algorithm POSMAXFLEXMOTIF is a memory and time efficient algorithm for discovering maximal flexible patterns.

Theorem 18 *Let Σ be an alphabet and $T \in \Sigma^*$ be an input string of length $n \geq 0$. Then, the algorithm POSMAXFLEXMOTIF in Fig. 3 enumerates all position-maximal patterns P in T without duplicates in $O(|\Sigma|kmn^2)$ delay per maximal pattern using $O(mn)$ space, where $m = |P|$ and $k = O(m)$ are the size and the number of variables of the pattern P to be enumerated.*

Corollary 19 *The maximal pattern enumeration problem for the class \mathcal{ERP} of flexible patterns (or erasing regular patterns) w.r.t. position-maximality is solvable in polynomial-space and polynomial-delay in the total input size.*

6. A Practical Algorithm for Discovering document-maximal flexible patterns in a set of input strings

Let Σ be a fixed alphabet of constants. An input string set is a collection of constant strings over Σ

$$S = \{s_1, \dots, s_d\} \subseteq \Sigma^*,$$

where each member $s_i \in \Sigma^*$ is called a *document* of S ($i = 1, \dots, d$). For a flexible pattern $P \in \mathcal{ERP}$, the *document-location list* of P is defined by the set $DO(P, S) = \{1 \leq i \leq d : P \sqsubseteq s_i\}$ of the indices of the documents in which P occurs. The *document frequency* of P in S is defined by $|DO(P, S)|$.

For $0 \leq \sigma \leq |S|$, a pattern P is *document σ -frequent* if $|DO(P, S)| \geq \sigma$. A pattern P is *document-maximal* in S if there is no proper specialization Q of P that has the same document-location list in S , i.e., $P \sqsubset Q$ and $DO(P, S) = DO(Q, S)$. The following lemma justifies the pruning strategy at Line 2 of Algorithm DOCMAXFLEXMOTIF in Fig. 4.

Lemma 20 (pruning by monotonicity) *If $P \sqsubseteq Q$ then $DO(P, S) \supseteq DO(Q, S)$.*

Let $S = \{s_1, \dots, s_k\} \subseteq \Sigma^*$ be an input document set and let $\# \notin \Sigma$ be a new delimiter symbol. Then, we define an input string $S = s_1\# \dots \#s_k$ obtained from S by concatenating all documents by the delimiter $\#$. The following lemma ensure the soundness of the pruning strategy at Line 3 of Algorithm DOCMAXFLEXMOTIF in Fig. 4.

Lemma 21 (pruning by position-maximality) *Let P be any flexible pattern over Σ . If P is document-maximal in S then P is also a position-maximal in S .*

Based on the above lemmas, we show the algorithm DOCMAXFLEXMOTIF for enumerating all document-maximal frequent flexible patterns in a set of strings in Fig. 4. Unfortunately, this algorithm is not shown to be of output-polynomial time.

Theorem 22 *Let Σ be an alphabet and $T \in \Sigma^*$ be an input string of length $n \geq 0$. Then, the algorithm DOCMAXFLEXMOTIF in Fig. 4 enumerates all document-maximal patterns P in T without duplicates using $O(mn)$ space, where $m = |P|$ is the size of the pattern P to be enumerated.*

7. Conclusion

In this paper, we consider the maximal pattern discovery problem for the class \mathcal{ERP} of flexible patterns [9], which is also known as erasing regular patterns in machine learning. The motivation of this study is applications to the optimal pattern discovery problem in machine learning and knowledge discovery. As a main result we present a polynomial-space and polynomial-delay algorithm for enumerating all maximal patterns appearing in a given string without duplicates in terms of position-maximality defined through the equivalence relation between the location lists. As another application, we presented a heuristics algorithm for enumerating document-maximal patterns in a collection of strings for the class \mathcal{ERP} with non-trivial pruning strategies.

References

- [1] D. Avis and K. Fukuda, Reverse Search for Enumeration, *Discrete Applied Mathematics*, Vol. 65, 21–46, 1996.
- [2] H. Arimura, R. Fujino, T. Shinozaki, Protein motif discovery from positive examples by minimal multiple generalization over regular patterns, Proc. GIW'94, 39-48, Dec. 1994.
- [3] H. Arimura, T. Uno, A polynomial space and polynomial delay algorithm for enumeration of maximal motifs in a sequence, Proc. ISAAC'05, LNCS 3827, Springer, Dec. 2005.
- [4] L. Devroy, L. Gyrfi and G. Lugosi, *A Probabilistic Theory of pattern Recognition*, Springer Verlag, 1996.
- [5] M. J. Kearns, R. E. Shapire, L. M. Sellie, Toward efficient agnostic learning, *Machine Learning*, 17(2–3), 115–141, 1994.
- [6] W. Maass, Efficient agnostic PAC-learning with simple hypothesis, In Proc. COLT94, 67–75, 1994.

Algorithm DOCMAXFLEXMOTIF($\Sigma, \mathcal{S}, \sigma$)

input: An alphabet Σ , an input set $\mathcal{S} = \{s_1, \dots, s_k\} \subseteq \Sigma^*$, $0 \leq \sigma \leq |\mathcal{S}|$;

output: All document-maximal patterns in \mathcal{DM} ;

- 1 Let \perp be the maximal pattern in T which equivalent to ε (the root pattern).
- 2 Let $S = s_1\# \cdots \#s_k$ be an input string ($\# \notin \Sigma$).
- 3 DOCENUMMAXIMAL($\perp, LO(\perp, S), \sigma$);

Procedure DOCENUMMAXIMAL($P, LO(P), \sigma$)

input: A maximal pattern P and its left-location list $LO(P, S)$.

output: All maximal patterns that are descendants of P .

- 1 Compute $DO(P, \mathcal{S})$ from $LO(P, S)$;
 - 2 **if** $|DO(P, \mathcal{S})| < \sigma$ **then return**;
 - 3 **if** P is not position-maximal in \mathcal{S} w.r.t. $LO(P, S)$ **then return**;
 - 4 **if** P is document-maximal in \mathcal{S} **then output** P ;
 - 5 **foreach** $a \in \Sigma$ **do begin**:
 - 6 DOCENUMMAXIMAL($aP, LO(aP, S), \sigma$);
 - 7 DOCENUMMAXIMAL($a*P, LO(a*P, S), \sigma$);
 - 7 **end**
-

Figure 4: An algorithm POSMAXFLEXMOTIF for enumerating all maximal flexible patterns in an input sequence.

- [7] H. Mannila, H. Toivonen, A. I. Verkamo, Discovery of frequent episodes in event sequences, *Data Min. Knowl. Discov.*, 1(3), 259–289, 1997.
- [8] S. Miyano, A. Shinohara, T. Shinohara, Polynomial-time learning of elementary formal systems, *New Generation Comput.* 18(3): 217–242, 2000.
- [9] L. Parida, I. Rigoutsos, *et al.*, Pattern discovery on character sets and real-valued data: linear-bound on irredundant motifs and efficient polynomial time algorithms, In *Proc. SODA'00*, 2000.
- [10] N. Pisanti, M. Crochemore, R. Gross, M.-F. Sagot, A basis of tiling motifs for generating repeated patterns and its complexity of higher quorum, In *Proc. MFCS'03*, 2003.
- [11] S. Shimozone, H. Arimura, S. Arikawa, Efficient discovery of optimal word-association patterns in large text databases, *New Generation Comput.* 18(1), 49–60, 2000.
- [12] T. Shinohara, Polynomial time inference of extended regular pattern Languages. *Proc. RIMS Symp. on Software Sci. & Eng.*, 115–127, 1982.
- [13] X. Yan and J. Han, R. Afshar, CloSpan: mining closed sequential patterns in large databases, In *Proc. SDM 2003*, SIAM, 2003.
- [14] J. Wang and J. Han, BIDE: efficient mining of frequent closed sequences, In *Proc. IEEE ICDE'04*, 2004.

Application of Truncated Suffix Trees to Finding Sentences from the Internet

Takuya Kida Takashi Uemura Hiroki Arimura
Hokkaido University
Graduate School of Information Science and Technology
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido,
060-0814, Japan

Abstract

We propose a space saving index structure based on the suffix tree, called the truncated suffix tree with k -words limitation, which is a tree structure obtained by eliminating the nodes that indicate the substrings longer than k words in a given text. Although such tree may confuse any substrings that occur at the different positions but the same string, it is suitable for counting short substrings. We presented it can be constructed on-line in $O(n)$ time and space. Our experimental result shows that the number of nodes falls broadly when k is small. The reduction ratio is almost half in comparison with the original suffix tree when $k = 3$ in practice. We also discuss about the application of it to Kiwi system which can help us to find keywords from the Internet.

1. Introduction

Since information created by people has increased rapidly after moving into the 21st century, efficient technologies for dealing with numerous data become more and more important. There are many documents and databases on the Internet now, and it is urgently necessary to develop information retrieval technologies that one can pick up the desired information.

Nowadays almost everyone uses search engines like Google or Yahoo to get information which one wants to know. Such search engines may display search results quickly when a user throws appropriate queries into them. However the user would not be able to obtain sufficient results if proper keywords are not come to mind. Therefore, to give users some clues for proper queries, or to extract semi-automatically the desired information from the Internet, many researchers have studied about *the web data mining*.

Kiwi system[13] developed by Tanaka and Nakagawa[14] is one of the solutions to help such situation. It is originally developed with the aim of presenting the users some

phrase examples that match into the inputted context. Though it is based on the same idea of the KWIC (Key Word in Context) tool, it can present phrases that match the user's needs better by using data on the Internet rather than a fixed dictionary. There are some similar systems like WebCorp[1], Google Fight[2], Google Duel[10], and so on, but Kiwi has a superior feature that can handle any languages and flexible queries.

The current Kiwi system gathers the search results for the temporal dictionary to be used by touching search engine's API for each query request. However, such mechanism may take several minutes before replying the final results. Although the response time can be reduced if the system gathers less search results, the precision of the final results will drop down. Since users usually can not stand for waiting more than several tens of seconds, the text data to be used must be in a local server, and the system must be able to retrieve keywords quickly from them, to guarantee the sufficient precision and response speed.

Suffix tree[5] is a useful data structure that can index every substring in a given text.

The size of the tree for the text is $O(n)$, where n is the text length. By using the suffix tree, searching a keyword can be done in $O(m)$ time, where m is the keyword length[16]. Several well-known optimal algorithms for constructing suffix trees exist[8, 15], and various extensions have been proposed.

Admitting that the suffix tree is a compact data structure which can deal with all substrings in $O(n)$ space as mentioned above, the required memory space tends to grow to an enormous size for practical purposes because the size n of the target data itself is often large. Therefore, more compact data structures that have almost the same function as the suffix tree, is desirable. From the viewpoint of application of the suffix tree, there are few cases that very long substrings are needed. For the purpose of searching short sentences, for example, it is sufficient to index only substrings that cover several words.

In order to speed up the response time, Ichii *et al.*¹ improved Kiwi system by using a *suffix tree for n-gram* that indexes all n-grams of the text data which are gathered from the Internet in advance and stored locally. Their proposed data structure is fundamentally the same as that of Na *et al.*[9]. They succeeded to shorten the response time dramatically. However, it can not retrieve the sentences longer than the fixed size n .

In this paper, we propose a new data structure which can index any substring shorter than k words in a given text, called the *truncated suffix tree with K-words limitation* (k -WST for short). We also present it can be constructed in $O(n)$ time. Assume that the input text T is split into the N words sequence with delimiter symbol #, that is, $T = w_1\#w_2\#\dots\#w_N$. Then, for T and a given integer k , k -WST represents all substrings of the string $w_i\#w_{i+1}\#\dots\#w_{i+k-1}$ for every i ($1 \leq i \leq N - k + 1$). The proposed algorithm is based on the online algorithm of Ukkonen[15]. k -WST can be con-

structed in $O(n)$ time and space for the input text of length n . The basic idea of our algorithm is to close partially the extensions of leaves in the suffix tree, which are automatically extended in the original Ukkonen's algorithm[15], whenever each delimiter is loaded. By applying k -WST to Kiwi system, there exists a possibility to make it more flexible and useful.

Related works Implementing data compression methods such as Ziv-Lempel family is one of the most important applications of suffix trees. The original suffix tree is not applicable to the LZ77 scheme because that the string can be referred as a dictionary is restricted by the sliding window. Several methods which can overcome this problem have been proposed[11, 4, 7]. Na *et al.*[9] defined the *truncated suffix tree* which can represent all substrings whose length is less than k , and proposed the constructing algorithm in proportion to the text length. Although our algorithm can be seen as an extension of their idea, closing leaves' extension is done at different time intervals for each leaves since the length of the suffixes registered with the tree is not fixed.

For Kiwi system, by adopting the several demands of the field of linguistics, the newly system, called *Tonguen*, has developed by Tanaka and Ishii[12].

2. Preliminaries

Let Σ be a finite alphabet and let Σ^* be the set of all strings over Σ . We denote the length of string $x \in \Sigma^*$ by $|x|$. The string whose length is 0 is denoted by ε , called *empty string*, that is $|\varepsilon| = 0$. We also define that $\Sigma^+ = \Sigma^* - \{\varepsilon\}$. The concatenation of two strings x_1 and $x_2 \in \Sigma^*$ is denoted by $x_1 \cdot x_2$, and also write it simply as x_1x_2 if no confusion occurs.

Strings x , y , and z are said to be a *prefix*, *factor*, and *suffix* of the string $w = zyz$, respectively. The i th symbol of a string w is denoted by $w[i]$, and the factor of u that begins at position i and ends at position j is denoted by $w[i \dots j]$. For convenience, let

¹This paper was published in the workshop proceeding written in Japanese

$w[i \dots j] = \varepsilon$ for $j < i$. We also denote by $Suf(w)$ the set of all suffixes of the string $w \in \Sigma^*$, and denote by $Fac(w)$ the set of all factors of $w \in \Sigma^*$.

2.1. Kiwi system

We will make a brief sketch of the processing of Kiwi system according to [13]. Kiwi system was developed as a consultation tool. It answers what kind of strings preceded/succeeded by the phrase which an user inputs, by using some search engines on the Internet. An input query phrase can include a meta character '*', for example, 'super *', '*-like man', 'ABC of * for', and so on. The system extracts the candidate strings that match at the place of '*' from the search results. The outline of the processing is as follows:

1. Receive a query from a user.
2. Tokenize the query and then send it to search engines.
3. Extract all fixed-length strings from the search results, which include the candidates.
4. Cut these strings to the proper length to extract candidates.
5. Rank them before answer.

The clipping positions can be straightforwardly determined when the '*' is used in the middle. In the case that a query is preceded or succeeded by '*', it is a problem to determine how long we must take as a candidate string. Cutting at each word separator for segmented languages, or cutting by fixed-length for non-segmented languages, are in common use for the other KWIC systems, while Kiwi system takes another approach.

Now we concentrate in the case that the input query ends with '*', since the case that it starts with '*' can be managed as the same way by viewing the target texts in reverse.

We call a string obtained by eliminating '*' from a query as a *clue string*. What we want

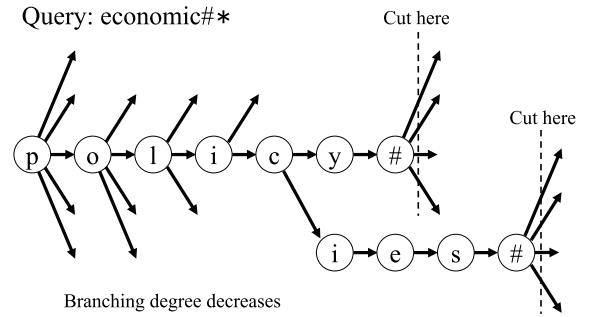


Figure 1: Method for clipping candidates

to do is to cut out the candidate from the string x which has the clue string P as a prefix. The candidates we expect must have the following natures:

- (i) It occurs frequently in the text.
- (ii) It has the moderate length.
- (iii) It is succeeded by various kind of words (namely characters).

Let denote by X_i the prefix $x[1 \dots i]$ ($i > |P|$) of x , and also denote by C_i the number of kinds of character $x[i+1]$ preceded by X_i in the target text. From the condition (iii), the Kiwi system extracts X_i as a candidate when

$$C_i > C_{i-1}.$$

In order to do this, an n-gram trie for the text is constructed and traversed. That is, the system determines the candidates by cutting at the position where the number of branches in the trie turns from downward to upward in the traversing (see Figure 1). Such candidate strings often satisfy the condition (ii).

This method is empirically and do not necessarily go well at any time. Actually, the additional efforts are required because it may fail when the candidates hardly appear in the text. However, it has an advantage that it is independent of language.

In the step of ranking, Kiwi system uses the formula

$$F(X) = freq(X) \log(|X| + 1)$$

as an evaluation function, where $freq(X)$ is the occurrence rate of the candidate X . This is to think that $freq(X)$ is important for the condition (i); on the other hand longer candidates are more important for the condition (ii). Although we can consider more complicated evaluation functions, it will be a trade-off between the precision and the performance.

2.2. Suffix tree

Suffix tree (ST for short) is a data structure that can represent $Fac(T)$ for the string T . It can be denoted by a quadruplet as $ST(T) = (V \cup \{\perp\}, root, E, suf)$, where V is a set of nodes and $\perp \notin V$, and E is a set of edges and $E \subseteq V^2$. Then, the graph (V, E) forms a rooted tree which have a root node $root \in V$ for $ST(T)$. That is, there exists just one path from $root$ to each node $s \in V$. If $(s, t) \in E$ for nodes $s, t \in V$, we call s the *parent* of t and t by the *child* of s . We also call the *internal node* of T if it has a child, and call the *leaf* of T if it has no children. Moreover, every node on the path from $root$ to the node $s \in V$ is called the *ancestor* of s .

Each edge in E is labeled with a factor of T , represented by a pair of integers (j, k) , that is, the *label string* is $T[j \dots k]$. We denote by $label(s)$ the label string for the edge into $s \in V$ since any child has just one parent. For any $s \in V$, we denote by \bar{s} the string spelled out every ancestor's label from $root$ to s , and call it the *string represented by s* . That is, assuming that $root, a_1, a_2, \dots, s$ is the sequence of the ancestor nodes of s , $\bar{s} = label(root) \cdot label(a_1) \cdot label(a_2) \cdots label(s)$.

For a given text T , each leaf of $ST(T\$)$ corresponds one-to-one with each suffix of T , and any internal node except for $root$ has more than two children, where we assume that $\$$, called *terminal symbol*, is not included in Σ . Then, there are $|T|$ leaves and less than $2|T| - 1$ nodes in $ST(T\$)$. For any two children x, y of each $s \in V$, \bar{x} starts a different character from \bar{y} , that is, $\bar{x}[1] \neq \bar{y}[1]$. Note that each node $s \in V$ corresponds one-to-one with a factor of T from the above. We treat each factor that has

no corresponding node in $ST(T)$ as a *virtual node*. To distinguish it, we call each $s \in V$ a *real node*.

We define $suf : V \rightarrow V$ as a function which returns a node representing the string eliminating the first character from \bar{s} for $s \in V$. For each real node $s, t \in V$ and a character $a \in \Sigma$, $suf(s) = t$ iff $\bar{s} = a \cdot \bar{t}$. Since this can be seen as an edge (s, t) from s to t , we call it the *suffix link* from s to t . For each real node $s \in V$, there exists just one path from s to $root$ by traversing suffix links. We call the path the *suffix path*.

2.3. Ukkonen's suffix tree construction

We will give a brief sketch of Ukkonen's suffix tree construction algorithm[15] below. The algorithm reads characters one by one from the given text $T = T[1 \dots n]$, and constructs $ST(T[1 \dots i])$ gradually for each step i ($1 \leq i \leq n$). The construction process is done by adding every suffix of $T[1 \dots i]$ into the tree in descending order of the length for each step i . Note that, however, the nodes corresponding to suffixes of $T[1 \dots i]$ are not necessarily to be leaves of $ST(T[1 \dots i])$ since it does not end with the terminal symbol in progress phase. Such an incomplete suffix tree is called an *implicit suffix tree*.

Consider that we are going to extend the suffix tree at a step i . We denote by ϕ the position on the tree to where a new leaf is going to be added, and we call it the *working node*. We regard the node at position ϕ as the node named ϕ if no confusion occurs. The working node can point a virtual node. Let $\phi = root$ when $i = 0$ for convenience. If $\bar{\phi} \cdot T[i]$ is not represented by $ST(T[1 \dots i])$, we add at ϕ the leaf whose label starts $T[i]$ as a child. If the node ϕ is a virtual node, we change it a real node and make a suffix link for it. Although the suffix link for the changed node is not stored explicitly in this case, we can traverse ST to the node representing $suf(\phi)$ by using the suffix link of the nearest neighbor ancestor that is also a real node.

When creating a leaf v and its edge, we la-

bel the edge as $label(v) = (i, *)$ by introducing the special symbol $*$, and regard $*$ as the same value as i for each step. Then, each leaf always represents a suffix of T . Now we finished the process to add the suffix $\bar{\phi} \cdot T[i]$ to the tree. Next we move the working node in order to add the shorter suffix and repeat the above process. Thus we create leaves if $\bar{\phi} \cdot T[i]$ is not represented by ST by traversing suffix path one by one until $\bar{\phi} \cdot T[i]$ has already represented. Then we update ϕ by traversing with the character $T[i]$, where is the starting working node position for the next step. Note that, we can stop the extension of ST when we encounter the node representing $\bar{\phi} \cdot T[i]$, because any suffix $Suf(\bar{\phi}) \cdot T[i]$ must appear at the same end position if $\bar{\phi} \cdot T[i]$ appears in T and thus any node on the suffix link which representing shorter suffix must be represented in the tree. For each step i , we call the portion of the suffix path *working path* where the leaves are added newly in the above creation.

We will omit the detail of the proof, the extension process can be done totally in time in proportion to the number of added nodes. Finally, we can obtain the complete suffix tree by adding the terminal symbol $\$$ to the implicit suffix tree $ST(T[1 \dots n])$ constructed as above.

3. Suffix tree with k words limitation

Let $\# \notin \Sigma$ be a symbol, called the *delimiter symbol*. Any string $x \in (\Sigma \cup \{\#\})^*$ can be written as $x = w_1\#w_2\#\dots\#w_N$ for the sequence of strings w_1, w_2, \dots, w_N ($\forall i, w_i \in \Sigma^*$). We call each w_i a *word*, and also call w the *string of N -words* divided by $\#$.

Assume that the given text is a string of N -words divided by $\#$. For convenience, we assume that $w_i \in \Sigma^+$ for any i , and that $\#$ never appears continuously. This means that we regard the consecutive $\#$ s as one delimiter symbol. The assumption is natural even in a practical use. For example, it is natural for English (or European language) texts to regard consecutive

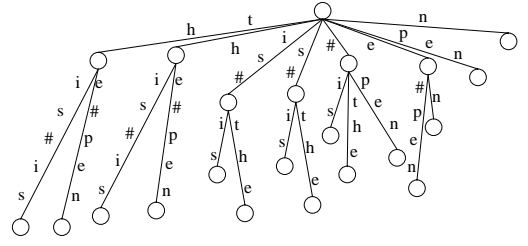


Figure 2: 2- $WST(T = \text{"this\#is\#the\#pen"})$

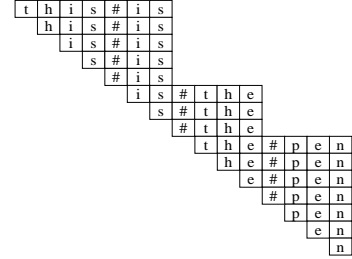


Figure 3: Suffixes represented by 2- $WST(T = \text{"this\#is\#the\#pen"})$

spaces and new line symbols as one delimiter. Although there is not delimiter symbols between words in phrases written in some oriental languages like Japanese, we can consider that it is separated into words with some NLP tools in advance. Then, for a given text $T = w_1\#w_2\#\dots\#w_N$ and a integer k ($1 \leq k \leq N$), we define the *suffix tree with k words limitation* (k - WST for short) as the rooted tree by doing path compression of the trie which represents $Fac(w_i\#w_{i+1}\#\dots\#w_{i+k-1})$ for any $1 \leq i \leq N - k + 1$ (Figure 2, Figure 3). Formally, k - $WST(T)$ is a rooted tree that satisfies the following conditions.

1. Each edge of k - $WST(T)$ is labeled with non-empty factor of T .
2. Each internal node of k - $WST(T)$ except for the root node has at least two children, and the edges from v to each children are labeled with strings that start with different characters mutually.
3. For each leaf v of k - $WST(T)$, \bar{v} corresponds one-to-one to a string $s \in Suf(w_i\#w_{i+1}\#\dots\#w_{i+k-1})$ ($1 \leq i \leq N - k + 1$).

We will discuss about the algorithm for constructing k - $WST(T)$, and present it can be

constructed in $O(n)$ time and space. Therefore, our claim is as follows.

Theorem 1. *Given a N -words text $T = w_1\#w_2\#\dots\#w_N$ whose length is n , assuming $\# \notin \Sigma$ and $\forall i, w_i \in \Sigma^+$, and a integer k ($1 \leq k \leq N$), the suffix tree with k words limitation of T can be constructed in $O(n)$ time and space.*

3.1. Constructing algorithm

The most important thing is not to process for redundant leaves and to close the extension of appropriate leaves just before they extend over $k + 1$ words when the tree is updated for each step. To close the extension of the leaf is done by replacing the label's end mark $*$ into $i - 1$. We call this process *closing a leaf*. Here, we consider that the label for each closed leaf is added virtually a terminal symbol $\natural \notin \Sigma$. From the above procedure, we can prevent from entering the suffixes longer than $k + 1$ words into the tree.

Figure 4 is the detail of the algorithm.

Procedure for updating the tree Consider that we have now k -WST($T[1 \dots i - 1]$) and process the updating the tree for $T[i]$. Let num_ϕ be the number of $\#$ which the suffix $\bar{\phi}$ includes, and also let $lpos_\phi$ be the start position of $\bar{\phi}$ in T , that is, $\bar{\phi} = T[lpos_\phi \dots i - 1]$. We increase num_ϕ by one when traversing the tree with $T[i] = \#$. On the other hand, we decrease num_ϕ by one when traversing the suffix path and $T[lpos_\phi] = \#$. Then we increase $lpos_\phi$ by one. We can manage correctly the number of words for $\bar{\phi}$ in this way.

The procedure for adding leaves on the working path is basically the same as that of Ukkonen's algorithm. However, we need additional effort if the k -words string contains another k -words string as a prefix like the case in Figure 5. Then, we must create a leaf which represents the end of the word for the shorter one, too. Although the label of it is ε , we assume that it represents a virtual terminal symbol $\natural \notin \Sigma$. On the other hand,

```

procedure update( $\phi, num_\phi, lpos_\phi, i$ );
method:
   $c = T[i]$ ;  $olldr = root$ ;
  while ( $\bar{\phi}c$  has not represented yet) {
    if ( $c \neq \#$  or
      there isn't a closed leaf representing  $\bar{\phi}$ ) {
      if ( $\phi$  is a virtual node)
        Change the node  $\phi$  into a real node;
      Create a child  $q$ ,  $label(q) = (i, *)$  of  $\phi$ ;
      if ( $olldr \neq root$ )
         $suf(olldr) = \phi$ ;
       $olldr = \phi$ ;
    }
     $\phi = suf(\phi)$ ;
    if ( $T[lpos_\phi] = \#$ )
       $num_\phi = num_\phi - 1$ ;
       $lpos_\phi = lpos_\phi + 1$ ;
    }
    if ( $olldr \neq root$ )
       $suf(olldr) = \phi$ ;
     $\phi = \bar{\phi}c$ ;
    if ( $c = \#$ )
       $num_\phi = num_\phi + 1$ ;
    return ( $\phi, num_\phi, lpos_\phi$ );
  }
end.

```

```

procedure close( $\psi, num_\psi, lpos_\psi, i$ );
method:
  if ( $T[i] = \#$ ) {
    while ( $num_\psi = k - 1$ ) {
      Close the leaf  $\psi$  with  $i - 1$ ;
      if ( $T[lpos_\psi] = \#$ )
         $num_\psi = num_\psi - 1$ ;
         $lpos_\psi = lpos_\psi + 1$ ;  $\psi = suf(\psi)$ ;
      }
       $num_\psi = num_\psi + 1$ ;
    }
    return ( $\psi, num_\psi, lpos_\psi$ );
  }
end.

```

Algorithm constructing k -WST;
input: $k, T[1 \dots n]$;
output: k -WST(T);
method:
 Create the root $root$ and the auxiliary node \perp ,
 and then add an edge $(\perp, root)$;
foreach $c \in \Sigma \cup \{\#, \$\}$ **do**
 Add an edge $(\perp, root)$ labeled with c ;
 $\phi = root$; $num_\phi = 0$; $lpos_\phi = 1$;
 $\psi = root$; $num_\psi = 0$; $lpos_\psi = 1$;
for ($i = 1 \dots n$) {
 ($\phi, num_\phi, lpos_\phi$) = update($\phi, num_\phi, lpos_\phi, i$);
 ($\psi, num_\psi, lpos_\psi$) = close($\psi, num_\psi, lpos_\psi, i$);
 }
 Report k -WST(T).

Figure 4: Algorithm for constructing k -WST

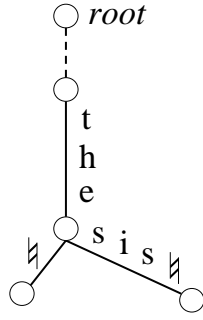


Figure 5: Virtual terminal symbol

we do not create a new leaf if there exists a closed leaf at the position ϕ .

Procedure for closing leaves Consider that we have now k -WST($T[1 \dots i-1]$) and process the updating the tree for $T[i] = \#$. Then, we need to close each leaf which represents the suffix that includes just $k-1$ $\#$ s. To do this, we close all relevant leaves from the leaf representing the longest suffix at this point in descending order of length.

Let ψ be a position of the leaf now we focused on. Also let num_ψ be the number of $\#$ within $\bar{\psi}$, and $lpos_\psi$ be the start position of $\bar{\psi}$ in T in a similar way of the updating procedure. We increase num_ψ by one when $T[i] = \#$, and close ψ with $i-1$ if $num_\psi = k$, and then repeat this after traversing the suffix link until $num_\psi < k$.

3.2. Complexity of the algorithm

Lemma 1. *The procedure for updating the tree can be done in $O(n)$ time totally.*

proof. The procedure create up to n leaves for each $i = 1, \dots, n$. Creating the internal node if necessary and then adding an edge into the new leaf, are done in constant time. Although the procedure may create up to n leaves at once, it can be done in $O(n)$ time totally since the number of the leaves in k -WST(T) is at most $O(n)$ finally. \square

Lemma 2. *The procedure for closing leaves can be done in $O(n)$ time totally.*

proof. The position ψ traverse the tree by one character after the closing loop for each

$T[i]$. Thus it moves n times totally. To update num_ψ and $lpos_\psi$ for each movement can be done in constant time. On the other hand, the position ψ traverse the suffix link only when $num_\psi = k-1$ and $T[i] = \#$, that is, when closing leaves. To close a leaf can be done in constant time. Therefore, the procedure for closing leaves can be done in $O(n)$ time totally since the number of the leaves is $O(n)$. \square

3.3. Managing the occurrence rate

We can accumulate the occurrence rate of any factor in T by using the original suffix tree $ST(T\$)$, where the occurrence rate of each suffix is exactly 1. A factor \bar{u} represented by an internal node u is a prefix of the strings represented by their children. Thus, for any children u' of u , if \bar{u}' occurs p times in T , then \bar{u} occurs p times at the same position. Therefore, the occurrence rate of \bar{u} can be calculated as the sum of the occurrence rate of every u' 's children.

However, for k -WST($T\$$), not all the occurrence rates for closed leaves are 1, because that it is not necessarily the case that each leaf v does not correspond to a suffix and thus \bar{v} may occur in T several times. To solve this problem, we add the frequency information $f(v)$ to each leaf $v \in V$, and increase $f(v)$ by one when \bar{v} occurs in T again. Then, we can calculate the occurrence rate for every node in k -WST($T\$$) by traversing the tree in depth first order.

3.4. Estimation of space efficiency

To estimate the space efficiency of k -WST, we implemented the algorithm and measured the change in the number of nodes created through the running. We used Reuters-21578 as a test data, where we removed all the SGML tags from the original data and concatenate only English phrases. The size of the data is about 18.8MB. We constructed k -WST for $k = 1, 2, \dots, 6$, and the original suffix tree in addition. The experimental result is shown in Figure 6. As we see, the number of nodes falls broadly when k

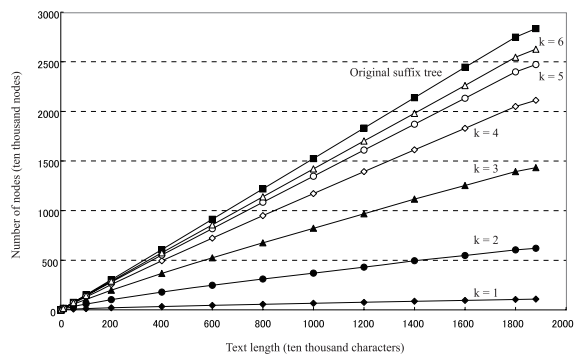


Figure 6: Change in the number of nodes

is small, while it is almost the same when $k = 6$.

3.5. Application to Kiwi

As mentioned in [13], the processing for clipping candidates which denoted in Section can be realized by using suffix trees but n-gram trie. For the case of Kiwi system, however, we do not need substrings longer than several words. Their positions in the text are also unnecessary. Moreover the longer candidates are not unnaturally broken when we use k -WST as an index structure but n-gram trie, since the lengths of the substrings which we can refer to are not fixed. From the above observation we conclude that our data structure k -WST is more suitable for Kiwi system rather than the original suffix tree or the n-gram trie.

4. Conclusion

In this paper, we proposed the truncated suffix tree with k -words limitation, which can represent any factor less than k -words in a given text $T = T[1 \dots n]$, and also presented it can be constructed in $O(n)$ time. In our implementation of the data structure, the experimental results showed that the number of nodes falls broadly when k is small, and it becomes about half when $k = 3$ comparing with the original suffix tree. We also mentioned about the possibility of application to Kiwi system.

Several researchers have proposed the techniques to enter the suffixes for a word-based sequence rather than character-based[3, 6]. It is our future work to combine the technique into our idea.

References

- [1] Webcorp home page, 1999. <http://www.webcorp.org.uk/>.
- [2] Googlefight : Make a fight with googlefight, 2000. <http://www.googlefight.com/>.
- [3] Arne Andersson, N. Jesper Larsson, and Kurt Swanson. Suffix trees on words. *Algorithmica*, 23(3):246–260, 1999.
- [4] E. R. Fiala and D. H. Greene. Data compression with finite windows. *Comm. ACM*, 32(4):490–505, 1989.
- [5] Dan Gusfield. *Algorithms on strings, trees, and sequences - Computer science and computational biology*. Cambridge University Press, 1997.
- [6] Shunsuke Inenaga and Masayuki Takeda. On-line linear-time construction of word suffix trees. In *In proc. of 17th Ann. Symp. on Combinatorial Pattern Matching*, 2006. (to appear).
- [7] N. J. Larsson. Extended application of suffix trees to data compression. In *In Proc. of Data Compression Conference*, pages 190–199, 1996.
- [8] E. M. McCreight. A space-economical suffix tree construction algorithm. *J. ACM*, 23:262–272, 1976.
- [9] J. C. Na, A. Apostolico, C. S. Iliopoulos, and K. Park. Truncated suffix trees and their application to data compression. *Theoretical Computer Science*, 304:87–101, 2003.
- [10] G. Peters. Geoff’s googleduel!, 2002. <http://www.googleduel.com/original.php>.
- [11] M. Rodeh, V. R. Pratt, and S. Even. Linear algorithm for data compression via string matching. *J. ACM*, 28(1):16–24, 1981.

- [12] Kumiko Tanaka-Ishii and Yuichiro Ishii. Tonguen - search your tongue's end, 2006. <http://cantor.ish.ci.i.u-tokyo.ac.jp/~kumiko/tonguen/tonguen.cgi>.
- [13] Kumiko Tanaka-Ishii and Hiroshi Nakagawa. Kiwi site, 2004. <http://kiwi.r.dl.itc.u-tokyo.ac.jp/>.
- [14] Kumiko Tanaka-Ishii and Hiroshi Nakagawa. A multilingual usage consultation tool based on internet searching: more than a search engine, less than qa. In *In Proc. of the 14th internat. WWW Conference*, pages 363–371, 2005.
- [15] E. Ukkonen. On-line construction of suffix-trees. *Algorithmica*, 14(3):249–260, 1995.
- [16] P. Weiner. Linear pattern matching algorithms. In *In Proc. of the 14th IEEE Symp. on Switching and Automata Theory*, pages 1–11, 1973.

Generating Frequent Closed Item Sets Using Zero-suppressed BDDs

Shin-ichi Minato

Graduate School of Information Science and Technology,
Hokkaido University
Sapporo, 060-0814 Japan.

Abstract

Frequent item set mining is one of the fundamental techniques for knowledge discovery and data mining. In the last decade, a number of efficient algorithms for frequent item set mining have been presented, but most of them focused on just enumerating the item set patterns which satisfy the given conditions, and it was a different matter how to store and index the result of patterns for efficient data analysis. Recently, we proposed a fast algorithm of extracting all frequent item set patterns from transaction databases and simultaneously indexing the result of huge patterns using Zero-suppressed BDDs (ZBDDs). That method, ZBDD-growth, is not only enumerating/listing the patterns efficiently, but also indexing the output data compactly on the memory to be analyzed with various algebraic operations. In this paper, we present a variation of ZBDD-growth algorithm to generate frequent closed item sets. This is a quite simple modification of ZBDD-growth, and additional computation cost is relatively small compared with the original algorithm for generating all patterns. Our method can conveniently be utilized in the environment of ZBDD-based pattern indexing.

1. Introduction

Frequent item set mining is one of the fundamental techniques for knowledge discovery and data mining. Since the introduction by Agrawal et al.[1], the frequent item set mining and association rule analysis have been received much attentions from many researchers, and a number of papers have been published about the new algorithms or improvements for solving such mining problems[4, 6, 11]. However, most of such item set mining algorithms focused on just enumerating or listing the item set patterns which satisfy the given conditions and it was a different matter how to store and index the result of patterns for efficient data analysis.

Recently, we proposed a fast algorithm[8] of extracting all frequent item set patterns from transaction databases, and simultaneously indexing the result of huge patterns on the computer memory using

Zero-suppressed BDDs. That method, called *ZBDD-growth*, does not only enumerate/list the patterns efficiently, but also indexes the output data compactly on the memory. After mining, the result of patterns can efficiently be analyzed by using algebraic operations.

The key of the method is to use BDD-based data structure for representing sets of patterns. BDDs[2] are graph-based representation of Boolean functions, now widely used in VLSI logic design and verification area. For the data mining applications, it is important to use Zero-suppressed BDDs (ZBDDs)[7], a special type of BDDs, which are suitable for handling large-scale sets of combinations. Using ZBDDs, we can implicitly enumerate combinatorial item set data and efficiently compute set operations over the ZBDDs.

In this paper, we present an interesting variation of ZBDD-growth algorithm to

generate frequent closed item sets. Closed item sets are the subset of item set patterns each of which is the unique representative for a group of sub-patterns relevant to the same set of transaction records. Our method is a quite simple modification of ZBDD-growth. We inserted several operations in the recursive procedure of ZBDD-growth, to filter the closed patterns from all frequent patterns. The experimental result shows that the additional computation cost is relatively small compared with the original algorithm for generating all patterns. Our method can conveniently be utilized in the environment of ZBDD-based data mining and knowledge indexing.

2. ZBDD-based item set representation

As the preliminary section, we describe the methods for efficiently indexing item set data based on Zero-suppressed BDDs.

2.1. Combinatorial item set and ZBDDs

A combinatorial item set consists of the elements each of which is a combination of a number of items. There are 2^n combinations chosen from n items, so we have 2^{2^n} variations of combinatorial item sets. For example, for a domain of five items $a, b, c, d,$ and $e,$ we can show examples of combinatorial item sets as:

$\{ab, e\}, \{abc, cde, bd, acde, e\}, \{1, cd\}, 0.$ Here “1” denotes a combination of null items, and 0 means an empty set. Combinatorial item sets are one of the basic data structure for various problems in computer science, including data mining.

A combinatorial item set can be mapped into Boolean space of n input variables. For example, Fig. 1 shows a truth table of Boolean function: $F = (a b \bar{c}) \vee (\bar{b} c),$ but also represents a combinatorial item set $S = \{ab, ac, c\}.$ Using BDDs for the corresponding Boolean functions, we can implicitly represent and manipulate combinatorial item set. In addition, we

a	b	c	F
0	0	0	0
1	0	0	0
0	1	0	0
1	1	0	1
0	0	1	1
1	0	1	1
0	1	1	0
1	1	1	0

As a Boolean function:
 $F(a,b,c) = (a b \sim c) \vee (\sim b c)$
 As a combinatorial item set:
 $S(a,b,c) = \{ab, ac, c\}$

→ ab
 → c
 → ac

Figure 1: A Boolean function and a combinatorial item set.

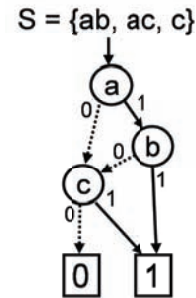


Figure 2: An example of ZBDD.

can enjoy more efficient manipulation using “Zero-suppressed BDDs” (ZBDD)[7], which are special type of BDDs optimized for handling combinatorial item sets. An example of ZBDD is shown in Fig. 2.

The detailed techniques of ZBDD manipulation are described in the articles[7]. A typical ZBDD package supports cofactoring operations to traverse 0-edge or 1-edge, and binary operations between two combinatorial item sets, such as union, intersection, and difference. The computation time for each operation is almost linear to the number of ZBDD nodes related to the operation.

2.2. Tuple-Histograms and ZBDD vectors

A *tuple-histogram* is the table for counting the number of appearance of each tuple in the given database. An example of tuple-histogram is shown in Fig. 3. This is just a compressed table of the database to combine the same tuples appearing more than once into one line with the frequency.

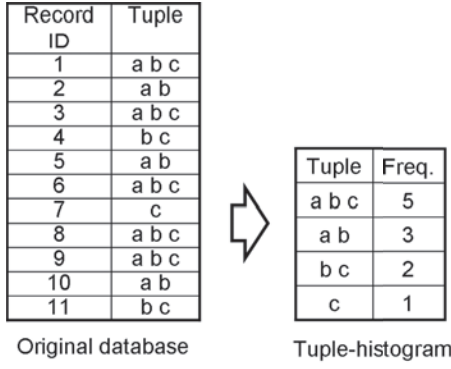


Figure 3: Example of tuple-histogram.

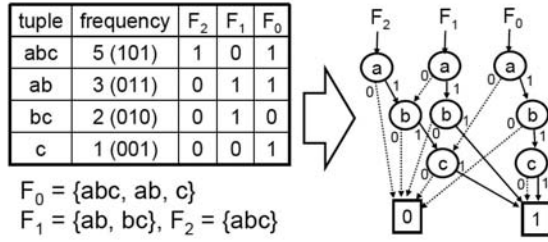


Figure 4: ZBDD vector for tuple-histogram.

Our item set mining algorithm manipulates ZBDD-based tuple-histogram representation as the internal data structure. Here we describe how to represent tuple-histograms using ZBDDs. Since ZBDDs are representation of sets of combinations, a simple ZBDD distinguishes only existence of each tuple in the database. In order to represent the numbers of tuple’s appearances, we decompose the number into m -digits of ZBDD vector $\{F_0, F_1, \dots, F_{m-1}\}$ to represent integers up to $(2^m - 1)$, as shown in Fig. 4. Namely, we encode the appearance numbers into binary digital code, as F_0 represents a set of tuples appearing odd times (LSB = 1), F_1 represents a set of tuples whose appearance number’s second lowest bit is 1, and similar way we define the set of each digit up to F_{m-1} .

In the example of Fig. 4, The tuple frequencies are decomposed as: $F_0 = \{abc, ab, c\}$, $F_1 = \{ab, bc\}$, $F_2 = \{abc\}$, and then each digit can be represented

by a simple ZBDD. The three ZBDDs are shared their sub-graphs each other.

Now we explain the procedure for constructing a ZBDD-based tuple-histogram from original database. We read a tuple data one by one from the database, and accumulate the single tuple data to the histogram. More concretely, we generate a ZBDD of T for a single tuple picked up from the database, and accumulate it to the ZBDD vector. The ZBDD of T can be obtained by starting from “1” (a null-combination), and applying “Change” operations several times to join the items in the tuple. Next, we compare T and F_0 , and if they have no common parts, we just add T to F_0 . If F_0 already contains T , we eliminate T from F_0 and carry up T to F_1 . This ripple carry procedure continues until T and F_k have no common part. After finishing accumulations for all data records, the tuple-histogram is completed.

Using the notation $F.add(T)$ for addition of a tuple T to the ZBDD vector F , we describe the procedure of generating tuple-histogram H for given database D .

```

H = 0
forall T ∈ D do
    H = H.add(T)
return H

```

When we construct a ZBDD vector of tuple-histogram, the number of ZBDD nodes in each digit is bounded by total appearance of items in all tuples. If there are many partially similar tuples in the database, the sub-graphs of ZBDDs are shared very well, and compact representation is obtained. The bit-width of ZBDD vector is bounded by $\log S_{max}$, where S_{max} is the appearance of most frequent items.

Once we have generated a ZBDD vector for the tuple-histogram, various operations can be executed efficiently. Here are the instances of operations used in our pattern mining algorithm.

- $H.factor0(v)$: Extracts sub-histogram of tuples without item v .

- $H.\text{factor1}(v)$: Extracts sub-histogram of tuples including item v and then delete v from the tuple combinations. (also considered as the quotient of H/v)
- $v \cdot H$: Attaches an item v on each tuple combinations in the histogram F .
- $H_1 + H_2$: Generates a new tuple-histogram with sum of the frequencies of corresponding tuples.
- $H.\text{tuplecount}$: The number of tuples appearing at least once.

These operations can be composed as a sequence of ZBDD operations. The result is also compactly represented by a ZBDD vector. The computation time is bounded by roughly linear to total ZBDD sizes.

3. ZBDD-growth Algorithm

Recently, we developed a ZBDD-based algorithm[8], ZBDD-growth, to generate “all” frequent item set patterns. Here we describe this algorithm as the basis of our method for “closed” item set mining.

ZBDD-growth is based on a recursive depth-first search over the ZBDD-based tuple-histogram representation. The basic algorithm is shown in Fig. 5.

In this algorithm, we choose an item v used in the tuple-histogram H , and compute the two sub-histograms H_1 and H_0 . (Namely, $H = (v \cdot H_1) \cup H_0$.) As v is the top item in the ZBDD vector, H_1 and H_0 can be obtained just by referring the 1-edge and 0-edge of the highest ZBDD-node, so the computation time is constant for each digit of ZBDD.

The algorithm consists of the two recursive calls, one of which computes the subset of patterns including v , and the other computes the patterns excluding v . The two subsets of patterns can be obtained as a pair of pointers to ZBDDs, and then the final result of ZBDD is computed.

```

ZBDDgrowth( $H, \alpha$ )
{
  if( $H$  has only one item  $v$ )
    if( $v$  appears more than  $\alpha$ )
      return  $v$ ;
    else return “0”;
   $F \leftarrow \text{Cache}(H)$ ;
  if( $F$  exists) return  $F$ ;
   $v \leftarrow H.\text{top}$ ; /* Top item in  $H$  */
   $H_1 \leftarrow H.\text{factor1}(v)$ ;
   $H_0 \leftarrow H.\text{factor0}(v)$ ;
   $F_1 \leftarrow \text{ZBDDgrowth}(H_1, \alpha)$ ;
   $F_0 \leftarrow \text{ZBDDgrowth}(H_0 + H_1, \alpha)$ ;
   $F \leftarrow (v \cdot F_1) \cup F_0$ ;
   $\text{Cache}(H) \leftarrow F$ ;
  return  $F$ ;
}

```

Figure 5: ZBDD-growth algorithm.

This procedure may require an exponential number of recursive calls, however, we prepare a hash-based cache to store the result of each recursive call. Each entry in the cache is formed as pair (H, F) , where H is the pointer to the ZBDD vector for a given tuple-histogram, and F is the pointer to the result of ZBDD. On each recursive call, we check the cache to see whether the same histogram H has already appeared, and if so, we can avoid duplicate processing and return the pointer to F directly. By using this technique, the computation time becomes almost linear to the total ZBDD sizes.

In our implementation, we use some simple techniques to prune the search space. For example, if H_1 and H_0 are equivalent, we may skip to compute F_0 . For another case, we can stop the recursive calls if total frequencies in H is no more than α . There are some other elaborate pruning techniques, but they needs additional computation cost for checking the conditions, so sometimes effective but not always.

4. Frequent closed item set mining

In frequent item set mining, we sometimes

faced with the problem that a huge number of frequent patterns are extracted and hard to find useful information. Closed item set mining is one of the techniques to filter important subset of patterns. In this section, we present a variation of ZBDD-growth algorithm to generate frequent closed item sets.

4.1. Closed item sets

Closed item sets are the subset of item set patterns each of which is the unique representative for a group of sub-patterns relevant to the same set of tuples. For more clear definition, we first define the *common item set* $Com(S_T)$ for the given set of tuples S_T , such that $Com(S_T)$ is the set of items commonly included in every tuple $T \in S_T$. Next, we define *occurrence* $Occ(D, X)$ for the given database D and item set X , such that $Occ(D, X)$ is the subset of tuples in D , each of which includes X . Using these notations, if an item set X satisfies $Com(Occ(D, X)) = X$, we call X is a closed item set in D .

For example, let us consider the database D as shown in Fig. 3. Here, all item set patterns with threshold $\alpha = 1$ is: $\{abc, ab, ac, a, bc, b, c\}$, but closed item sets are: $\{abc, ab, bc, b, c\}$. In this example, “ ac ” is eliminated from a closed pattern because $Occ(D, “ac”) = Occ(D, “abc”)$.

In recent years, many researchers discuss the efficient algorithms for closed item set mining. One of the remarkable result is *LCM* algorithm[10] presented by Uno et. al. LCM is a depth-first search algorithm to extract closed item sets. It features that the computation time is bounded by linear to the output data length. Our ZBDD-based algorithm is also based on a depth-first search manner, so, it has similar properties as LCM. The major difference is in the data structure of output data. Our method generates ZBDDs for the set of closed patterns, ready to go for more flexible analysis using ZBDD operations.

4.2. Eliminating non-closed patterns

Our method is a quite simple modification of ZBDD-growth shown in Fig. 5. We inserted several operations in the recursive procedure of ZBDD-growth, to filter the closed patterns from all frequent patterns. The ZBDD-growth algorithm is starting from the given tuple-histogram H , and compute the two sub-histograms H_1 and H_0 , such that $H = (v \cdot H_1) \cup H_0$. Then ZBDD-growth(H_1) and ZBDD-growth($H_1 + H_0$) is recursively executed.

Here, we consider the way to eliminate non-closed patterns in this algorithm. We call the new algorithm ZBDD-growthC(H). It is obvious that $(v \cdot ZBDD-growthC(H_1))$ generates (a part of) closed patterns for H each of which includes v , because the occurrence of any closed pattern with v is limited in $(v \cdot H_1)$, thus we may search only for H_1 . Next, we consider the second recursive call ZBDD-growthC($H_1 + H_0$) to generate the closed patterns without v . Important point is that some of patterns generated by ZBDD-growthC($H_1 + H_0$) may have the same occurrence as one of the pattern with v already found in H_1 . The condition of such duplicate pattern is that it appears only in H_1 but irrelevant to H_0 . In other words, we eliminate the patterns from ZBDD-growthC($H_1 + H_0$) such that the patterns are already found in ZBDD-growthC(H_1) but not included in any tuples in H_0 .

For checking the condition for closed patterns, we can use a ZBDD-based operation, called *permit* operation by Okuno et al.[9].¹ $P.permit(Q)$ returns a set of combinations in P each of which is a subset of some combinations in Q . For example, when $P = \{ab, abc, bcd\}$ and $Q = \{abc, bc\}$, then $P.permit(Q)$ returns $\{ab, abc\}$. The permit operation is efficiently implemented as a recursive proce-

¹Permit operation is basically same as *SubSet* operation by Coudert et al.[3], defined for ordinary BDDs.

```

P.permit(Q)
{
  if(P = "0" or Q = "0") return "0" ;
  if(P = Q) return F ;
  if(P = "1") return "1" ;
  if(Q = "1")
    if(P include "1" ) return "1" ;
    else return "0" ;
  R ← Cache(P, Q) ;
  if(R exists) return R ;
  v ← TopItem(P, Q) ; /* Top item in P, Q */
  (P0, P1) ← factors of P by v ;
  (Q0, Q1) ← factors of Q by v ;
  R ← (v · P1.permit(Q1))
    ∪ (P0.permit(Q0 ∪ Q1)) ;
  Cache(P, Q) ← R ;
  return R ;
}

```

Figure 6: Permit operation.

ture of ZBDD manipulation, as shown in Fig. 6. The computation time of permit operation is almost linear to the ZBDD size.

Finally, we describe the ZBDD-growthC algorithm using the permit operation, as shown in Fig. 7. The difference from the original algorithm is only one line, written in the frame box.

5. Experimental Results

Here we show the experimental results to evaluate our new method. We used a Pentium-4 PC, 800MHz, 1.5GB of main memory, with SuSE Linux 9. We can deal with up to 40,000,000 nodes of ZBDDs in this machine.

Table 1 shows the time and space for generating ZBDD vectors of tuple-histograms for the FIMI2003 benchmark databases[5]. This table shows the computation time and space for providing input data for ZBDD-growth algorithm. In this table, $\#T$ shows the number of tuples, $total|T|$ is the total of tuple sizes (total appearances of items), and $|ZBDD|$ is the number of ZBDD nodes for the tuple-histograms. We can see that tuple-

```

ZBDDgrowthC(H, α)
{
  if(H has only one item v)
    if(v appears more than α )
      return v ;
    else return "0" ;
  F ← Cache(H) ;
  if(F exists) return F ;
  v ← H.top ; /* Top item in H */
  H1 ← H.factor1(v) ;
  H0 ← H.factor0(v) ;
  F1 ← ZBDDgrowthC(H1, α) ;
  F0 ← ZBDDgrowthC(H0 + H1, α) ;
  F ← (v · F1) ∪
    (F0 - (F1 - F1.permit(H0))) ;
  Cache(H) ← F ;
  return F ;
}

```

Figure 7: ZBDD-growthC algorithm.

histograms can be constructed for all instances in a feasible time and space. The ZBDD sizes are almost same or less than $total|T|$.

After generating ZBDD vectors for the tuple-histograms, we applied ZBDD-growth algorithm to generate frequent patterns. Table 2 show the results of the original ZBDD-growth algorithm[8] for the selected benchmark examples, "mushroom," "T10I4D100K," and "BMS-WebView-1." The execution time includes the time for generating the initial ZBDD vectors for tuple-histograms.

The results shows that the ZBDD size is exponentially smaller than the number of patterns for "mushroom." This is a significant effect of using the ZBDD data structure. On the other hand, no remarkable reduction is seen in "T10I4D100K." "T10I4D100K" is known as an artificial database, consists of randomly generated combinations, so there are almost no relationship between the tuples. In such cases, ZBDD nodes cannot be shared well, and only the overhead factor is revealed. For the third example, "BMS-WebView-1," the ZBDD size is almost linear to the

Table 1: Generation of tuple-histograms[8].

Data name	# T	total $ T $	ZBDD Vector	Time(s)
T10I4D100K	100,000	1,010,228	552,429	43.2
T40I10D100K	100,000	3,960,507	3,396,395	150.2
chess	3,196	118,252	40,028	1.4
connect	67,557	2,904,951	309,075	58.1
mushroom	8,124	186,852	8,006	1.2
pumsb	49,046	3,629,404	1,750,883	188.5
pumsb_star	49,046	2,475,947	1,324,502	123.6
BMS-POS	515,597	3,367,020	1,350,970	895.0
BMS-WebView-1	59,602	149,639	46,148	18.3
BMS-WebView-2	77,512	358,278	198,471	138.0
accidents	340,183	11,500,870	3,877,333	107.0

Table 2: Result of the original ZBDD-growth[8].

Data name: Min. freq. α	#Frequent patterns	(output) ZBDD	Time(sec)
mushroom: 5,000	41	11	1.2
1,000	123,277	1,417	3.7
200	18,094,821	12,340	9.7
50	198,169,865	36,652	10.2
16	1,176,182,553	53,804	7.7
4	3,786,792,695	59,970	4.3
1	5,574,930,437	40,557	1.8
T10I4D100K: 5,000	10	10	81.3
1,000	385	382	135.5
200	13,255	4,288	279.4
50	53,385	20,364	408.7
16	175,915	89,423	543.3
4	3,159,067	1,108,723	646.0
BMS-WebView1: 1,000	31	31	27.8
200	372	309	31.3
50	8,191	3,753	49.0
40	48,543	12,176	46.6
36	461,521	34,790	102.4
35	1,177,607	47,457	111.4
34	4,849,465	64,601	120.8
33	69,417,073	80,604	130.0
32	1,531,980,297	97,692	133.7
31	8,796,564,756,112	117,101	138.1
30	35,349,566,550,691	152,431	143.9

number of patterns when the output size is small, however, an exponential factor of reduction is observed for the cases of generating huge patterns.

Next, we show the experimental results of frequent closed pattern mining using ZBDD-growthC algorithm. In Table 3, we show the results for the same examples as used in the experiment of the original ZBDD-growth. The last column $Time_{(closed)}/Time_{(all)}$ shows the ratio of computation time between the ZBDD-growthC and the original ZBDD-growth algorithm. We can observe that the computation time is almost the same order as the original algorithms for “mushroom” and “BMS-WebView-1,”

but some additional factor is observed for “T10I4D100K.” Anyway, filtering closed item sets has been regarded as not a easy task. We can say that the ZBDD-growthC algorithm can generate closed item sets with a relatively small additional cost from the original ZBDD-growth.

6. Conclusion

In this paper, we presented an interesting variation of ZBDD-growth algorithm to generate frequent closed item sets. Our method is a quite simple modification of ZBDD-growth. We inserted several operations in the recursive procedure of ZBDD-growth, to filter the closed pat-

Table 3: Results of ZBDD-based closed pattern mining.

Data name: Min. freq. α	#Freq. closed patterns	(output) ZBDD	ZBDD- growthC Time(s)	$Time_{(closed)}$ $/Time_{(all)}$
mushroom:				
5,000	16	16	1.2	1.00
1,000	3,427	1,660	3.8	1.02
200	26,968	9,826	9.9	1.02
50	68,468	19,054	13.0	1.27
16	124,411	24,841	13.3	1.73
4	203,882	26,325	13.2	3.06
1	238,709	20,392	12.9	7.19
T10I4D100K:				
5,000	10	10	104.8	1.29
1,000	385	382	208.1	1.54
200	13,108	4,312	2713.6	9.71
50	46,993	20,581	4600.1	11.25
16	142,520	89,185	5798.5	10.67
4	1,023,614	691,154	18573.0	28.75
BMS-WebView-1:				
1,000	31	31	30.1	1.08
200	372	309	36.8	1.18
50	7,811	3,796	71.9	1.47
40	29,489	11,748	111.4	2.39
36	64,762	25,117	153.7	1.50
35	76,260	30,011	169.2	1.52
34	87,982	35,392	186.5	1.54
33	99,696	40,915	207.7	1.60
32	110,800	46,424	221.7	1.66
31	120,190	51,369	247.7	1.79
30	127,131	55,407	271.5	1.89

terns from all frequent patterns. The experimental result shows that the additional computation cost is relatively small compared with the original algorithm for generating all patterns.

A ZBDD can be regarded as a compressed trie for representing a set of patterns. ZBDD-based method will be useful as a fundamental technique for database analysis and knowledge indexing, and will be utilized for various data mining applications.

References

- [1] R. Agrawal, T. Imielinski, and A. N. Swami, Mining Association rules between sets of items in large databases, In P. Buneman and S. Jajodia, editors, *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, Vol. 22(2) of SIGMOD Record, pp. 207–216, ACM Press, 1993.
- [2] Bryant, R. E., Graph-based algorithms for Boolean function manipulation, *IEEE Trans. Comput.*, C-35, 8 (1986), 677–691.
- [3] O. Coudert, J. C. Madre, H. Fraisse, A new viewpoint on two-level logic minimization, in *Proc. of 30th ACM/IEEE Design Automation Conference*, pp. 625-630, 1993.
- [4] B. Goethals, “Survey on Frequent Pattern Mining”, Manuscript, 2003. <http://www.cs.helsinki.fi/u/goethals/publications/survey.ps>
- [5] B. Goethals, M. Javeed Zaki (Eds.), Frequent Itemset Mining Dataset Repository, Frequent Itemset Mining Implementations (FIMI’03), 2003. <http://fimi.cs.helsinki.fi/data/>
- [6] J. Han, J. Pei, Y. Yin, R. Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach, *Data Mining and Knowledge Discovery*, 8(1), 53–87, 2004.
- [7] S. Minato: Zero-suppressed BDDs for set manipulation in combinatorial problems, In *Proc. 30th ACM/IEEE Design Automation Conf. (DAC-93)*, (1993), 272–277.
- [8] S. Minato, H. Arimura: ZBDD-growth: An Efficient Method for Frequent Pattern Mining and Knowledge Indexing, *Hokkaido University, Division of Computer Science, TCS Technical Reports*, TCS-TR-A-06-12, Apr. 2006. <http://www-alg.ist.hokudai.ac.jp/tra.html>
- [9] H. Okuno, S. Minato, and H. Isozaki: On the Properties of Combination Set

Operations, Information Processing Letters, Elsevier, 66 (1998), pp. 195-199, 1998.

- [10] T. Uno, Y. Uchida, T. Asai, and H. Arimura: "LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets," Proc. *Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Mohammed J. Zaki and Bart Goethals (eds.), 2003.
<http://fimi.cs.helsinki.fi/fimi03/>
- [11] M. J. Zaki, Scalable Algorithms for Association Mining, IEEE Trans. Knowl. Data Eng. 12(2), 372-390, 2000.

Data Mining in Amino Acid Sequences of H3N2 Influenza Viruses Isolated during 1968 to 2006

Kimihito Ito, Manabu Igarashi, Ayato Takada
Hokkaido University Research Center for Zoonosis Control
Sapporo 060-0818, Japan

Abstract

The hemagglutinin (HA) of influenza viruses undergoes antigenic drift to escape from antibody-mediated immune pressure. In order to predict possible structural changes of the HA molecules in future, it is important to understand the patterns of amino acid mutations and structural changes in the past. We performed a retrospective and comprehensive analysis of structural changes in H3 hemagglutinins of human influenza viruses isolated during 1968 to 2006. Amino acid sequence data of more than 2000 strains have been collected from NCBI Influenza virus resources. Information theoretic analysis of the collected sequences revealed a number of simultaneous mutations of amino acids at two or more different positions (correlated mutations). We also calculated the net charge of the HA1 subunit, based on the number of charged amino acid residues, and found that the net charge increased linearly from 1968 to 1984 and, after then, has been saturated. This level of the net charge may be an upper limit for H3 HA to be functional. It is noted that "correlated mutations" with the conversion of acidic and basic amino acid residues between two different positions were frequently found after 1984, suggesting that these mutations contributed to counterbalancing effect to keep the net charge of the HA. These approaches may open the way to find the direction of future antigenic drift of influenza viruses.

1. Introduction

Influenza A virus causes highly contagious, acute respiratory illness, and continues to be a major cause of morbidity and mortality. The viral strains mutate from year to year, causing the annual epidemic world wide.

The hemagglutinin (HA) is the major surface glycoprotein of influenza viruses and plays an important role in virus entry into host cells. The HA undergoes antigenic drifts which occur by accumulation of a series of amino acid substitutions. Influenza viruses that escape from antibody-mediated immune pressure of human population acquire a new antigenic structure and continuously cause epidemic in the world.

In order to predict possible structural changes of the HA molecules in future, it is important to understand the patterns of antigenic drift of influenza viruses in the past. We have been studying computational meth-

ods that include information theoretic analysis to find patterns of amino acid substitutions, molecular modeling to reveal the changes in 3D structure of antigenically different HAs, and molecular simulation to investigate the antigen-antibody interaction.

We performed a retrospective and comprehensive analysis of structural changes in H3 HAs of human influenza viruses isolated during 1968 to 2006. Amino acid sequence data of more than 2000 strains have been collected from NCBI Influenza virus resources and used for the analysis. These approaches may open the way to find the direction of future antigenic drift of influenza viruses.

2. Information Theoretic Analysis of the Past HA Sequences

The entropy and mutual information are used to find simultaneous mutations of

amino acids at two or more different positions (correlated mutations). The entropy of an amino acid position represents the uncertainty of the amino acids in the position. The amino acid positions with frequent mutations are expected to have higher entropy. The mutual information between two amino acid positions represents the average reduction in uncertainty about one amino acid position that results from learning the amino acid of another position. The pairs of amino acid positions that tend to be involved in correlated mutations are expected to have higher mutual information values.

Definition 1 Entropy: The entropy of X is defined to be the average Shannon information of an outcome.

$$H(X) = \sum_{x \in Ax} P(x) \log \frac{1}{P(x)}$$

Example 1 Suppose we have 6 protein sequences in which the amino acid of positions 1, 2, 3 are the following.

position 1	D	D	D	D	D	D
position 2	Q	Q	Q	Q	Q	I
position 3	D	D	I	I	N	V

The entropy of position 1,2, and 3 are

$$\begin{aligned} H(X_1) &= 0.00, \\ H(X_2) &= 0.65, \\ H(X_3) &= 1.91, \end{aligned}$$

respectively.

Definition 2 Mutual Information: The mutual information between X and Y is defined to be the average reduction in uncertainty about x by knowing the value of y .

$$I(X; Y) = H(X) - H(X|Y)$$

$$H(X|Y) = \sum_{xy \in AxAy} P(x, y) \log \frac{1}{P(x|y)}$$

Example 2 Suppose we have 6 protein sequences in which the amino acid of positions 1, 2, 3 are the following.

position 1	Q	Q	Q	Q	Q	Q
position 2	Q	Q	Q	I	I	V
position 3	D	D	D	N	N	K
position 4	Q	I	Q	Q	I	Q

The Mutual Information of among position 1, 2, 3 include

$$\begin{aligned} I(X_1; X_2) &= 0.00, \\ I(X_2; X_3) &= 1.45, \\ I(X_3; X_4) &= 0.12. \end{aligned}$$

3. Correlated Mutations Found by the Analysis

X	Y	$I(X; Y)$
144	156	1.0687064407
156	276	1.0606584462
62	158	1.0568431741
135	262	1.0462450293
156	158	0.9904144166
196	276	0.9548333157
62	276	0.9544171044
62	156	0.9440218619
158	276	0.9313234350
156	196	0.9269716745

Table 1: The top 10 pairs of amino acid positions in HA1 subunit that have high mutual information values.

2183 amino acid sequences of HA1s of human influenza viruses have been collected from NCBI Influenza virus resources. Mutual information for every pair of two amino acid positions in the HA1 subunit is calculated. Table 1 shows the top 10 pairs that have high mutual information values. The analysis finds a number of the correlated mutations in amino acid positions of the HA1, and three of them are shown in Figure 1.

The 3D structure model in Figure 2 depicts the location of amino acid positions that appear in the pairs in Table 1.

The correlated mutation between amino acid position 163 and 248, which have mutual information 0.61 during 1968 to 1989, resulted

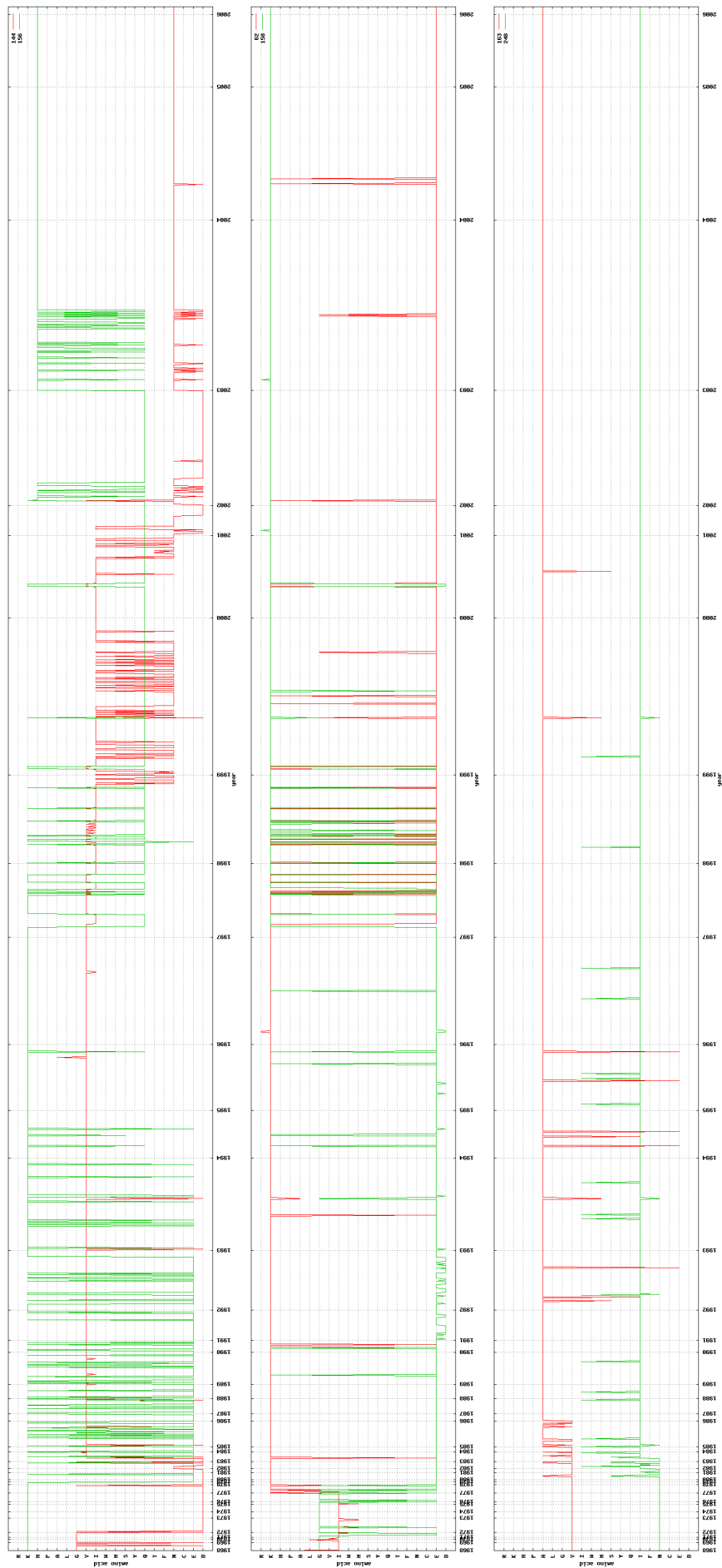


Figure 1: Correlated mutations found by the analysis. Series of amino acid changes in the positions 144-156(left), 62-158(middle), and 163-248(right) are shown.

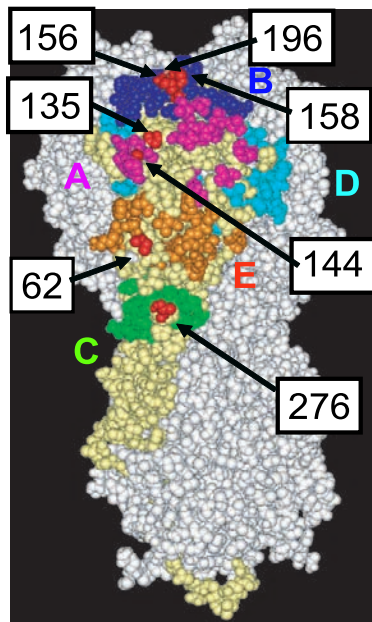


Figure 2: The 3D structure model of an HA molecule. The locations of amino acid positions that are involved in the correlated mutations in Table 1 are highlighted. The commonly used definitions of the five antigenic sites (A-E) are shown in different colors.

from the addition of an N-linked glycosylation site. The acquisition of new oligosaccharide chains is known to be significant for the antigenic drift of HA. Figure 3(a)(b) shows that the oligosaccharide chain that is added at position 246 produced a steric hindrance which likely required the substitution of valine at position 163 with the smaller amino acid alanine.

4. Retrospective Net Charge Analysis of Hemagglutinins

To study the change of the net charge of HAs, which may be associated with antigenic properties, we have also analyzed the HA1 subunits of influenza viruses isolated from human, swine, and avian hosts.

The net charge at neutral pH are calculated, assuming that glutamic and aspartic acid have -1 charge, and arginine and lysin have +1 charge at this pH. Figure 4 shows that the net charge of HA1s increased linearly from 1968 to 1984, while avian virus HAs have retained their net charge

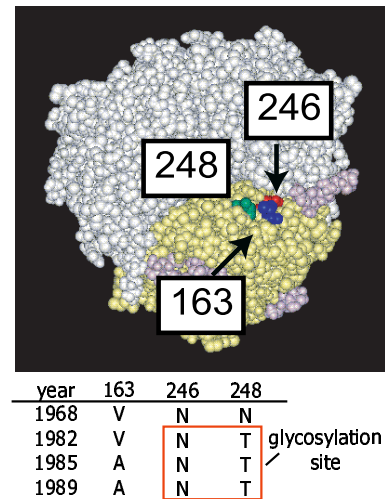


Figure 3: The correlated mutation among amino acids at 163, 246, and 248. (a) The 3D structure model of an HA. The locations of the amino acids at 163, 246, and 248 are indicated. (b) The amino acids in the glycosylation site during 1968 to 1989.

since 1963. This result suggests that the mutations that increased the charge of HA molecule were required for the adaptation of avian virus to human population. After 1984, the increase in the net charge in human virus HA has been saturated. This level of the net charge might be an upper limit for H3 HA to be functional. Correlated mutations that exchange the charges among two or more different amino acid positions were frequently found after 1984. These co-mutations include E82K/K83E(1989), N145K/G135E(1991), E135K&E156K/S133D&K145N&R189S-(1993), D124G/G172D&R197Q(1995), N145K/K135T(1996), E158K&N276K/-K62E&K156Q(1998), D271N/K92T(2000), T92K/N271D(2001), E83K&D144N&-W222R/R50G&N126D&G225D(2003), and D126N/K145N(2004). It suggests that these co-mutations contributed to counter balancing effect to keep the net charge of the HA.

For further investigation, we performed homology modeling to build 3D structure models of HAs of antigenically different influenza viruses (Figure 5). We found that antigenic drifts with frequent substitutions

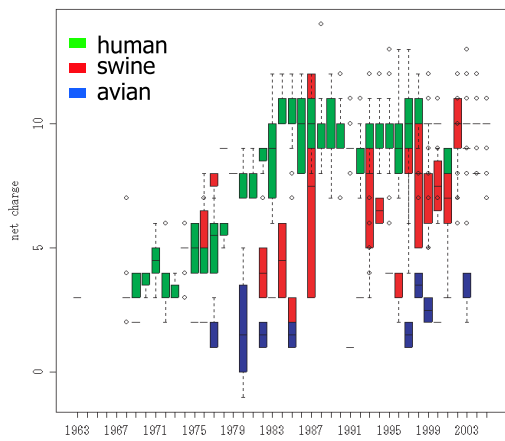


Figure 4: The net charges of H3 HA1 of influenza viruses isolated from human, swine, and avian.

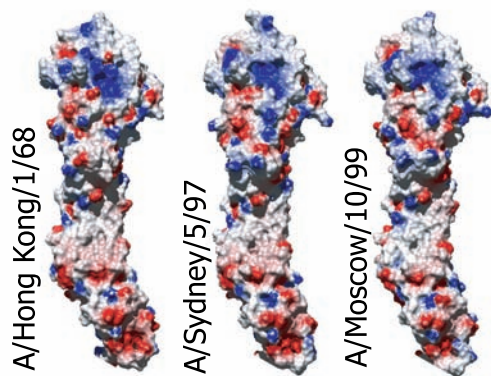


Figure 5: 3D structure model of three antigenically different HAs of human H3N2 viruses. The electrostatic potential on the surface of HAs are shown.

of charged amino acids resulted in the change of the charge distribution on the HA surface.

5. Conclusion

A number of correlated mutations in HAs are found by analyzing large scale sequence data of influenza viruses. A retrospective net charge analysis shows the increase of the net charge in HAs of human H3N2 viruses. It might be required for avian influenza viruses to increase the positive charge in order to adapt to human population. There seems to be an upper bound on possible net charge of hemagglutinin. Charge compensations have

been made by co-mutations since 1984 and these co-mutations might contribute to keep the net charge constant in HA molecules.

References

- [1] Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., Dress, A.W.: Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis, *Molecular Biology and Evolution* 17,pp.164-178 (2000)
- [2] Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J., Fitch, W.M.: Predicting the evolution of human influenza A, *Science*, 286(5446),pp.1921-1925.(1999)
- [3] Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., Derinovoy, D., Tatusova, T., Bao, Y., St, George, K., Taylor, J., Lipman, D.J., Fraser, C.M., Taubenberger, J.K., Salzberg, S.L.: Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution, *Nature*, 437(7062),pp.1162-1166 (2005)
- [4] Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22),pp.4116-4124, (2005)

The Links Between GIScience and Information Science

By Nigel Waters
Department of Geography
University of Calgary
2500 University Dr. NW
Calgary, Alberta, Canada T2N 1N4
e-mail; nwaters@ucalgary.ca
Tel: 403-220-5367 Fax: 403-282-6561

Introduction

This paper explores the links between GIScience and Information Science and in so doing suggests areas where future collaboration may be both warranted and productive.

1. The Rise of a Geographic Information Science

Geographic Information Systems (GIS) are a relatively new technology. The term was first used in 1968 by Roger Tomlinson [3]. Although the market for GIS technology has been growing steadily estimates of the size of the industry vary greatly. One recent report from industry assessment specialists, Daratech (<http://www.daratech.com/press/2004/-041019/>), suggested that the core revenue of the GIS industry in 2003 was USD \$1.84 billion and that it had grown at a rate of 5.1% over 2002. At the end of 2004 the core revenue was estimated to be USD \$2.02 with a corresponding annual growth rate of 9.7%. Core business revenue was stated to include software (64%), hardware (4% and declining), services (24%) and data products (8%). By any measure, in the early years of this decade, the GIS industry was doing well.

Yet in the early 1990s many academics had been less than satisfied with the current state-of-affairs in GIS research.

An argument was made for a new discipline of Geographic Information Science, or GIScience as it was soon renamed. This new discipline was promulgated in a paper by Goodchild [6] and leading GIS academic journals (including the one in which the paper was published) began renaming themselves with Science replacing Systems. It was argued by Goodchild that information science studies issues relating to the creation, handling, storage, analysis and use of information. Similarly Geographic Information Science, Goodchild suggested, should study those issues that arise from the creation, handling, storage, analysis and use of geographic or spatial information.

So the links are there, to a certain extent. With respect to Geographic Information Systems we may ask: how were the links exploited? With respect to Geographic Information Science let us ask: how will the links be exploited?

2. The Links Between Geographic Information Systems and Information Science

The early approach to GISystems implemented by Environmental Systems Research Institute (ESRI) in their pioneering ArcInfo software was to have one solution for handling the spatial data (Arc) and a second solution for handling the attribute data (the proprietary Info database) in their GISystem software. The Info part used traditional database technology that was imported directly from Information Science. Originally ESRI used their own database technology but links with commercial relational database technology soon followed.

Towards the end of the 1990s the relational database approach was supplemented by object-oriented solutions made available in the ArcGIS generation of products. The current state of the art is described in *Designing Geodatabases: Case Studies in GIS Data Modeling* by Arctur and Zeiler [1]. They identify ten steps in designing and building an exemplary geodatabase:

“Conceptual Design”:

1. Identify the information products that will be produced with your GIS.
2. Identify the key thematic layers based on your information requirements.
3. Specify the scale ranges and spatial representation for each thematic layer.
4. Group representation into datasets.

“Logical design”:

5. Define the tabular database structure and behavior for descriptive attributes.

6. Define the spatial properties of your datasets.
7. Propose a geodatabase design.

“Physical Design”:

8. Implement, prototype, review and refine your design.
9. Design workflows for building and maintaining each layer.
10. Document your design using appropriate methods.” [1]

Under step 7 Arctur and Zeiler suggest studying existing design for examples and ESRI provides a website that includes a collection of downloadable case studies that provide detailed geodatabase models (<http://support.esri.com/index.cfm?fa=downloads.dataModels.caseStudies>).

Most areas of application have their own distinct challenges and issues. Two of particular note are the transportation and atmospheric data models. Problems with the transportation data model arose even in the era of the use of relational database models. Two issues of concern were the need to develop linear referencing systems and the need for the ability to change data attributes, for example speed limits, along a single link in the network. The latter problem was resolved with the introduction of dynamic segmentation approaches ([17, 18], see also the ArcGIS Transportation Data Model for New York State available at <http://support.esri.com/index.cfm?fa=downloads.dataModels.caseStudies>).

One particular challenge remains and this is the handling of continuously varying data. This is a problem along transportation networks and makes for difficulties in calculating the spatial autocorrelation indices of, for example, accident data [2]. It is a problem for data that it is distributed over surfaces and even more so for continuously varying data in three

or even four dimensions, for example atmospheric data. This is a problem that is still not handled well as a data model in GIS (although the *analytical tools* are there in abundance). These concerns were noted by Mike Goodchild in a Key-note Address to the Annual GEOIDE Conference held in Banff June 1st, 2006. The problem is addressed from the perspective of scientific analysis of atmospheric data by Nativi et al. [10]). They note the difficulty in integrating atmospheric data within a GIS framework and suggest interim remedies in their article.

3. Selected Issues in Information Science

Recently Tanaka [15] has described some of the challenges facing Information Science with respect to developing a knowledge federation over the Web. He begins by noting that the Web may be seen as a depository of distributed intelligent resources. These resources include traditional server-client applications as well as mobile applications. Tanaka continues by defining pervasive computing as an open system of intelligent resources that a user can interact with in a seamless interoperable environment. Pervasive computing is assumed to link a wide variety of intelligent resources that are distributed in both virtual environments (the Web) and the geographical environment of the real world.

The ad hoc definition and implementation of pervasive computer in either or both environments is defined as a federation. The intelligent resources that form part of this definition are assumed to be in the form of documents. Suggested uses include scientific simulations, digital libraries and research activities. Tanaka [15] continues by noting that a federation of intelligent resources may be defined by programs that access such

resources using web service technologies (the autonomic approach) or they can be defined by users (the ad hoc approach). Noting that there is extensive literature exploring the former, the autonomic approach, he focuses on the latter, ad hoc methodologies. He thus advocates the use of meme-media technologies to extract intelligent resources from web documents and applications. The approach is similar to a structured query to a relational database but in this case the query is defined by HTML pathway expressions to Web documents and/or applications.

It will be seen below that there are many broad areas of application in a GIS environment where Professor Tanaka's approach will provide new research opportunities.

Spyratos [14] has presented a functional model for the analysis of large volumes of transactional data. The model presented used a data schema with an acyclic graph and a single root. An example was provided for product purchases from a store. Since the store is located in geographical space the example has obvious implications for GIS analysis in the field of Business Geographics [9]. Geographers refer to this as panel data.

Professor Tanaka and his colleagues at the Meme Media Laboratory have designed an interface to the Digital Library. The metaphor that has been used is that of a City. So that the user can move through a network of streets and then into a building, a room and then, finally, remove a book from a virtual shelf. Additional challenges have included providing a cell phone interface to the library.

Geographers have considered similar problems in trying to build digital librar-

ies of maps (<http://www.alexandria-ucsb.edu/>). The digital map library provides additional challenges in storing, referencing and accessing the spatial information. Tanaka [15] addressed problems with which geographers are also concerned. For example, he showed a movie in which fish were linked to a map. We have similarly experimented with the use of case-based reasoning software to link the resources of indigenous peoples in northern Canada with a GIS and map based information [4]. Similarly we have developed Visual Basic templates to input indigenous knowledge into a GIS.

Algorithms for searching the web for documents are a major area of research in information science. Several papers at the Third International Symposium concentrated on these technologies. These procedures are of interest to GIScientists because they too need faster, more efficient web search algorithms and because they use similar algorithms for the analysis of spatial and socio-economic data. Examples of new Web Document search algorithms are provided by Lamonova and Tanaka [8] and Poland and Zeugmann [11]. Both papers describe approaches that are of great interest to GIScientists. For example, similar fuzzy clustering algorithms to those developed by Lamonova and Tanaka are now being used to classify commuters in transportation planning models and the spectral approaches used by Poland and Zeugmann are of great interest to spatial analysts as are the neural nets and wavelet approaches mentioned by Lamonova in her presentation.

4. The Nexus of GIS and IS

The use of Tanaka's approach to the acquisition of documents over the Web has obvious implications for our Mapping the Media in the Americas (MMA) Project ([18] and www.mediamap.info for a complete description). The MMA Project is a collaboration between the University of Calgary, the Carter center in Atlanta and the Canadian Foundation for the Americas (FOCAL) in Ottawa. The three-year project began in September 2004 and will conclude in August, 2007. The Project seeks to portray and explain the impact of the media (TV, radio and newspapers) on the electoral process in 12 countries in the Americas. To-date we have completed research on Canada and three Latin American countries (Peru, Guatemala, and the Dominican Republic).

The research process involves scouring the internet for spatial and aspatial data on socio-economic variables for the countries question. We also need to map the locations of TV and radio antenna and the distribution points for newspapers. Spatial data on the results of the most recent national or federal elections are obtained. Although the research process usually begins with a visit to the capital city of each country in question, in addition, we search the Web intensively for appropriate resources. This research process would obviously be more efficient if the resources that we require were intelligent resources, as envisaged by Tanaka, that could be searched with more intelligent algorithms. In one sense our research aims to bring together the information in one location thus removing the need for such searches. Moreover, Tanaka's [15] C3W framework that allows the clipping of arbitrary HTML elements from web pages and then the subsequent pasting of

these clips on to a special C3Wsheet Pad would appear to have enormous potential for increasing the efficiency of our research process.

The data, once acquired, is assembled into a GIS. This GIS is then migrated to an interactive, GIS-based web server. For our work the GIS is built using ESRI's ArcGIS 9.1 (www.esri.com). Likewise we use an ESRI product, ArcIMS (Internet Map Server), to produce the web-based GIS. The final product may be found at www.mediamap.info. The splash screen for the website provides an introduction to the project and, like the rest of the site is available in two languages: Spanish and English for the Latin American countries and French and English for Canada.

One goal of the MMA Project is to show the impact of the media on the electoral process. This can be achieved, in part, through the use of visualization techniques. Thus we can map the results of the elections and we can map socio-economic data, and the locations of the media. We can visualize the results as individual layers or we can map two or more layers together. This is most effective where just one of the layers is a polygon layer and the additional layers are point files such as the locations of two or more different types of media. We have experimented with different types of shading for two different polygon layers e.g. solid colors that are overlain by hatched shading for the second variable. This, however, has met with limited success.

Essentially our work has brought together disparate data sets in an interactive, web-based GIS. We have achieved the goal suggested by Tanaka ([15], and see discussion above) but we have done this by physically locating the resources

on a single website rather than allowing the user to access the data on-the-fly from a federated knowledge network.

Our website does fulfill the third of Tanaka's goals, listed above, namely that of allowing for enhanced research activities. These may be carried out either by those campaigning for political office or by their party strategists. It is also a resource for statistical analysis for academics. Thus we have built Ordinary Least Squares regression models to predict electoral outcomes based on the socio-economic variables included in the website. As might be expected such models have met with limited success because political support and the independent variables that can be used to predict that support vary geographically. Geographers have proposed various solutions to this problem. These include the use of Spatial Regressions and Geographically Weighted Regression (GWR; [5]).

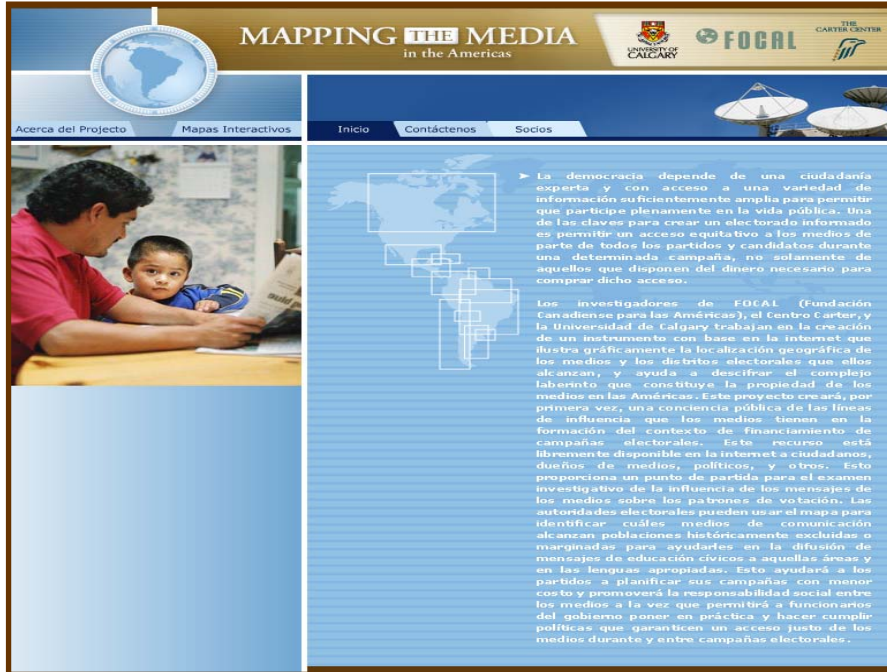


Figure 1: Splash Screen for the Mapping the Media in the Americas Project GIS-Based Interactive Website: www.mediamap.info

The original OLS model, using five independent variables that represented socio-economic status, education, literacy, and languages spoken, achieved goodness-of-fit statistics (R-squared) of only about 35%. When we used the GWR model, with a spatially varying kernel of about 100 points, the R-squared values ranged from 27% to 83%. Our work here is reminiscent of the work of Taniguchi and Haraguchi [16]. They noted that correlations in their global database were lower than in a local database. The result appeared to be due to temporal autocorrelation in the US Census data with which they experimented.

Essentially, our results are the spatial equivalent of their temporal autocorrelations. It would appear that a collaboration in this area would be mutually beneficial to both Information Scientists and GIScientists.

Since in the MMA Project we are concerned with the reach of the broadcast media we have been building models of wave propagation from radio antenna towers. Based on the height of the towers and by incorporating their power these broadcast models, in conjunction with digital elevation models, provide an estimate of the reach of the radio station in mountainous countries such as Peru. We combine these models with data on population distribution and language spoken from the census data incorporated into our web-based GIS and this then provides an estimate of the reach of the radio station. In turn this can be combined with information from past elections to determine if it is worthwhile for a political candidate to invest advertising resources in radio or TV broadcasts. Figure 3 shows one such propagation model from ReMartinez [13].

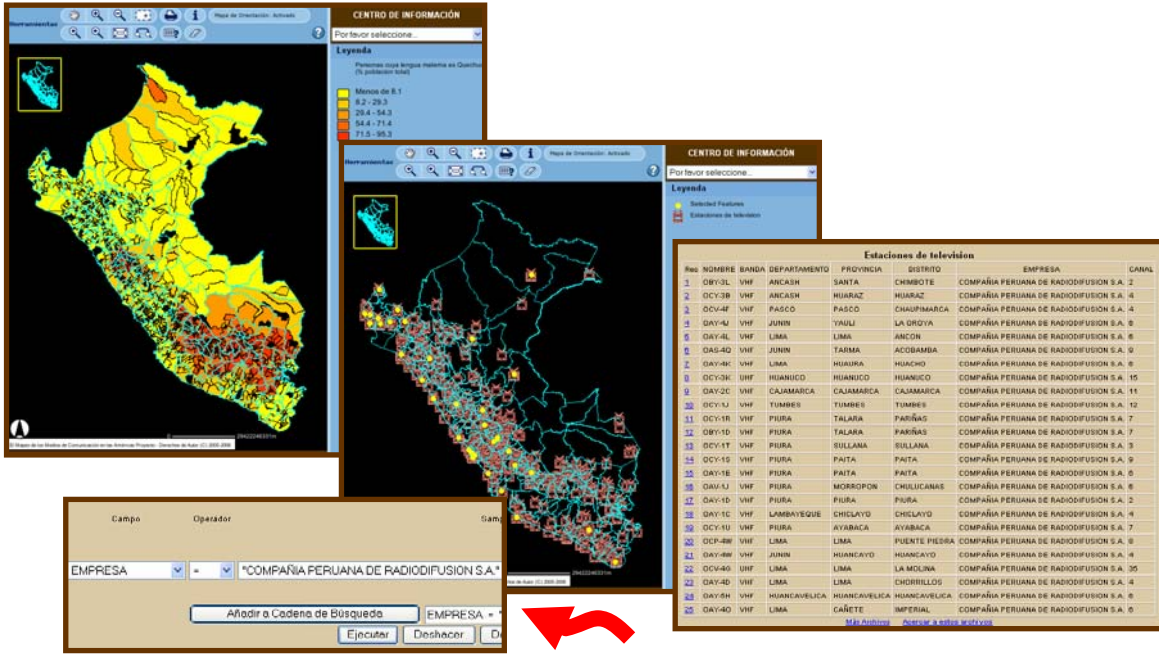


Figure 2: A Query to the TV Antenna Database for the Peruvian Website Showing How One National Television Network Does Not Reach an Area Where the Indigenous Quechuan Speaking Population Predominate.

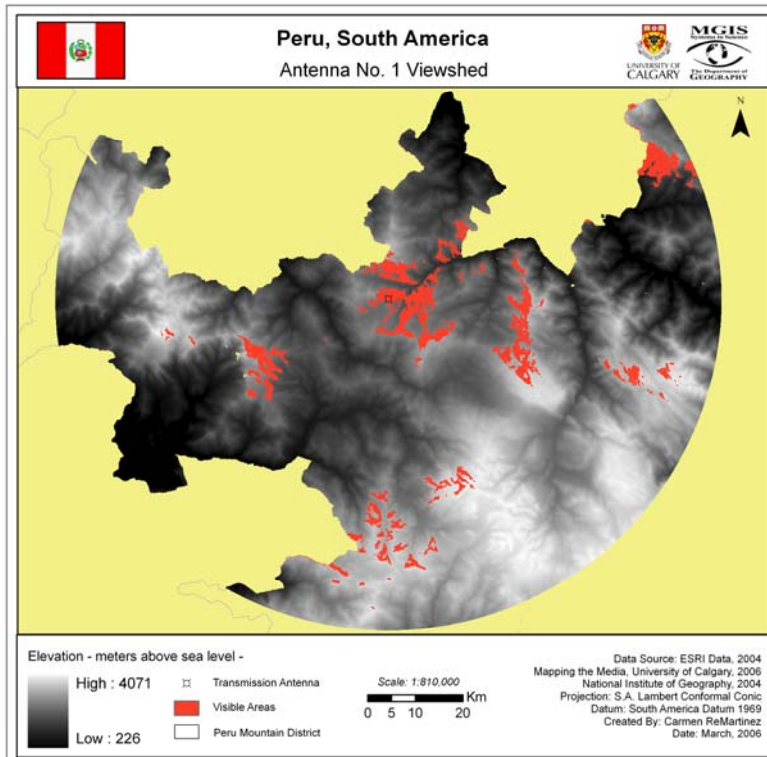


Figure 3: Antenna Viewshed [13]

concern over the difference in performance in line-of-sight and nonline-of-sight environments was the subject of Ogawa's [12] presentation at the Third International Symposium on Ubiquitous Knowledge: Network Environment. The possibility for synergy between work in GIScience and the work of Ogawa should be apparent. Spyratos' proposed data model, discussed above, produces panel data. Such data, while difficult to acquire, has long been of interest to retail geographers and those developing business applications [19, 20]).

5. Future Directions

One of the next steps in the integration of Information and Geographical Science, especially as the former was represented at Professor Tanaka's Third International Symposium on Ubiquitous Knowledge: Network Environment, would be to extend his C3W framework

to other markup languages. Specifically GIScience practitioners would want to extend Tanaka's approach to include the Geography Markup Language (GML). This is even more important now that on June 5th, 2006, the Open Geospatial Consortium has approved and released a simple features profile specification for GML (<http://www.opengeospatial.org/press/?page=pressrelease&year=0&prid=260>).

One area of future research in GIScience appears to be in the use of interactive web-based GIS for decision support. In 2005 we began a GEOIDE supported research project for *Promoting Sustainable Communities Through Participatory Spatial Decision Support* (http://www.geoide.ulaval.ca/files/17_E.jpg). Our research involves the Town of Canmore in Alberta, Canada. We are

building a GIS that will include key variables that of interest to both the

planners and the citizens in the town. These GIS maps will then be entered into an interactive GIS based website using one of two technologies: MapChat or ArguMap. The latter is already being used for a similar research experiment at University College London (http://ernie.ge.ucl.ac.uk:8080/Web_System/faces/2_Results.jsp?argumap=yes). One of the components of any such system might be the ability to bring in new material from the Web and to use in subsequent analysis. Meme media extended to geographical objects would be extremely useful, as would fast internet search algorithms.

The research that Jantke [7] presented at the Meme Media workshop in Sapporo in early 2006 included a discussion of how a movie such as *The Kingdom of Heaven* could have digital tags attached to items in the screen shot. These tags allow the viewer to access additional information relation to the tagged object. This type of technology could be used effectively within an environment such as MapChat or ArguMap. Thus the planner or citizen working within the interactive environment could query an object in the GIS or perhaps on a Google Earth or orthophoto backdrop and be provided with additional information that would lead to a more informed planning choice.

6. Conclusion

This paper has attempted to show the links between GIScience and Information Science. It began with a short introduction to Geographic Information Systems and then discussed the rise of GIScience as a distinct discipline. It then provided a small subset of examples that show some of the more intriguing links between the GIScience research that is being conducted at the University of Calgary and the Information Science at

the Institute for Media and Information Science at Ilmenau in Germany and at the Meme Media Lab at the University of Hokkaido in Sapporo.

References

- [1] D. Arctur and M. Zeiler. *Designing Geodatabases: Case Studies in GIS Data Modeling*. ESRI Press, Redlands, CA., 2004.
- [2] W. Black. and I. Thomas. Accidents on the Belgium Motorway System: A Network Autocorrelation Analysis. *Journal of Transport Geography*, 6, 23-31, 1998.
- [3] N. Chrisman. *Charting the Unknown: How Computer Mapping at Harvard became GIS*. ESRI Press, Redlands, CA, 2006.
- [4] D. Clayton and N. Waters. Distributed Knowledge, Distributed Processing, Distributed Users: Integrating Case Based Reasoning and GIS for Multicriteria Decision Making. Chapter in *Multicriteria Decision-Making and Analysis: A Geographic Information Sciences Approach*, edited by Jean-Claude Thill, Ashgate, Brookfield, USA, 275-308, 1999.
- [5] A. S. Fotheringham, C. Brunson and M. Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester, UK, 2002.
- [6] M. F. Goodchild. Geographical Information Science. *International Journal of Geographical*

- Information Systems*, 6, 31-45, 1992.
- [7] K. Jantke. Media Integration for Film in the Digital Bibliothek City. Paper presented at the Meme Media Workshop, associated with the 21st Center of Excellence Program in Information, Electrics and Electronic at Hokkaido University, Sapporo, February 28 – March 1, 2006.
- [8] N. Lamonova and Y. Tanaka. A Robust Probabilistic Fuzzy Clustering Algorithm with Application to Web Document. Paper presented at the Meme Media Workshop, associated with the 21st Center of Excellence Program in Information, Electrics and Electronic at Hokkaido University, Sapporo, February 28 – March 1, 2006.
- [9] P. A. Longley and G. Clarke. GIS for Business and Service Planning. Wiley, New York, 1996
- [10] S. Nativi, M. B. Blementhal, J. Caron, B. Domenico, T. Habermann, D. Hertzmann, Y. Ho, R. Raskin and J. Weber. Differences Among the Data Models Used by the Geographic Information Systems and Atmospheric Science Communities, 2003. Retrieved from <http://support.esri.com/index.cfm?fa=downloads.dataModels.caseStudies>
- [11] J. Poland and T. Zeugmann. Spectral Clustering of the Google Distance. Paper presented at the Meme Media Workshop, associated with the 21st Center of Excellence Program in Information, Electrics and Electronic at Hokkaido University, Sapporo, February 28 – March 1, 2006.
- [12] Y. Ogawa. Multiple Antenna Systems and Their Applications. In Proceedings of the Third International Symposium on Ubiquitous Knowledge: Network Environment, February 27 – March 1, 2006, Sapporo Convention Center, Sapporo, Japan, pp. 143-151, 2006.
- [13] C. ReMartinez. A GIS Coverage Broadcast Prediction Model for Transmitted FM Radio Waves in Peru. MGIS Research Project, Department of Geography, University of Calgary, Calgary, Alberta, Canada, 2006.
- [14] N. Spyrtos. A Functional Model for Data Analysis. In Proceedings of the Third International Symposium on Ubiquitous Knowledge: Network Environment, February 27 – March 1, 2006, Sapporo Convention Center, Sapporo, Japan, pp. 37-48, 2006.
- [15] Y. Tanaka. Knowledge Federation Over the Web Based on Meme Media Technologies. In Proceedings of the Third International Symposium on Ubiquitous Knowledge: Network Environment, February 27 – March 1, 2006, Sapporo Convention Center, Sapporo, Japan, pp. 13-36, 2006.
- [16] T. Taniguchi and M. Haraguchi. Discovery of Implicit Correlations Based on Differences of

- Correlations. Paper presented at the Meme Media Workshop, associated with the 21st Center of Excellence Program in Information, Electrics and Electronic at Hokkaido University, Sapporo, February 28 – March 1, 2006.
- [17] N. Waters. Transportation GIS: GIS-T. In *Geographical Information Systems: Principles, Techniques, Applications and Management*, 2nd Edition, Edited by Paul Longley, Michael Goodchild, David Maguire, and David Rhind, John Wiley and Sons, pp. 827-844, 1999.
- [18] N. M. Waters. Transportation GIS: GIS-T. In *Geographical Information Systems, Second Edition, Abridged*, edited by Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W.; Wiley, New York, 2005.
- [19] N. Wrigley. *Categorical data Analysis for Geographers and Environmental Scientists*. Longman, London, 1985.
- [20] N. Wrigley and M. S. Lowe. *Reading Retail: A Geographical Perspective on Retailing and Consumption Spaces*. Arnold, New York, 2002.

Storyboarding – An AI Technology to Represent, Process, Evaluate, and Refine Didactic Knowledge

Rainer Knauf¹ and Klaus P. Jantke²

Technical University Ilmenau

¹Faculty of Computer Science and Automation

Chair of Artificial Intelligence

²Institute of Media and Communication Sciences

PO Box 100565

98684 Ilmenau

Abstract

The current state of affair in learning systems in general and in e-learning in particular suffers from a lack of an explicit and adaptable didactic design. Students complain about the insufficient adaptability of e-learning to the learners' needs. Learning content and services need to reach their audience properly. That is, according to their different prerequisites, needs, and different learning conditions. After a short introduction to the storyboard concept, which is a way to address these concerns, we present an example of using storyboards for the didactic design of a university course on Intelligent Systems. In particular, we show the way to express didactic variants and didactic intentions within storyboards. Finally, we sketch ideas to for a machine supported knowledge processing, knowledge evaluation, knowledge refinement, and knowledge engineering with storyboards.

1. Introduction

Successful university instructors are often not those with the very best scientific background or outstanding research results. The most successful ones are typically those that successfully utilize didactical experiences as well as 'soft skills' in dealing with other actors in the teaching process. Besides the students and colleagues, such actors include e-learning systems as well as the large amount of active (desirable and undesirable, conscious and unconscious) 'content presenters' that include news, web sources and advertisements.

The design of learning activities in collegiate instruction is a very interdisciplinary process. Besides deep, topical knowledge in the subject being thought, an instructor needs knowledge and skills in many other subjects. This includes IT-related skills to

use today's presentation equipment, didactic skills to effectively present the topical content, plus skills in fields like social sciences, psychology and ergonomics.

University instruction, however, often suffers from a lack of didactic knowledge. Since universities are also research institutions, their professors are usually hired based on their topical skills. Didactic skills are often underestimated in the recruiting process. At German universities, e.g., there is no didactic education required as a prerequisite for a professorship. Such skills are only checked by asking (usually just two) students about their impression on this issue after watching a single talk given by the applicant.

Here, we refrain from further discussing reasons for that, but focus the issue of involving it a little more. We propose an approach to

improve the didactic content within the large variety of skills needed for successful teaching by discussing questions like

- How to include didactic variants of presenting materials?
- What determines the variants?
- How to choose an appropriate variant according to particular circumstances?

Our approach to facing problems like these is a modeling concept for didactic knowledge called Storyboarding. A storyboard, as the name implies, provides a roadmap for a course, including possible detours if certain concepts to be learned need reinforcement. Using modern media technology, a storyboard also plays the role of a server that provides the appropriate content material when deemed required. Our suggestion to ensure a wide dissemination of this concept is to use a standard tool to develop and process this model, which is MicrosoftTM Visio. Here, we present an application of this approach for a particular university course.

The paper is organized as follows. Section 2 is a short introduction to the storyboard concept as introduced in [1]. Section 3 introduces an exemplary storyboard. Because we used ourselves as the very first ‘experimental subjects’ we selected an AI course at the University of Central Florida as our example. In this section, we show the opportunities of storyboards to express didactic intentions and variants. Section 4 outlines some conceptual refinements towards a machine supported knowledge processing with storyboards. In section 5, we summarize the research undertaken so far and outline our future research in three time horizons, (1) the short term objective of storyboard dissemination, (2) the medium term objective of storyboard evaluation, and (3) the vision of identifying and finally utilizing successful didactic patterns.

2. Storyboarding - the concept

The basic approach behind storyboarding is that teaching consists of further structured episodes and atomic scenes (with no meaningful structure) - just like traditional story-

boards on shows, plays, or movies. The material of the storyboarded learning activities (e.g., text books, scripts, slides, models) is something comparable to the requisites of a show.

Basic differences of our storyboards to those used to ‘specify’ a show are:

1. the primary purpose (learning vs. entertainment)¹,
2. the degree of formalization, and, as a consequence of being semi-formal, and
3. the opportunity to formally represent, process, evaluate, and refine our storyboards.

Our storyboard concept [1] is built upon standard concepts which enjoy (1) *clarity* by providing a high-level modeling approach, (2) *simplicity*, which enables everybody to become a storyboard author, and (3) *visual appearance* as graphs.

A storyboard is a nested hierarchy of directed graphs with annotated nodes and annotated edges. Nodes are scenes or episodes; scenes denote leaves and episodes denote a sub-graph. Edges specify transitions between nodes. Nodes and edges have (pre-defined) key attributes and may have free attributes. These terms are described below. Figure 1 shows an example storyboard on the present paper. The representation as a graph (instead of a linear sequence of sections) reflects the fact that different readers trace the paper in different manners according to their particular interests, prerequisites, a current situation (like being under time pressure), and other circumstances. For example,

- members of our research group will skip the sections 1 and 2, because they are already motivated and familiar with the concept,
- reviewers of this paper, on the other hand, will (hopefully) read the complete paper, but they might skip the *Acknowledgement* section, because its content doesn’t matter for their work.

¹There is no ambivalence between these purposes. To include some entertainment into learning is one of the keys of successful learning and thus, also an ultimate objective of storyboarding learning processes.

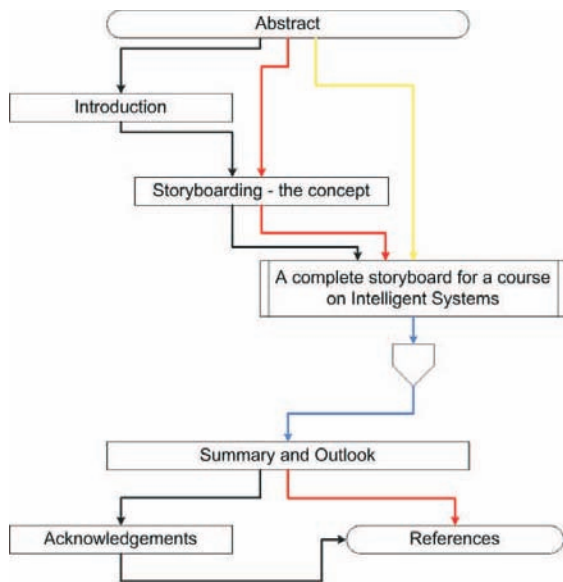


Fig. 1: An exemplary storyboard

So a storyboard can be traversed in different manners according to (1) users' interests, objectives, and desires, (2) didactic preferences, (3) the sequence of nodes (and other storyboards) visited before (i.e. according to the educational history), (4) available resources (like time, money, equipment to present material, and so on) and (5) other application driven circumstances.

A storyboard is interpreted as follows:

- *Scenes* denote a non-decomposable learning activity, which can be implemented in any way: It can be the presentation of a (media) document, or an informal description of the activity. There is no formalism at and below the scene level.
- *Episodes* denote a sub-graph.
- *Edges* denote transitions between nodes.
- *Key attributes* specify actors and locations. Depending on the application, more key attributes may be defined.
- *Free attributes* can specify whatever the storyboard author wants the user to know: didactic intentions, useful methods, necessary equipment, and so on.
- (Key and free) attributes to nodes are inherited by each node of the related super-graph.

In fact, the storyboard is a semi-formal knowledge representation for the didactics

of a teaching subject and thus, a firm base for processing, evaluating and refining this knowledge.

Not all the following supplements are really necessary. In fact, the following list is a proposal of further supplements. Conversely, some of the features might not be applicable to particular storyboards. Many of them are implicit in the general concept. We discuss those details only for the readers' convenience, to become a little more familiar with our ideas, aims and intuition. In *italics* we provide details about our current technological representation of storyboards in MicrosoftTM Visio [3]. Readers may use any other appropriate tool at hand.

- Because those nodes called episodes may be expanded by sub-graphs, storyboards are hierarchically structured graphs by their nature. *Double clicking on an episode opens the corresponding sub-graph on a separate sheet.*
- Comments to nodes and edges are intended to carry information about didactics. Goals are expressed and variants are sketched. *Clicking to a comment opens a window with the text, including author information and date.*
- As far as it applies to a node, educational meta-data, such as a degree of difficulty (e.g., basic or advanced) or a style of presentation (e.g., theory-based or illustrated) may be added as key attributes. *Visio built-in object properties are used to represent general information and meta-data.*
- Edges are colored to carry information about activation constraints, conditions, or recommendations to follow them. Certain colors may have some fixed meaning like usage for certain educational difficulties. *Clicking on edges opens didactic comments and meta-data for adaptive behavior.*
- Actors and locations, including those in the real world, are assigned to scenes only. *Through programming, actor and location information may be propagated automatically.*

- Certain scenes represent documents of different media types like pictures, videos, PDF files, Power Point slides, Excel Tables, and so on. *Double clicking on a scene opens the media object in a viewer, e.g., plays the film.*

Clearly, the sophistication of storyboards can go very far. The concept allows for deeply nested structures involving different forms of learning, getting many actors involved and permitting a large variety of alternatives. Though this is possible, in principle, the emphasis of this concept – driven by the goal of dissemination – is on simple storyboards designed quickly by almost anyone.

For the intended purpose of storyboarding a university course, we developed storyboards that contain (besides the Scenes and Episodes) additionally

- *To Do* –s, which define anything to do for the final grade, and
- *Off Page References*, which are points to jump back and forth between sub– and related super–graphs.

Each node type has different meanings and behaviors. In MicrosoftTM Visio, so called hyperlinks can be defined on any graph object to open either a local file of any media type with the appropriate tool or to open the standard browser with a specified URL or mail tool if it is an e-mail address. We made use of this opportunity for the *Scenes*, *Episodes* and the *To Do* –s.

In particular, the nodes, their behavior and their key attributes are as specified in Tables 1, 2, 4, and 3.

For edges, it is not meaningful² to define double click actions or hyperlinks. In our storyboard, they have exactly one key attribute: a field, where the author can specify a condition to follow this edge. Edges may have different colors. A storyboard author is free in the choice of colors. There

²Also the edges are not intended to carry topical subject content, but didactics of a (mandatory, conditioned, or recommended) switch between the nodes of the graph.

is only one requirement to meet: the colors must mean something. Wherever a new color is introduced for several edges going out of one node, it must be noted as the ‘key attribute’ which color of the upcoming path is recommended under which conditions.

The storyboarding practice of the authors indicated that introducing new colors is useful, if the ‘fork-situation’ continues for a path of several nodes. In case both ways merge back to one after visiting one node, we did only mark the condition as a key attribute, but did not use a new color and did not see a lack of visual overview and clarity. We feel the opposite is true; too many colors cause lack of overview.

3. A complete storyboard for a course on Intelligent Systems

Here, we outline a storyboard that we developed for a course on Intelligent Systems at the University of Central Florida.

3.1. Resources initially available

At the starting point of the storyboard development, we had the course material and “experience sources” as used so far. This included:

- a course syllabus including references, grading rules, class policy, and a tentative schedule as a linear sequence of topics over the semester,
- the books referenced in this syllabus,
- a huge amount of slights (MicrosoftTM Power Point files) for each part, and
- two individuals with some topical background: the lecturer and author of this paper.

In fact, we intentionally storyboarded a course on our own (teaching and research) subject to gain some experience with the use, evaluation and refinement of particular courses as well as with the concept itself.

3.2. Course structure

Figure 2 shows the top level storyboard of the course.


Symbol	
Behavior when double clicked	<ul style="list-style-type: none"> opening a material document nothing, if just verbally described scene
Behavior when following a hyperlink	<ul style="list-style-type: none"> opening a material document visiting a website with the standard browser, if URL opening the standard mail tool, if it is an e-mail address
Key annotations	
Scene	scene name: <i>free text</i>
Key words	key words: <i>free text</i>
Educational difficulty	degree of complexity: <i>any</i> <i>basic</i> <i>advanced</i>
Educational presentation	style: <i>any</i> <i>theory based</i> <i>illustrated</i>
Media	media documents needed: <i>free text</i>
Unspecified media	media documents to be specified: <i>free text</i>
Actors	human actors: <i>instructor</i> <i>students</i> <i>both</i>
# slides	# of slides: <i>integer</i>
Estimated time consumption	approx. time needed: [h]:[min]
Location	location and equipment requirements: <i>free text</i>

Table 1: Scenes


Symbol	
Behavior when double clicked	opening the sub-graph that specifies the episode
Behavior when following a hyperlink	<ul style="list-style-type: none"> opening a material document visiting a website with the standard browser, if URL opening the standard mail tool, if it is an e-mail address
Key annotations	
Episode	episode name: <i>free text</i>
Key words	key words: <i>free text</i>
Educational difficulty	degree of complexity: <i>any</i> <i>basic</i> <i>advanced</i>
Slides	slide file in the conventional course: [file name].ppt
Equipment	equipment needed for the episode: <i>free text</i>
Media	media documents needed: <i>free text</i>
Unspecified media	media documents to be specified: <i>free text</i>

Table 2: Episodes



Symbols	 
Behavior when double clicked	jumping back and forth
Behavior when following a hyperlink	<ul style="list-style-type: none"> opening a material document visiting a website with the standard browser, if URL opening the standard mail tool, if it is an e-mail address
Key annotation:	name of sub- or super/graph: <i>free text</i>

Table 3: Off-Page References

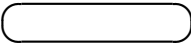
Symbol	
Behavior when double clicked	<ul style="list-style-type: none"> • opening a material document • nothing, if just verbally described scene
Behavior when following a hyperlink	<ul style="list-style-type: none"> • opening a material document • visiting a website with the standard browser, if URL • opening the standard mail tool, if it is an e-mail address
Key annotations	
Type	type of activity (homework #, midterm exam, final exam, ...): <i>free text</i>
Topics	topic of homework or exam: <i>free text</i>
Material allowed for students	things allowed to use (calculator, dictionary, ...): <i>free text</i>
Material needed by instructor	things needed by instructor (docs with tasks, ...): <i>free text</i>
Relative worth	% of worth w.r.t. the final grade: <i>free text</i>
Date	scheduled date and time for activity: <i>free text</i>
Location	location of activity (home, seminar room, ...): <i>free text</i>
Maximal time consumption	max. time given for activity: [h]:[min]

Table 4: ToDo-s

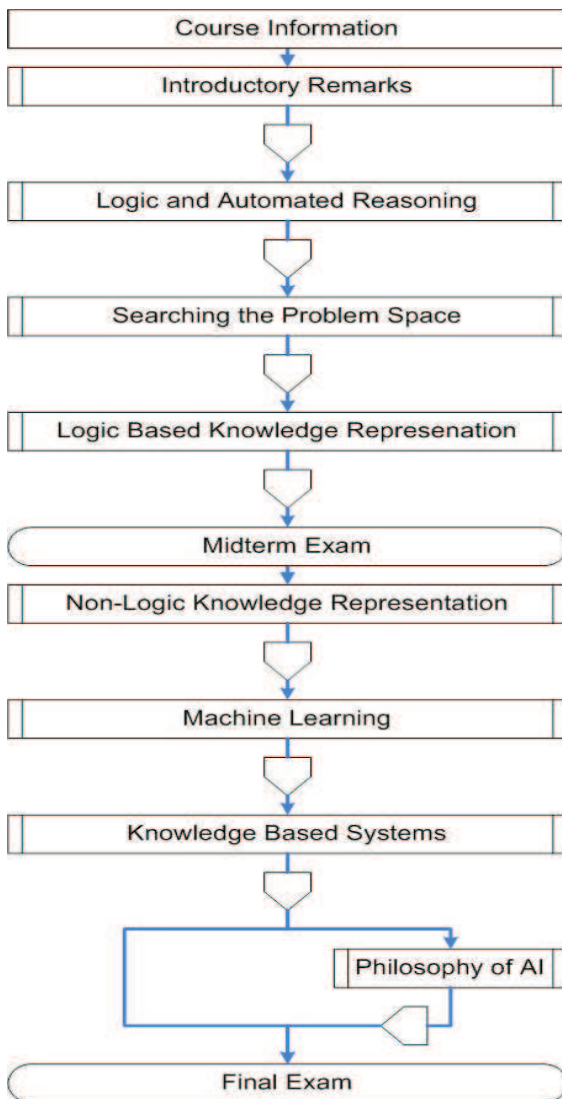


Fig. 2: Top level storyboard of the course

At this level, there are not many alternative paths. In our experience in building storyboards, the top level storyboard of a subject to be taught is usually of this kind. The storyboards below this level contain alternative paths according to some ‘tactical variants’ of teaching (like a recommended inclusion of an example). The storyboards above this level contain alternative paths according to some strategy of teaching (like recommended subject combinations).

To show the opportunities of including didactic variants and intentions, we next have a closer look at the 2nd level storyboard on *Machine Learning* respectively selected storyboards below this level.

3.3. Examples from the Machine Learning chapter

Two sub-graphs of this chapter are dedicated to *Case Based Reasoning* (CBR, for short) and *Inductive Inference*.

The first one (CBR, see Figure 3), shows a little more structure than the top level. Here, different paths are defined depending on a tactical decision: If the instructor realizes that the CBR process is well understood, he can continue with search technologies in *Case Libraries*. Otherwise he needs to include the example before doing that. Both

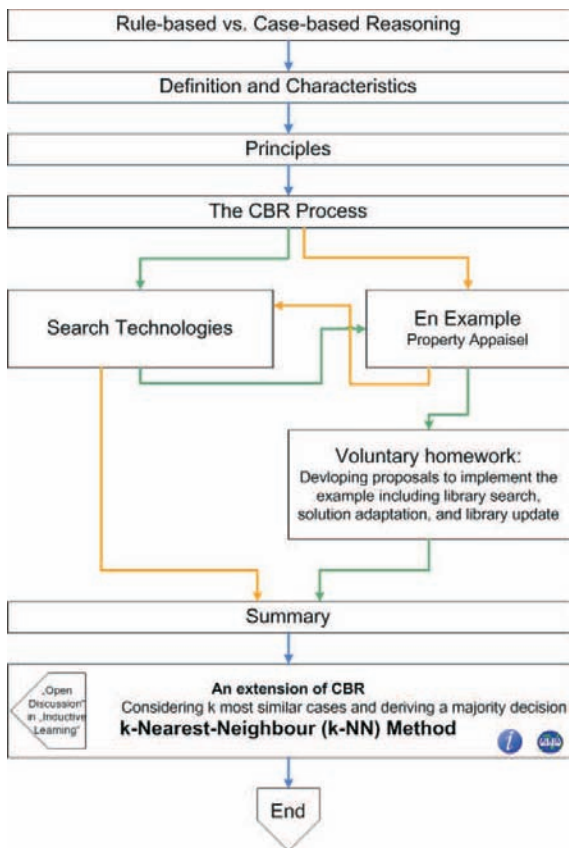


Fig. 3: Storyboard on Case Based Reasoning

paths are distinguished by different colors and by a description of conditions to follow the one or the other as an annotation to the respective edge at the ‘fork point’ of these different ways to continue. Moreover, for the ‘advanced students’ who follow the green path, an additional optional node is visited. Since the *k-NN Method* (see last scene in Figure 4) is some sort of CBR, which overcomes some drawbacks in Inductive Inference (which is another 3rd level sub-graph besides CBR), there is a reference back and forth. Depending on whether this scene is visited as a supplement to Inductive Inference or as a regular part of CBR, the related *Off-Page Reference* needs to be clicked to jump back.

A nice example for including didactic intentions can be seen in the storyboard of Inductive Inference (see Figure 4). Here, the simple play ‘guess my number’ is intended to give the students an idea of what the term *Information Entropy* means, because this concept is utilized in the subsequent scene. An instructor might decide whether such a (en-

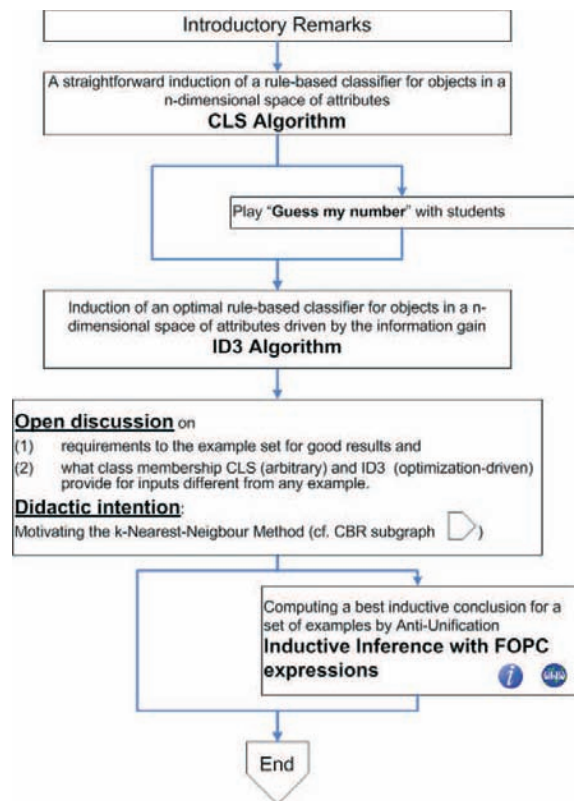


Fig. 4: Storyboard on Inductive Inference

tertaining) bridge towards the next topic is appropriate. Moreover, the open discussion on requirements the Example Set in *Inductive Inference* should meet for *CLS* and *ID3* and the impact of these sets on the quality of inference results, if they are not met properly, leads directly to a concept to address these drawbacks: the *k-NN Method*. Since *k-NN* is not a method of *Inductive Inference*, but closely related to CBR, an *Off-Page Reference* to the related node in the CBR storyboard is placed here.

A scene that hasn’t been a part of the *Machine Learning* episode in the course before its storyboarding, is *Inductive Inference with FOPC expressions* (see Figure 4), which is a technology to compute a so-called *best inductive conclusion* on a set of expressions of the *First Order Predicate Calculus (FOPC)*. This is a technology that is not presented in any AI textbook, but included in a related course of the first author. By including it as an optional scene in the second author’s course on Intelligent Systems, we utilized the storyboard as a medium of collaboration in teaching *Artificial Intelligence*. As the double click action, we implemented the

opening of a MicrosoftTM Power Point presentation to this topic as used in the first author's course³. Furthermore, clicking the www-symbol in the lower left corner of the scene opens the webpage of the first author, at which all slides and scripts of the first author for teaching AI can be downloaded by interested users (instructors or students). Additionally, there is an information symbol *i* in the lower left corner, which indicates a source of further information on this learning content. What it does is open the standard mail tool of the PC under use with the recipient address of somebody who is able to answer questions on this learning content: the first author.

A similar use of the storyboard for sharing teaching material has been done with the scene on the *CLS Algorithm*. Here, two hyperlinks are defined, one that opens the original slides of the course on Intelligent Systems and another one that opens (an English translation of) the slides for the same topic in the first authors' course on *Artificial Intelligence*. Generally, the number of hyperlinks is unlimited, so that a scene can carry many different material to provide its content, including web addresses.

Another didactic measure that helps to realize the nature of the various Machine Learning technologies on the very first view is the labeling of the scenes (1) *CLS Algorithm*, (2) *ID3 Algorithm*, and (3) *Inductive Inference with FOPC expressions*. By the comments above these scenes are titled

- (1) *A straightforward induction of a rule-based classifier in a n-dimensional space of attributes*
- (2) *Induction of an optimal rule-based classifier for objects in a n-dimensional space of attributes driven by the information gain, respectively*
- (3) *Computing a best inductive conclusion for a set of examples by Anti-Unification*

The storyboard user (instructor and student) receives some knowledge on how these three

topics are related to each other. Providing this information in conventional course material exclusively (e.g., books, scripts, slides), which is sequentially organized, bags the risk, that the user doesn't recognize that this information is some sort of meta-knowledge about these three technologies. The nesting of storyboards supports implementing such knowledge, effectively.

4. Current efforts towards formalizing the concept

With the objectives of

1. a software realization of *knowledge processing* (i.e. deductive inference on storyboards),
2. *evaluating* storyboards by case-based validation technologies such as described in [2],
3. *machine learning* (i.e. the identification of didactic patterns by some sort of data mining and inductive inference),
4. *knowledge refinement* by improving particular storyboards due to didactic insights that became explicit by storyboarding, and finally
5. *knowledge engineering* by supporting the storyboard development with a toolbox of didactic patterns that are proven to be appropriate in former use,

the overall concept is currently under revision.

In opposition to classical knowledge processing in Artificial Intelligence, deduction on the knowledge modelled by storyboards is not 100 % performed by machines. Albeit the traversing can be software-guided, there are some decisions left to humans respectively depend from an input parameter that might be provided by humans. Furthermore, the knowledge within (and below) the Scenes is usually informal and thus, absconds itself from formal processing.

However, we made some conceptual refinements that improve the opportunities for a software supported (deductive) knowledge processing as a very first step towards the first one of the visionary objectives above.

³Of course, this is a version translated in English.

4.1. Conceptual refinements of nodes

A view at figure 3 reveals that jumping back into a related super-graph is not uniquely defined, formally. For better opportunities to support the storyboard traversing by an appropriate software, we additionally limited the expressiveness by requesting, that the graph hierarchy has to be a tree of graphs. In other words, a graph that implements an episode can only be a sub-graph of exactly one super-graph. This implies, that, whenever a sub-graph has been completely traversed by reaching its final node, there is a unique super-graph into which the storyboard interpreter has to jump back.

Also, the definition of exactly one starting point and exactly one end point is necessary to interpret graphs without any elbowroom for human interpretation respectively misunderstandings. Other references between graphs than start- and end-nodes have to be forbidden to enable a software supported storyboard traversing.

4.2. Conceptual refinements of edges

Since humans are still an essential part of a storyboard interpreter, the modelling language must be easy to read and to understand. So we shifted the representation of conditions to follow an edge from the (rather hidden) key annotation of it to an obvious annotation within the arrow that represents the edge.

To ensure decidable conditions for edges, we limit these conditions to

- dynamic conditions regarding the traversing history (visited nodes, e.g.) and
- input parameter with case-specific values that don't change after their evaluation.

Furthermore, if node-outgoing edges carry a condition, there have to be at least two of those to avoid dead ends of traversing.

Fork situations If there are several edges that leave a node, three cases have to be dis-

tinguished: (1) they are alternatives that exclude each other, (2) they are edges that need to be traversed concurrently, and (3) there is a rule that determines something in-between both ('follow at least three of 10 edges' e.g.).

These cases are syntactically distinguished in the way these edges branch and merge after finishing the alternative or concurrent paths: (1) alternative paths start and end with different edges (2) concurrent paths start and end with one edge that spreads at the branch point and re-unites at the merge point in the graph. (3) The third case is treated as a special case of the second one by expressing the rules at the branch and merge point of the paths.

Edge coloration Additionally, rules of coloring edges have been established. So far, a color specifies a path to follow under a certain condition. This raises a problem, if there are several edges of different colors towards a node and also several edges of different colors leaving a node. So the color concept so far was not able to express the rules how to determine possible outgoing colors depending on the incoming color.

By introducing the concept of bi-colored edges that change their color within the edge, these rules can be expressed in a very simple and formally decidable way: All edges having the same color as the incoming edge, are the only possible outgoing edges.

5. Summary and outlook

Storyboards in general, and the one introduced in this paper in particular, are an approach to make the didactic design of university courses explicit. Since their scenes are not limited to the presentation of electronic material, but may represent *any* learning activity, the application of this concept goes far beyond the IT approaches to support learning so far.

The idea to represent knowledge at a high level with a modeling concept that is appropriate to be used by topical experts (university instructors, in this case) without the

need of an IT- or even software technological background is very much AI-driven. Here, the term ‘topical knowledge’ is not related to the learning content, but to its didactics instead. In particular, didactical intentions and variants can be specified as a nested graph-structure.

To validate the usefulness in practice, we developed a storyboard on an AI related course at the second author’s university. After its development, we presented the result to the current instructor of this course (as well as other university instructors). There was, in fact, no need to convince him in using this storyboard; he was quite happy to have such a useful tool to support his teaching.

One essential property of this concept and its implementation is its simplicity in terms of both the concept itself and the tool we used to implement it. Everybody, also university instructors of subjects that are far removed from information technology, are able to develop storyboards.

Of course, we won’t stop this research after having a high-level modeling concept for didactic design and asking university instructors to perform their ‘knowledge processing’ with it.

Our upcoming research on this issue is as follows:

1. A *short term objective* is, of course, promoting the development and use of this concept.
2. After that, as a *medium term objective*, we plan to develop an evaluation concept for storyboards based on the learning results of the students as acquired from the final grade they achieve for the storyboarded courses as well as the students’ specific comments in a questionnaire.
3. Our *long term objective* is to identify typical didactic patterns of successful storyboards. Since the learning result of a particular student is associated to a particular path through the storyboard, we should be able to identify successful storyboards in general, but also suc-

cessful paths within storyboards in particular. Through the use of Machine Learning methods, we finally might be able to find out what these successful storyboards respectively paths have in common and in which properties they differ from the less successful ones. Thus, we might be able to identify successful didactic patterns.

The latter is, in fact, the vision of knowledge discovery in didactics. By utilizing the didactic insights acquired by this approach for the upcoming storyboards, we intend to close the loop of the never ending storyboard development spiral.

Acknowledgements The authors are grateful for the generous support from the Laboratory of Interdisciplinary Information Science and Technology (I²-Lab) at the School of Electrical Engineering and Computer Science at the University of Central Florida. Thanks to this support we could build the exemplary storyboard, which is (1) an important source of experience for the authors, (2) a starting point for future research to refine the concept and, most importantly (3) a launch pad of a campaign that aims at the wide dissemination of this concept.

References

- [1] K.P. Jantke and R. Knauf. Didactic design through storyboarding: Standard concepts for standard tools. In *Proc. of the 4th International Symposium on Information and Communication Technologies, Workshop on Dissemination of e-Learning Technologies and Applications, Cape Town, South Africa, 2005*, pages 20–25, 2005.
- [2] R. Knauf, A.J. Gonzalez, and T. Abel. A framework for validation of rule-based systems. *IEEE Transactions of Systems, Man and Cybernetics - Part B: Cybernetics*, 32(3):281–295, June 2002.
- [3] M.H. Walker and N. Eaton. *icrosoft Office Visio 2003 Inside Out*. Redmond, Washington: Microsoft Press, 2004.

Data Mining on Traffic Accident Data

Jürgen Cleve[†], Christian Andersch^{*}, Stefan Wissuwa[†]

[†]University of Wismar, Dept. of Economics
Philipp-Müller-Str., D-23952 Wismar, Germany
Phone: +49 3841 753-527
E-mail: {j.cleve,s.wissuwa}@wi.hs-wismar.de
<http://www.wi.hs-wismar.de/kiwi>

^{*}Current affiliation: MIT – Management Intelligenter Technologien GmbH
Pascalstr. 69, D-52076 Aachen, Germany
E-mail: christian@andersch.net

Abstract

A large number of people is killed or injured in traffic accidents every year. We describe an approach to determine the seriousness of potential accidents. Therefore, we use data mining techniques to analyze data from traffic accidents and to build models for prediction. We use freely available data mining tools and self-developed software for data preprocessing and automation.

Keywords: data analysis, data mining, machine learning, traffic accident data

1 Traffic Accident Data

The number of people killed in traffic accidents per year is about 1.2 million worldwide. Even more get injured: 50 million people [6]. It is important to lower these numbers.

Analyzing the accidents may help to understand and isolate facts with significant influence. In this paper we describe the analysis of traffic accident data from the Rostock¹ area. The question is: If there is an accident, how serious will it be?²

¹City located in the northern part of Germany; important Baltic seaport

²See also: “Improving road safety with data mining”, <http://soleunet.ijs.si/website/html/cocasesolutions.html>

We have access to official statistics covering several years of the Rostock area. There are a total of 10,813 anonymized records available. Each is marked with a ‘score’ that indicates the seriousness of the accident [2]. The score is an integer number on a scale from 1 to 9 with 9 as maximum. It was originally calculated from the number of injured or killed people and the material damage as shown in Table 1. It is worth to notice that the rules for the score were created by physicians. Since the material damage is not included in the data provided, the mapping from material damage and personal injury to the score cannot be reversed.

Each record contains 52 attributes describing the circumstances of the accident. The attributes include weather

Table 1: Score definition

Material damage (€)	Personal injury	Score
[0; 250)	none	1
[250; 500)	none	2
[500; 2,000)	none	3
[2,000; 5,000)	barely injured	4
[5,000; 10,000)	≤ 2 badly injured	5
[10,000; 25,000)	> 2 badly injured	6
[25,000; 50,000)	1 dead	7
≥ 50,000	several dead and badly injured	8
massive accident with ≥ 10 vehicles and/or ≥ 5 badly injured		9

conditions like rain or darkness, violations of traffic regulations like speed limits or alcohol, character of the surrounding area like trees or sharp turns, type and number of vehicles and people involved, type and number of injured people, and the age and gender of the person who caused the accident.

From these attributes are 36 binary sparse data. Eight of them, the attributes that indicate violation of traffic regulations, are mutually exclusive, so they might be interpreted as one attribute.

The values for age class and gender are missing in 194 records (1.79%) because of hit-and-run drivers.

The data are ambiguous, which means that several records with identical attribute values can have different scores. This is a major problem for machine learning because contradictions cause diametrical learning effects. It also raises questions about data quality or missing attributes. On the other hand, this may be the result of highly chaotic behavior in complex dynamic systems as traffic accidents usually are.

We used a relational DBMS to hold the data and to perform early analysis, WEKA [9] for data mining because of its high number of included algorithms, and our self-developed software

environment ‘Eddie’, described in Section 3, for data preprocessing.

1.1 Preliminary considerations

The main goal was to predict the score of (un)known data. Therefore, to identify over-fitting, the comparison of algorithms was based on three different methods for training and test data.

- Test on training data: This method shows how good the training data are learned. It is typically not suited for prediction. If not stated otherwise, test on training data in this paper means training and testing on all 10,813 records.
- The $\frac{2}{3}$ -rule means a random split of the data into $\frac{2}{3}$ for training and $\frac{1}{3}$ for testing.
- Cross validation means multiple stratified splitting into test and training data, where the errors are averaged. We based our main criteria to judge the result’s quality on 10-times cross validation (CV10).

The algorithms used for classification include decision trees such as ID3 or Random Forest [3], rule based algorithms, and SMO [5] as representative

for Support Vector Machines (SVM). We used the standard parameters for all algorithms.³ As additional step, the boosting algorithms Vote, AdaBoost, and Bagging were applied for further optimizations.

Since ID3 ‘stores’ all distinguishable input data, it will give an upper limit on the training data and can therefore be used as reference for how good the training data can be learned. ZeroR always predicts the mean/mode value and therefore gives the lower limit on CV10, which should be outperformed by all other algorithms.

1.2 Methods and Procedures

Classification predicts a target attribute from a set of non-target attributes. This equals the function

$$f(x_1, \dots, x_{n-1}) = x_n$$

where x_n is the target attribute.

To evaluate the results of different experiments, a quality indicator is needed. The simplest indicator is the total recognition rate:

$$\bar{R} = \frac{\# \text{ correctly classified records}}{\text{total } \# \text{ records}}$$

The classification results for test data are often represented as square confusion matrix M that shows the number of records from each class (score) and how they are classified. This allows us to define recall, precision, and f-measure for each score.

Let i, j be the row and column index of the confusion matrix M . Recall R is the normalized amount of correctly classified records for each class:

³Except for SMO, where we used radial basis function (RBF) with $c = 1$, $\gamma = 0.3$.

$$R(x) = \frac{m_{x,x}}{\sum_{j=1}^n m_{x,j}}$$

Precision P defines the accuracy of the predicted class:

$$P(x) = \frac{m_{x,x}}{\sum_{i=1}^n m_{i,x}}$$

Both, precision and recall are equal to the proportion of the main diagonal element versus the column or row sum. F-measure F is the ‘average’ of R and P :

$$\frac{1}{F} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{R} \right) \rightarrow F = 2 \cdot \frac{P \cdot R}{P + R}$$

Many data mining algorithms try to minimize F . This is especially useful if no cost matrix C is used. In case a cost matrix is used, its elements $c_{i,j}$ are multiplied with the corresponding elements of the confusion matrix $m_{i,j}$. In our experiments we used the following cost matrix:

$$C_9 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 \\ 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 6 & 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 1 \\ 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

The main diagonal elements $c_{i,i}$ represent the correctly classified records. As costs for incorrectly classified records we chose the difference between real and predicted score. The cost function is the sum of all singular costs, which needs to be minimized:

$$\text{Costs} = \sum_{i=1}^n \sum_{j=1}^n m_{i,j} \cdot c_{i,j} \rightarrow \min$$

Table 2: Identical records

Score difference	Identical records	
	Absolute #	%
0	3,729	34.5
1	2,795	25.8
2	1,953	18.1
3	537	5.0
4	108	1.0
5	14	0.1
6-8	—	—

2 First Results

2.1 Classification using original data

The experiments were performed using the original data and score. For all values either a nominal or binary encoding was used. Missing values were set to 0. Further, we removed all attributes that used information ascertained after the accident already happened, such as number of injured people.

The results are shown in Table 3. ID3 and Random Forest have similar recognition rate and costs, but differ in cross validation. The best recognition of the training data with 77.5% by ID3 generally indicates that it may be hard to learn the data.

2.2 Classification using unique data

Because of the relatively bad results in our previous experiments we did further investigation of the data. We figured out that the data are partially ambiguous, which means that there are identical records but with different scores.

Ignoring the score attribute, there are a total of 5,993 (55.42%) different records. With score, there are 7,084 (65.51%) different records. Table 2

shows the distribution of score differences over the entire dataset. For example, there are 14 identical records with a score difference of 5.

Based on this we split the data into test and training data so that the training set contains only unique records. The remaining records were used as test data. Incomplete records were also removed from the training set, although this may be argued.

We used the same cost function as before. From the records with different scores we chose the one which was nearest to the average score. In case of doubt we chose the higher score because we considered higher scores as more important.

The training set contained 5,891 (54.48%) unique records and the testing set contained the remaining 4,992 (45.52%). All data were nominal encoded. Missing values for age and gender were set to NULL (non-existing). Violations of traffic rules were grouped into one attribute.

We got the best result using Random Forest. All results are shown in Table 4.

2.3 Classification using new scores

Previous results showed problems in predicting the score. In the last experiment, unique data was used for training, but not for testing. For the current experiment, we defined new scores to minimize ambiguousness. These new scores were directly calculated from the personal injury, ignoring the material damage, which was not provided in the original data. In several steps the personal injury was aggregated from ScoreB with 6 score classes (ScoreB = 1: only material damage; ScoreB = 2 to 6: personal injury according to the

Table 3: Original data—total recognition rate and costs per algorithm

Algorithm	Total recognition rate (%)			Costs		
	Training	CV10	$\frac{2}{3}$ -rule	Training	CV10	$\frac{2}{3}$ -rule
Random Forest	77.3	46.9	44.4	3,697	9,005	3,208
J48	60.7	44.5	44.2	6,507	9,303	3,100
REPTree	49.2	41.6	41.1	8,352	9,496	3,291
ZeroR	38.4	38.4	39.4	9,776	9,776	3,266
ID3	77.5	43.8	40.8	3,678	10,426	3,858

Table 4: Unique data—total recognition rate and costs per algorithm

Algorithm	Total recognition rate (%)				Costs			
	Train.	Test	Total	CV10*	Train.	Test	Total	CV10*
Rand. Forest	93.2	48.4	72.8	35.5	686	3,439	4,125	6,109
J48	60.1	43.8	52.7	40.7	3,715	3,804	7,519	5,405
REPTree	47.0	42.1	44.8	38.6	4,808	3,898	8,706	5,483
ZeroR	34.4	43.1	38.4	34.4	5,830	3,946	9,776	5,830
ID3	93.4	47.3	72.4	28.6	680	4,370	5,050	8,108

*CV10 on unique data, therefore reduced costs compared to Table 3, CV10

original score) to ScoreF (ScoreF = 1: only material damage; ScoreF = 2: also personal injury). Results for ScoreB as the most complex and ScoreF as the most simple differentiation are described here, for other scores see [1].

We used nominal encoding and set missing values to a special value. Attributes for violations of traffic rules were not grouped into one attribute as before. All attributes collected after the accident were removed. Altogether, we used 44 attributes including the score.

The results for total recognition shows Table 5. The total recognition rate is in inverse proportion to the number of classes used. ID3 on training data recognizes exactly the number of distinguishable records, as verified by analyzing the database. Using CV10 as criteria, SMO gives the best results.

Table 7 shows a detailed overview of single and total recognition as well as costs when using CV10.

Most algorithms did not predict the higher scores at all. Furthermore, the recognition rate of a score is nearly proportional to the number of appearances in the original data. This cannot be generalized for score 4 to 6 because of their very low appearance. Random Forest and SMO have the lowest costs and highest total recognition rate, but only Random Forest predicts higher scores.

Directly learning score variants with fewer classes shows slightly better results than learning ScoreB and aggregating to ScoreF afterwards [1]. Details for ScoreF and one algorithm shows Table 6.

ScoreF only distinguishes the binary decision problem material damage vs. personal injury. Therefore, the cost matrix leads to costs that equal the number of wrongly classified records:

$$C_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Table 5: Total recognition rate (%)

Alg.	Test	Score	ScoreB	ScoreF
ID3	Train.	77.5	97.0	97.4
	CV10	45.2	87.0	89.4
	$\frac{2}{3}$ -rule	42.7	85.7	88.8
J48	Train.	59.9	90.7	91.9
	CV10	44.4	88.0	90.8
	$\frac{2}{3}$ -rule	43.6	87.9	90.6
SMO	Train.	58.3	91.6	93.4
	CV10	46.5	88.7	91.1
	$\frac{2}{3}$ -rule	45.5	88.5	91.1

Table 6: Details of ScoreF for J48, CV10; R is recall, P is precision (%)

S	1	2	R
1	9,047	203	97.81
2	797	766	49.01
P	91.90	79.05	90.75

As new optimization criteria, f-measure for ScoreF = 2 (personal injury) was used, since it shows the quality and precision of the classification of serious accidents. The reduced complexity allowed us to further use several boosting algorithms. The best results are shown in Table 8 including recall and precision for ScoreF = 2, total recognition rate, and costs.

The best result with the highest f-measure of 0.667, almost 60% recognition and little more than 21% false-positive was achieved with voting using Random Forest and the rather simple algorithms Decision Stump and ID3. Replacing ID3 with PART resulted in slightly lower f-measure, but higher total recognition rate of 92.0% and lower costs. For J48, Table 8 compared to Table 5 also shows that boosting algorithms do not always lead to a better total recognition rate.

3 Data Mining Environment for Preprocessing and Automation

We are developing a data mining environment for data preprocessing, automation and integration of data mining software, called Eddie: Extensible Dynamic Data Interchange Environment. The system was used in part for data preprocessing and table operations. It is designed to support scientific experiments, which need a flexible and in-depth capability to adjust parameters, to handle very large amounts of data, and to perform multiple experiments automatically [7], [8], [4].

3.1 Motivation

When running data mining experiments, we often have to use several software applications. Reasons are that not every algorithm is supported by every software, non-transparent algorithms complicate the interpretation of results or models (especially true for Neural Networks), and visualization capabilities are not always appropriate. It is necessary to exchange data between different applications and to manually perform various formatting steps because of proprietary data formats. While for a single experiment this is not a limiting factor, it becomes very time consuming and error prone for larger series of experiments.

To solve these issues we decided to develop a set of data transformation tools that convert data from proprietary formats into a single exchange format and back, all using standard input and output streams so they can easily be appended. We chose XML as standard exchange format since it is easy to parse, to edit, and human

Table 7: New score—recognition rate (%) and costs for ScoreB, CV10

ScoreB	%	ID3	Dec.Table	PART	J48	NBTree	SMO	C.Rule	Rand.For.
1	85.5	94.8	98.2	96.4	98.1	97.3	98.6	98.1	97.3
2	9.5	44.4	39.3	36.9	33.1	41.9	40.1	42.3	43.5
3	4.2	38.9	13.1	26.0	21.4	8.1	12.0	—	38.4
4	0.3	18.2	9.1	—	—	—	—	—	18.2
5	0.3	36.1	—	—	—	—	—	—	39.5
6	0.1	50.0	8.3	—	—	5.6	—	—	50.0
uncl.	—	0.2	—	—	—	—	—	—	—
Total:	100.0	87.0	88.3	87.0	88.0	87.5	88.7	87.9	89.2
Costs:		1,878	1,720	1,890	1,772	1,857	1,671	1,764	1,545

Table 8: Best boosted ScoreF, CV10; based on f-measure for ScoreF = 2 (F-M2)

Algorithm	F-M2	R2 (%)	P2 (%)	Total (%)	Costs
Vote: Rand.For., Dec.Stump, ID3	0.677	59.4	78.6	91.8	887
Vote: Rand.For., Dec.Stump, PART	0.663	54.4	84.6	92.0	867
Random Forest: Bagging	0.646	56.0	76.1	91.1	962
PART: AdaBoost	0.645	58.3	72.0	90.7	1,005
J48: AdaBoost	0.640	57.6	72.0	90.6	1,013
...					
SMO*	0.621	50.6	80.3	91.1	966
...					

*not boosted because of much higher CPU consumption

readable. Because this set of programs was already very flexible, we found that using the underlying architecture could be used to describe and automate general data flows and preprocessing steps.

3.2 Architecture

Its main concept is to split functionality into several stand-alone programs for well-defined tasks, which communicate with each other using a flexible XML-based protocol, which we call SXML. These modules can be connected via input and output streams to build a workflow or to exchange data with other applications, as shown in Fig. 1. This allows us to build complex data flows, to store intermediate results at any time for further investigation, and to integrate the functionality of existing applications, such as WEKA for data mining and SNNS [10]

as specialized software for Neural Networks.

Each module reads and writes SXML data, which contains the data to be processed as well as configuration information. Therefore, a workflow can be described as configuration for a special module, which then reads its configuration, executes programs, and manages the data flow between them. Because the workflow module itself reads and writes SXML code, a workflow can easily be integrated within another workflow. Data import and export as well as integration of existing applications are handled this way using special modules.

4 Conclusion

A large number of correctly predicted, but less serious accidents resulted in a relatively high total recognition rate.

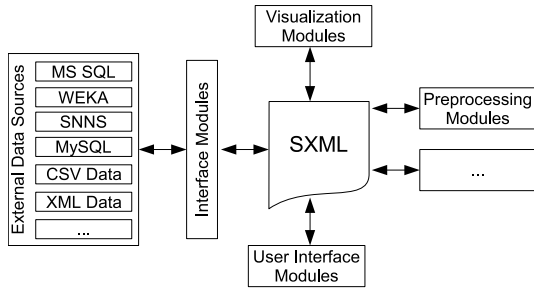


Fig. 1: System architecture of Eddie

The best result for the separation of high and low score achieved a voting algorithm with an f-measure of 0.677 for the high score. Such higher scores are more important as they indicate situations with matter of life and dead.

Better results may be possible by decreasing the ambiguity of the data combined with a more specific subset of training data. Especially the score definition is not mathematically derived and very subjective. Further, an asymmetric cost matrix with higher penalty for too low classified records should produce better models.

Our experiments show that many difficulties in building data mining models are caused by inadequate data, which had been raised and partially pre-processed without taking information-theoretical aspects into account.

References

- [1] C. Andersch and J. Cleve. *Data Mining auf Unfalldaten*. University of Wismar, January 2006. Wismar Discussion Paper 01/2006, <http://www.wi.hs-wismar.de/fbw/aktuelles/wdp/>.
- [2] D. Bastian and R. Stoll. *Risikopotenziale und Risikomanagement im Straßenverkehr*. Technical report, Forschungsinstitut für Verkehrssicherheit GmbH Schwerin, University of Rostock, July 2004.
- [3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [4] J. Cleve, U. Lämmel, and S. Wisuwa. Data Mining on Transaction Data. In Nejdet Delener and Chiang-Nan Chiao, editors, *Global Markets in Dynamic Environments*, Lisboa, 2005. GBATA.
- [5] J. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998. Microsoft Research Technical Report MSR-TR-98-14.
- [6] WHO. World Health Day: Road safety is no accident!, 2004. Press release for the World Health Day 2004, <http://www.who.int/mediacentre/news/releases/2004/pr24/en/>.
- [7] S. Wisuwa. *Data Mining und XML – Modularisierung und Automatisierung von Verarbeitungsschritten*. University of Wismar, 2003. Wismar Discussion Paper 12/2003, <http://www.wi.hs-wismar.de/fbw/aktuelles/wdp/>.
- [8] S. Wisuwa, U. Lämmel, and J. Cleve. XML-basierte Beschreibung von Data-Mining-Prozessen. In J. Biethahn, A. Lackner, and V. Nissen, editors, *Information-Mining und Wissensmanagement in Wissenschaft und Wirtschaft, Proc. AFN-Jahrestagung*, pages 37–48, Göttingen, June 2004. University of Göttingen.
- [9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.
- [10] Andreas Zell. *Simulation Neuronaler Netze*. Oldenbourg, München, 2003.

The Digital Mechanism and Gear Library – a Modern Knowledge Space

Torsten Brix, Ulf Döring, Sabine Trott, Rike Brecht, Hendrik Thomas
Technical University of Ilmenau, PO Box 10 05 65
98684 Ilmenau, Germany
Torsten.Brix@TU-Ilmenau.de

Abstract

Mechanisms and gears are an essential part of technical products in industry. However, the worldwide existing knowledge about mechanisms in theory and practice is mostly scattered and only fragmentarily accessible for users like students, engineers and scientist. It does not comply with today's requirements concerning a rapid information retrieval. This paper presents the "Digital Mechanism and Gear Library" (DMG-Lib). In this interdisciplinary project of the Technical Universities of Ilmenau, Dresden and the RWTH Aachen a new digital, internet-based library (www.dmg-lib.de) is built to collect, preserve and present the knowledge of mechanism and gear science on a new level of quality. The DMG-Lib contains a wide range of digitalized information resources in very heterogeneous media types. The resources are enriched with various additional information like animations and simulations. Combined with innovative multimedia applications and a semantic information retrieval environment, the DMG-Lib provides an efficient access to this knowledge space of mechanism and gear science.

1. Introduction

In the middle of the 19th century in Germany the systematic research on mechanisms and gears started as a result of the fast growing engine building industry in this time [5]. Especially the theoretical reflections and practical works of the German engineer F. Reuleaux [15] became important. Mechanism and gear technology is today still essential for industry and it will become even more important due to the introduction of new technologies like nanotechnology and corresponding new fields of application.

The existing knowledge about mechanisms in theory and practice is worldwide scattered in hand- and textbooks, photographs, solid functional models, engineering drawings, etc. It is only limited and very fragmentarily accessible and does not comply with today's requirements concerning a rapid information retrieval [5]. However, industrial companies and research institutes demand an efficient access to the whole mechanism and gear theory [7]. Existing activities to pro-

vide such access are promising (e. g. [3]) but by far insufficiently. Today in Germany only 12 university institutes with focus on mechanism and gear science are left. More and more didactical experiences and valuable training material are lost because experts on this field of application retire or through economy measures specialized institutes are closed. Also old and unique literature with only a few numbers left are quite difficult to access like the publications of Reuleaux. They have to be digitized and online presented so that this still important knowledge becomes accessible for the public again.

A solution of these problems is the collection and presentation of all relevant information resources for mechanism and gear science in a centralized worldwide accessible platform [5, 8]. The research and education in various ingenious disciplines would certainly benefit from such a comprehensive library of knowledge.

In 2004 the development of the worldwide accessible "Digital Mechanism and Gear Li-

brary” (DMG-Lib) was started to prevent this sneaking lose of knowledge. The DMG-Lib is an interdisciplinary project of different departments of the Technical Universities of Ilmenau, Dresden and the RWTH Aachen. It is financed by the “German Research Foundation” in the program “Scientific Library Services and Information Systems” (project number: LIS 4-554975).

The aim of this project is the collection, integration, preservation, systematization and adequate presentation of the worldwide knowledge about mechanisms and gears. The gained results and experiences of this project will hopefully help in future other digital libraries in different application domains as well.

The digital library is designed to satisfy the requirements of different user groups like engineers, scientists, teachers, students, librarians, historians and others. To offer users a wide variety of opportunities for retrieval and utilization the digitized resources are extensively post-processed and enriched with various information like animations, meta-data, references and constraint based models. The focus is not only on textual documents, images and animations. Also functional models are digitalized, which exists in thousands of unique models with no or only very limited access for the public.

This huge amount of available heterogeneous information resources in the DMG-Lib implies a key challenge of this project: the implementation of an efficient, uniform and user-satisfying information retrieval [8, 14].

In the following section the concept of the DMG-Lib project is introduced. Afterwards the implementation of the DMG-Lib is presented. Thereby the digitalization and enrichment of the information resources and the DMG-Lib online portal are discussed. Also developed multimedia applications and a semantic information retrieval environment for innovative ways of presentation and retrieval in the knowledge space are described. Finally, the paper concludes with a summary and an evaluation of the project.

2. Concept of the DMG-Lib

The DMG-Lib contains a vast amount of very heterogeneous information resources (see Fig. 1) like books, publications, functional models, gear catalogues, videos, images, technical reports, etc. The original sources are procured, digitized and converted to suitable data formats.

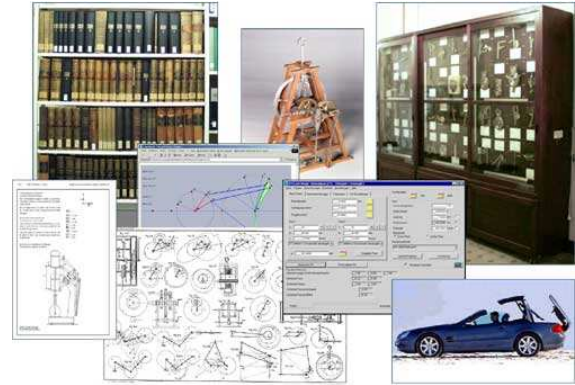


Fig. 1: Examples of information resources in the DMG-Lib

The information resources can be accessed worldwide on the DMG-Lib internet portal. This simplifies the access and distribution of these information resources, but does not directly enhance a goal-oriented usage and retrieval for solutions of technical tasks in research and industry. Furthermore the common storage method for knowledge, mainly in static texts and images, does not comply with requirements concerning an efficient and fast information retrieval. The advantages of functional models for a better understanding of complex construction and function principles are well known. Today the necessary techniques are available to provide an easy access to such helpful demonstration models for a broad public. Computer based methods enable the generation of multimedia documents which describe the function and other relevant attributes of mechanisms and gears. These can easily be distributed and enriched with extensive additional information [7].

Therefore in contrast to other digital libraries projects, which often provide only access to the digital raw data [4], in the DMG-Lib project the digitized resources

are extensively post-processed and enriched with various information like animations, constraint-based models or various verbal descriptions. Also further simulations and analyses are possible, because constraint-based models can be used in external analysis, synthesis and optimization systems. Such approaches are necessary to move from a static to a dynamic problem oriented supply of knowledge for a wide range of application domains and user requirements.

An overview of the complex production workflow for the identification, digitalization, enrichment, storage and presentation of information resources in the DMG-Lib is displayed in the following figure (see Fig. 2).

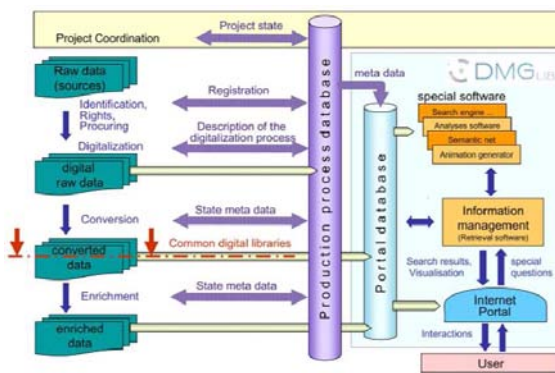


Fig. 2: Production workflow in the DMG-Lib

Based on the vast amount of available heterogeneous information resources in the library and the extensive enrichment, the DMG-Lib is able to provide an efficient retrieval as well as various utilization options for users. Following these considerations several additional aims of the DMG-Lib project can be derived:

- Constraint based modeling of mechanisms and gears as base for generation of further description forms [7]
- Supply of descriptions of mechanism and gear knowledge in various forms to ensure a flexible, adaptive and long term usability (verbal, images, constraint-based descriptions, 2D and 3D animations)

- Cross-platform presentation in the internet for different user-groups and different use-cases like research, product development or self-study
- Development of information retrieval systems, which allow a structural selection and type syntheses of mechanisms and gears
- Support of automated access options for the library content using various applied descriptors or meta-data (e. g. OAI-PMH service)
- Support for researchers and developers during the development of solutions for special synthesis or optimizing problems

3. Implementation of the DMG-Lib

For the implementation of this ambitious concept a consequent cooperation of information, computer and usability scientists as well as engineers, librarians and experts of mechanism and gear science is necessary. This is the only way to collect, enrich and present the complex domain specific heterogeneous information resources according to user requirements.

3.1. Enrichment of the information resources

The following information sources are digitized and integrated in the digital library:

- Literature relevant for mechanism and gear technology (monographs, journal articles, etc.) from different libraries and private collections
- Solid mechanism and gear models of the TU-Ilmenau, the TU-Dresden and the RWTH Aachen
- Images and slides of gears available in the project partners archives
- Technical drawings (outlines, technical blueprints, technical principles and calculation instructions)

- Training materials of the departments involved in the DMG-Lib project

The literature sources are usually scanned with 300 dpi resolution and 256 greyscales and are saved as TIFF files. For the scanned resources meta-data according to the Dublin Core standard are stored in the production database [1]. In addition the documents are classified according to technical aspects.

For further processing of the digital raw data a layout and text analysis is necessary. For the identification of the physical structure (text blocks, images etc.) as well as the individual characters in the documents the commercial software ABBYY-Finereader is used. The software is embedded in a self developed application framework called AnAnAS (**Analyse-Anreicherungs-Aufbereitungs-Software**).

Other applications developed in the DMG-Lib project identify the logical structure (headlines, labels of figures etc.) more and more automatically. The storage of the meta-data in AnAnAS is based on the METS-Standard [2]. Different meta-data are added to the documents like administrative (who scanned the document, document source), descriptive (e. g. Dublin Core) and structural (connection between the content and other meta-data like figure references). The result of the structural and layout analysis is the identified logical structure of the document. This information can be used in further processing steps like the automated generation of links and tables of contents as well as in the ranking of full text search results.

For the enrichment of the scanned documents an animation generator was developed which allows the simulation and the variation of drawings, images and models in an easy and fast way (see Fig. 3).

An export to CAD and special analysis software systems will be available as well. Base for the export and the animation generation is a special XML based file format in which the description of the displayed gear is stored [7]. These abstract model descrip-

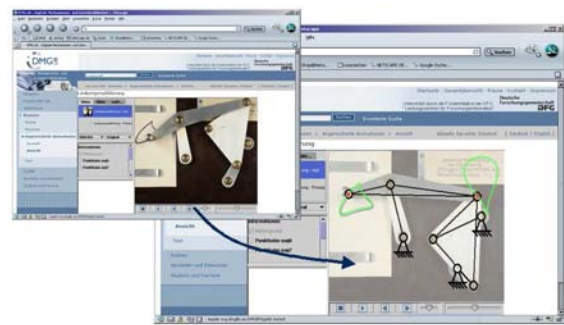


Fig. 3: Enhancement of videos by an overlaid simulation-based animation

tions provide rich information for various search criteria for example the number of elements of the gear. The analysis of the simulation results provides further information describing the function of the gear like the transmission behavior. This functional information is important for the implementation of a problem oriented information retrieval.

To the individual models, animations, images and literature resources experts can attach further meta-data like detailed descriptions, cross-links and other annotations. This information will be edited either in the AnAnAS system during the processing of the digital raw data or in special designed interfaces directly in the production database.

A first version of the production database was developed using MySQL and content is now continually added. In June 2006 the DMG-Lib portal included about 30 books, 700 demonstration models, 45 bibliographic entries and more than 40 enhanced images and videos. However in the production database over 900 documents and 400 persons relevant to the DMG-Lib are listed. In the next years thousands of resources will be provided in the portal.

3.2. DMG-Lib Online Portal

The portal is the internet based communication and presentation interface between the user and the DMG-Lib (see Fig. 4). For an user adequate design and implementation an evaluation of the usability was performed which is oriented on the Usability Engineering Lifecycle developed by Deborah J.

Mayhew [11]. According to this method a requirement analysis and expert interviews have been carried out to develop a conceptual model of the DMG-Lib portal.

In March 2006 the prototypic online portal on www.dmg-lib.org was activated. It currently serves as a platform for usability tests. Beside the interactive search option in the web portal the content of the DMG-Lib can be accessed with an OAI-PMH web service as well.

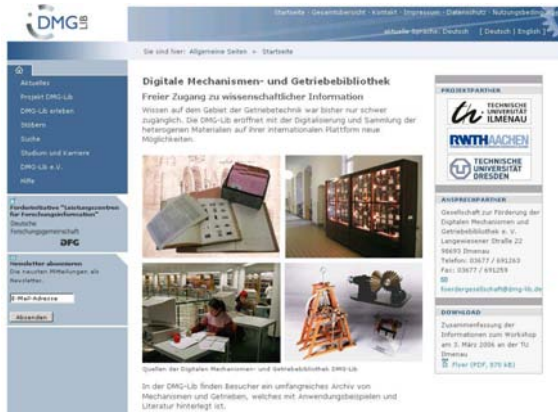


Fig. 4: DMG-Lib portal

3.3. Multimedia Applications of the DMG-Lib

Parallel to the internet portal interface other interactive multimedia applications are developed like the multimedia timeline (see Fig. 5) and the virtual mechanism and gear museum.

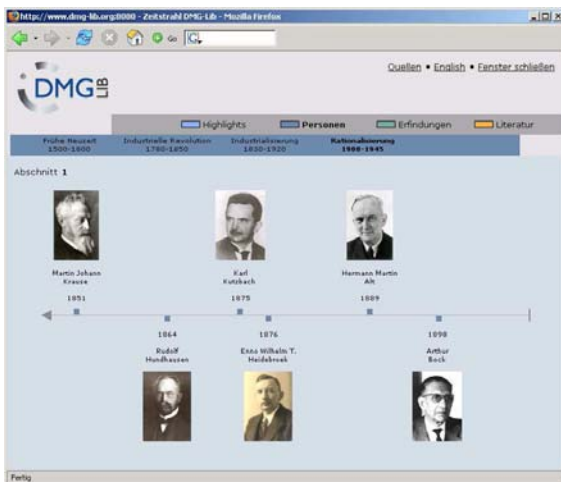


Fig. 5: Timeline of mechanism and gear development

The timeline application gives users a multimedia overview of important persons, inventions and publications in the historical development of mechanism and gear science. Users will be able to directly access corresponding information resources, for example available books of selected persons in the library.

Beside traditional browsing and retrieval methods these applications provide alternative ways to access the knowledge stored in the library. Prototypes of these applications are integrated in the DMG-Lib portal and are currently tested by the user community.

3.4. Semantic Information Retrieval

A further field of research is the retrieval in heterogeneous information resources using different mechanism and gear hierarchies like the structural system of Reuleaux [15] or other classification systems of well-known publications (e. g. [6]).

A visualization and an efficient navigation over these different categories of gears could help users to get a systematic overview over the huge amount of existing mechanism and gear constructions. However, the identification and modeling of these classifications and relations between the different technical terms are quite complicated, because different opinions of experts and authors have to be considered.

To solve this problem semantic web technologies can be used. With the help of Topic Maps, as a special kind of semantic networks, the knowledge of mechanism and gear science can be generalized and explicit modeled in a semantic meta-layer [12, 13]. With the extensive descriptive power of Topic Maps, all relevant concepts and relations between the concepts of this application domain can be modeled. Additionally, valid contexts, alternative names and other relevant semantic information can be included. Furthermore each concept in the semantic meta-layer is linked to all relevant information resources available in the library (see Fig. 6).

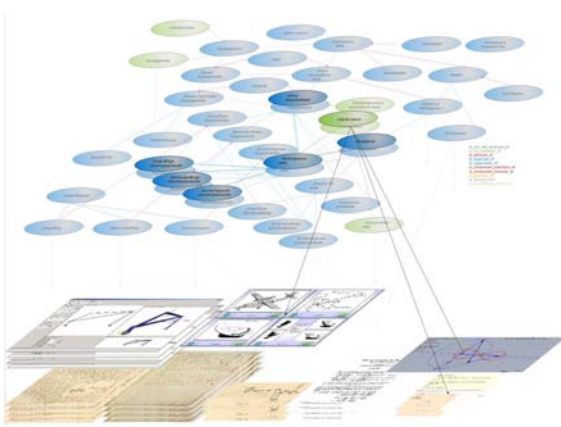


Fig. 6: Semantic meta-layer

With the help of this semantic meta-layer the different hierarchies can be modeled and visualized. This enables a user to decide which structuring system he wants to use for navigation.

Currently a Topic Map based “Semantic Information Retrieval Environment for digital libraries” (SIREN) is developed to support the complex development process of the semantic meta-layer and the information retrieval process. SIREN consists of three prototypical systems, which are developed as part of the DMG-Lib:

- **TMwiki** (Topic Map Wiki) [10, 16] enables a collaborative development of semantic meta-layers in a wiki environment.
- **TMV** (Topic Map Visualizer) [10] provides a user-friendly interface for visualization, presentation and navigation in the semantic meta-layer, a graphical topic-based definition of information needs and the presentation of the search results in the semantic context.
- **MERLINO** (Method for extraction and retrieval of links for occurrences) [9, 16] is able to identify relevant information resources for a defined information need automatically. The prototype identifies relevant information resources by querying the database of the digital library based on the knowledge stored in the semantic meta-layer.

Based on the semantic information and with the help of SIREN the structuring and the retrieval in the available heterogeneous information resources of the DMG-Lib can be enhanced.

4. Conclusion

In this paper the DMG-Lib project is presented, a digital and interactive library for mechanism and gear science. Aim of this project is the collection, preservation and suitable presentation of the worldwide existing knowledge about mechanisms and gears. Outstanding features of the digital library are the powerful and user-oriented internet portal and the integration of a high amount of very heterogeneous information resources relevant for this field of application. The extensively post-processing and enrichment of the digital data with various additional information like animations or constraint-based models is also important. Combined with the development of new interactive multimedia applications and a semantic information retrieval environment, the DMG-Lib provides users with an innovative access to the stored knowledge in the library.

The DMG-Lib project is an example for a modern knowledge space, which tries to satisfy the users’ needs for an efficient access to required information as one of the key tasks in our today’s information society.

References

- [1] *The Dublin Core Metadata Initiative*. <http://dublincore.org> (2006-20-06), 2006.
- [2] *Metadata Encoding and Transmission Standard*. Library of Congress. <http://www.loc.gov/standards/mets> (2006-06-20), 2006.
- [3] *Web Resource of the Kinematic Models for Design Digital Library*. <http://kmoddl.library.cornell.edu> (2006-06-20), 2006.
- [4] Christine Borgman and Lászlóc Kovcs Ingeborg Sølvsberg and editors.

- Proceedings of the Fourth DELOS Workshop Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics. Budapest, Hungary, June 6-7, 2002.* <http://www.sztaki.hu/conferences/deval/presentations.html> (2006-20-06), 2002.
- [5] Torsten Brix, Ulf Döring, and Sabine Trott. DMG-Lib – ein moderner Wissensraum für die Getriebetechnik. In *Knowledge extended: die Kooperation von Wissenschaftlern und Bibliothekaren und IT-Spezialisten, 3. Konferenz der Zentralbibliothek (KX'05), November 2-4, 2005, Jülich, Germany*, pages 251–262. Jülich: Schriften des Forschungszentrums Jülich, 2005.
- [6] Kammer der Technik. *Begriffe und Darstellungsmittel der Mechanismentechnik*. Suhl: Kammer der Technik, 1978.
- [7] Ulf Döring, Torsten Brix, and Michael Reeßing. Application of Computational Kinematics in the Digital Mechanism and Gear Library DMG-Lib. Special issue on CK2005, International Workshop on Computational Kinematics. *Mechanism and Machine Theory*, 41(8):1003–1015, August 2006.
- [8] George A. Goodall. *A Time for Digital Libraries*. <http://www.dereguilo.com/facertationpdfs/aTimeForDigitalLibraries.pdf> (2006-20-06), 2003.
- [9] Bernd Markscheffel, Hendrik Thomas, and Dirk Stelzer. Merlino – a Prototype for semi automated Generation of Occurrences in Topic Maps using Internet Search Engines. In *Poster and Demos of the 3rd European Semantic Web Conference (ESWC2005). Heraklion, Greece, 29. Mai - 01. June, 2005.* http://topic-maps.org/lib/exe/fetch.php?cache=cache&media=member%3Aht%3Amerlinoabstract_eswc2005_greece.pdf (2006-20-06), 2005.
- [10] Bernd Markscheffel, Hendrik Thomas, and Dirk Stelzer. TMwiki – a Collaborative Environment for Topic Map Development. In *Poster and Demos of the 3rd European Semantic Web Conference (ESWC 2006). Budv, Montenegro, June 10-14, 2006.* http://topic-maps.org/lib/exe/fetch.php?cache=cache&media=member%3Aht%3Atmwiki_abstract_eswc2006_budva.pdf (2006-20-06), 2006.
- [11] Deborah J. Mayhew. *The Usability Engineering Lifecycle – A Practitioner's Handbook for User Interface Design*. San Francisco: Morgan Kaufmann Publishers Inc., 1999.
- [12] Jack Park and Sam Hunting. *XML Topic Maps: Creating and using topic maps for the web*. New Jersey: Pearson Education Inc., 2003.
- [13] Steve Pepper. *The TAO of Topic Maps*. <http://www.ontopia.net/topicmaps/materials/tao.html> (2006-20-06), 2000.
- [14] Edie Rasmussen. Information Retrieval Challenges for Digital Libraries. In *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL'04), December 13–17, 2004, Shanghai, China*, pages 93–103. New York: Springer, 2005.
- [15] Franz Reuleaux. *Lehrbuch der Kinetik*. Braunschweig: Vieweg, 1875.
- [16] Alexander Sigel. Report on the Open Space Session. In Lutz Maicher and Jack Park, editors, *Charting the Topic Maps Research and Application Landscape*, pages 270–280. Berlin: Springer, 2005.

The Use of GIS in Avalanche Modeling

Donna M. Delparte, PhD Candidate
University of Calgary, 2500 University Drive NW,
Calgary, Alberta, T2N 1N4, Canada,
dmdelparte@gmail.com

Abstract

Geographic Information Systems have the potential to aid in modeling snow avalanche terrain in order to identify areas of varying hazard. At the heart of terrain modeling is the digital elevation model (DEM). For this research a higher resolution DEM has been generated for selected areas within Glacier National Park in the Rogers Pass area of British Columbia. An avalanche database of the Rogers Pass highway corridor has been input into GIS using 3D mapping techniques and contains detailed information based upon expert knowledge. Terrain parameters have been extracted from the known avalanche paths to identify similar terrain in lesser known areas in the backcountry. Start zone and runout zone models aid in the process of identifying avalanche terrain. Visualization of the initial results has been integrated into Google Earth with the goal to allow potential backcountry users to recognize the snow avalanche hazard and reduce their level of risk.

1. Introduction

Snow avalanches are a significant natural hazard that impact roads, structures and threaten human lives in mountainous terrain. Modeling of terrain in a GIS is typically done by utilizing a digital elevation model (DEM). To evaluate what terrain parameters are most likely to contribute to high frequency, known avalanche paths are typically documented and evaluated for these key factors [2-5].

An avalanche path describes terrain boundaries of known or potential avalanches [1]. An avalanche path is characterized by: a start zone, track and runout zone (Figure 1). The starting zone of an avalanche path is where avalanches begin with a slope ranging from 30° - 50° , the track is where avalanches achieve maximum velocity and mass with slopes between 15° - 30° and the runout zone is where avalanches begin to decelerate and the deposition occurs.

This paper highlights the process of building a DEM and the methods for avalanche terrain modeling using a GIS. Initial results are featured along with a discussion of visualization for the general backcountry recreationist.

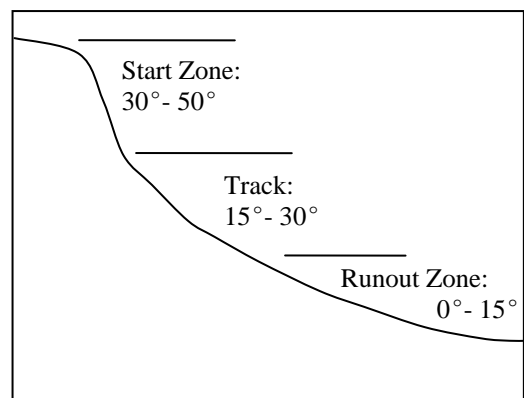


Figure 1. Anatomy of an Avalanche Path [1]

2. Methodology

A high resolution DEM for the study area was created using a procedure of digital stereo photogrammetry. This technology allows a GIS operator wearing stereo goggles to resample surface heights using stereo-airphoto pairs. The horizontal resolution of the DEM obtained from the 1:30,000 stereo-airphotos was 9 m. Vertical accuracy is within 1 m. Topographic parameters such as slope, surface area, aspect, runout length, and profile shape were derived from the DEM to evaluate start zone, track and runout characteristics that are most likely to contribute to avalanche frequency as determined by

known records from over a 130 avalanche paths along the Rogers Pass highway. The results are useful in conducting and improving avalanche hazard mapping as well as a tool for risk assessment. Terrain parameters from avalanche paths in the database along the highway corridor can be used to identify comparable areas in the backcountry.

The DEM facilitated analysis of start zone terrain and ground characteristics as well as runout parameters based on the alpha-beta statistical approach (Figure 2) first used by the Norwegian Geotechnical Institute [6].

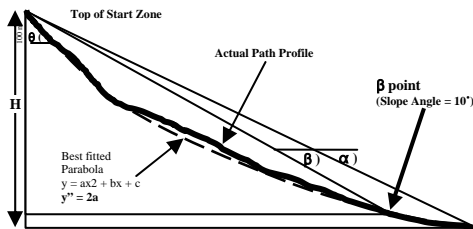


Figure 2. Alpha-Beta Statistical Model for Avalanche Runout

The Norwegian approach involves four specific terrain parameters that are represented in the regression equation below (1).

$$\alpha = 0.92\beta - 7.9 \times 10^{-4} H + 2.4 \times 10^{-2} Hy''\theta + 0.04 \quad (1)$$

$$\alpha = 0.92\beta - 1.4^\circ \quad (2)$$

Equation 1 represents a four parameter model:

α – represents the angle between the maximum runout and the top of the slide

H – is the vertical distance from the top of slide to the base as estimated by the best fitted parabola

β – represents the angle from a line of sight where the slope is 10° to the top of the slide

y'' – is the curvature of the slope based on a second derivative of a second degree polynomial

θ – is the average inclination of the starting zone as measured within the first vertical 100 m of the path

Equation 2 represents a simplified two parameter (β) model.

To apply terrain parameters for risk assessment, the information derived from the GIS analysis has been mapped according to the Parks Canada, Avalanche Terrain Exposure Scale (ATES) introduced in 2004 to reduce risk of backcountry users in National Parks and act as guidelines for backcountry use by custodial groups. The terrain based guidelines from ATES were used in the GIS to develop maps displaying backcountry areas based upon simple, challenging and complex terrain.

3. Initial Results

Figure 3 highlights the information digitized in stereo based upon avalanche expert confirmation. The thick bounding line indicates the boundaries of the snow avalanche revised to accurately represent the path, the dashed line represents the typical centerline of avalanche travel down the path, the shaded grey area indicates the start zone for the path and the dotted line indicates the typical start of avalanche runout or area where the avalanche begins to slow and deposit its snow load.



Figure 3. Avalanche Path 7

Geographic Information Systems allow for the extraction of terrain parameters from the DEM. Table 1 shows the extraction of terrain parameters from the start zone of

avalanche path 7. For each avalanche path along the highway corridor, parameters are extracted for the entire path, the start zone, the track and the runout. In the profile graph (Figure 4) the line segment to the triangle highlights the start zone portion of the entire path. The table reveals the start zone terrain parameters with measurements in degrees and meters. It is interesting to note that the mean slope for the start zone falls close to the commonly recognized 38° as a slope that is highly typical for avalanche activity.

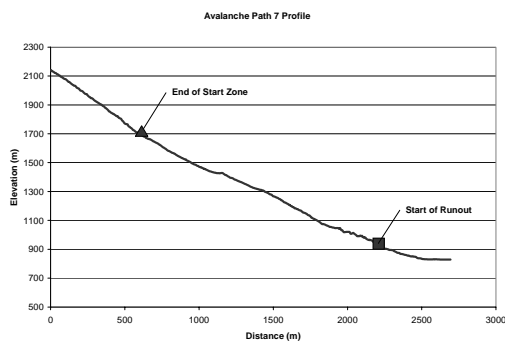


Figure 4. Profile of Avalanche Path 7

Table 1. Terrain Parameters from Avalanche Path 7

Terrain Parameter	Measure
Surface Length	756.8 m
Flat Length	572.8 m
Length Ratio	1.3 m
Mean Slope	37.6 °
Min Slope	0.2 °
Max Slope	79.9 °
Low Elevation	1730.2 m
High Elevation	2151.9 m
Average Elevation	1947.7 m
Range	421.7 m
Cum Z	442.4 m
Surface Area	230024.5 m ²
Flat Area	149535.2 m ²
Surf ratio	1.5
Aspect	NE

In the winter of 2005-2006, Avalanche Path 7 experienced a significant avalanche event. A photo was taken subsequent to the event (Figure 5). The photo reveals the crown

fracture line near the top of the slope. The thick bounding line indicates the path perimeter and the dashed line represents the main gully through which the avalanche path traveled.



Figure 5. Start Zone of Avalanche Path 7

4. Discussion-GIS Visualization

The capacity for Geographic Information Systems (GIS) to produce maps visualizing terrain features and improvements in web-based mapping software has led to an interest in providing avalanche risk based maps for the public, particularly in the form of on-line solutions. Google Earth is becoming a ubiquitous product that is expanding its reach to the general Internet user. Google Earth has the capacity to allow GIS overlays. It is online solutions such as Google Earth (Figure 6) that have the capacity to allow backcountry enthusiasts to view avalanche terrain and aid in their decision making.

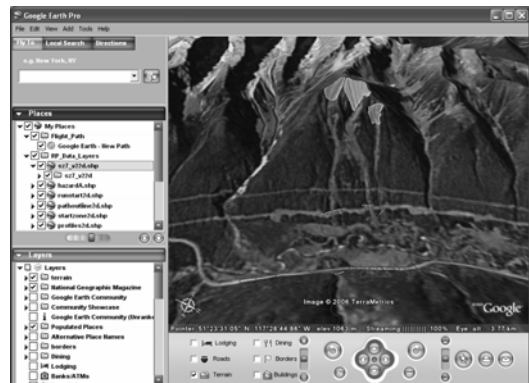


Figure 6. Google Earth

5. References

1. Mears, A.I., *Snow Avalanche Hazard Analysis for Land-Use Planning and Engineering*. Vol. Bulletin 49. 1992, Denver, Colorado: Colorado Geological Survey, Geological Survey, Dept. of Natural Resources. 54.
2. Maggioni, M., U. Gruber, and A. Stoffel. *Definition and characterisation of potential avalanche release areas*. in *Proceedings of the 2001 ESRI International User Conference*. 2001. San Diego.
3. Maggioni, M. and U. Gruber, *The influence of topographic parameters on avalanche release dimension and frequency*. *Cold Regions Science and Technology*, 2003. **37**: p. 407-419.
4. Gruber, U. *Using GIS for avalanche hazard mapping in Switzerland*. in *Proceedings of the 2001 ESRI International User Conference*. 2001. San Diego.
5. Gruber, U. and S. Sardemann. *High frequency avalanches: Release area characteristics and runout distances*. in *International Snow Science Workshop (ISSW)*. 2002. Penticton, BC, Canada.
6. Lied, K. and S. Bakkehoi, *Empirical calculation of snow-avalanche run-out distance based on topographic parameters*. *Journal of Glaciology*, 1980. **26**(94): p. 165-177.

A Case Study for Knowledge Externalization using an Interactive Video Environment on E-learning

Jochen Felix Boehm
Saarland University
Information sciences
Saarbruecken, Germany
jo.boehm@mx.uni-saarland.de

Jun Fujima
Hokkaido University
Meme Media Laboratory
Sapporo, Japan
fujima@meme.hokudai.ac.jp

Abstract

E-learning technology today should provide different means for the user to interact with the presented knowledge. In this case study an electronic instruction manual based on an instructional movie is presented and enhanced with interactive components. The integration of novel technologies – IntelligentPad and DVDconnector – is presented in this paper and the new possibilities for E-learning by these technologies is shown. The instruction manual application takes advantage of dynamic high quality videos and the possibility to interact with the learning environment. Also a new dimension of interaction will be shown. In standard E-learning scenarios the user normally acts with the information system and the environment separately. Using a special kind of pad, the system is now able to receive information via a sensor from the environment and can store this information as knowledge based items in form of IntelligentPads within the E-learning application. This paper shows a case study for externalization of knowledge in an interactive video environment based on an E-learning application designed as an electronic instruction manual for putting strings on a violin and tuning-up the violin. The user obtains all necessary information to correctly perform each step and can verify the correctness via a sensor interface showing the frequency of the violin.

1. Introduction

Nowadays E-learning is still used in ways of mediating knowledge from the teacher to the student, often only in textual form and sometimes supported with audio-visual media. The main forms of interaction given to the user are means of testing sequences to verify if the content has been successfully mediated and means of choosing his own learning path. However, in standard E-learning scenarios the user has to carry out these tasks separately, since the information system and the external environment are independent.

In order to provide a flexible and interactive E-Learning environment, various kinds of knowledge should be externalized as knowledge objects which can be accessed or ma-

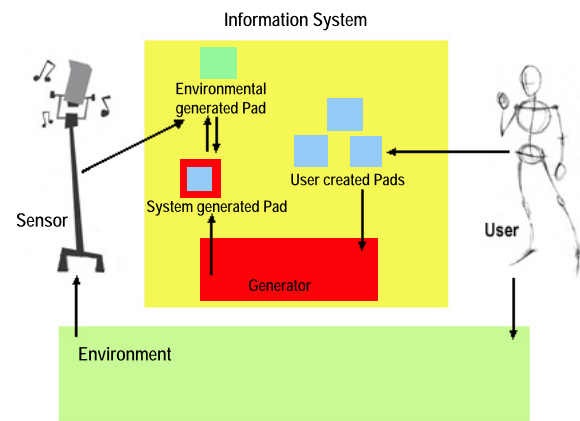


Figure 1: Model for representing knowledge objects in information systems

nipulated by the user. If the learning system is more advanced, new knowledge objects can also be created dynamically through some types of processes. Figure 1 shows the model for representing knowledge as knowledge objects in information systems. Each knowledge object can be created through three different processes. One can be generated automatically by the system. Another can be created by the user himself through operations with the system. The third possibility is the generation of knowledge objects by processing input data from the external environment through some kind of sensor devices.

This paper shows a case study for representation of knowledge in an E-learning application. The user can obtain necessary information through knowledge objects in different processes, such as interactive video, online/offline contents described by HTML or interactive object connected to sensor devices.

In this project, we use DVDconnector technology for providing an interactive video environment with dynamical contents and Meme Media [6] technology for representing interactive knowledge objects in information systems. DVDconnector provides DVD-based high quality videos which have the ability to access online/offline contents. On the other hand, Meme Media is a media technology for externalizing various kinds of intellectual information resources and distributing them among computer users. The integration of DVDconnector and Meme Media technology provides flexible and interactive environments for E-learning.

Following the basic approach of learning by imitation [2], the authors developed an E-Learning module about putting on strings on a violin and tuning-up the instrument. In addition to providing a step-by-step instruction via audio-visual and textual information, the user is also able to interact with the video employing so called Hotspots in the video. These Hotspots enable the user to draw the sound of a tuned-up string out

of the video and place it in the TunerPad window. This pad resembles a knowledge object. The exact frequency of the sound of the corresponding string is contained in it. This object is provided by the system. The user also will be able to play his violin and record the played sounds with the use of a microphone to verify that the tuning process has been correct. The frequency of the played sound will be displayed on the TunerPad, using the microphone as a sensor input device to pass that information from the environment into the system.

First, the basic concept of this project shall be described, while drawing a parallel between complex instructions and E-learning applications. Subsequently, we will focus on the design and the technical execution of the project, as well as a description of the used technology. In a conclusion, we discuss other possibilities of using the modules of content and future work.

2. Project scenario

Basic idea of the project is to create a multimedia manual for fastening the strings of a violin and tuning the instrument. While the instruction should appeal to the user, it should also offer the possibility of verifying the learning success, and designed in such a way as to allow its use for other applications, thereby ensuring its sustainability.

Conventional manuals describe in longer or shorter text sequences the steps necessary to achieve a certain result. Illustrations facilitate understanding of complex manuals and, in addition, offer the possibility of an overview of the respective step. Thus sectioned, the user is always able to again return to the manual from any point and, with the help of the illustrations, he is able to at least visually verify whether the steps taken have been performed correctly. In complex manuals, the purely textual form is mostly supplemented by graphs. This can help the user in mentally simulating real processes [4]. Mental simulation, however, has its limits. Complex processes require the iterative completion of approximation cycles [5].

Often, if the application level is more complex, a training film is applied, e.g., in displaying an instruction for using a certain software. Usually, this is realized with the help of Flash,- Director,- Authorware-films or other software tools, which enable the sequential procedure being visualized as true to original as possible.

The authors therefore want to explore in this case study, in which way known technology can be applied in order to make manuals more appealing for the user, while at the same time creating added value compared to traditional formats for manuals. Videos are helpful in depicting complex chains of actions and convey information in a compact format. Furthermore, the interaction with a certain medium furthers the interest of users. While conventional manuals are usually not used at all, the authors want to rouse the interest of users with the help of appealing videos and possibilities for interaction.

For this purpose, the authors apply high-resolution film material on a DVD, which is equipped with hyper-video elements (see further [3]) provided by the DVDconnector Technology by the firm micromonics. Moreover, so-called IntelligentPads [6], developed by the Meme Media Laboratory of Hokkaido University, offer the possibility of practical interaction with the instructional film.

The mix of different intertwined media and applications in this project make parallels to E-learning scenarios apparent. Not only is an instruction offered for each step, but it is provided in modules and enriched with further information.

The use of instruction videos basically corresponds to the use of teaching videos. By imitating the action depicted in the video, the user experiences learning success, due to his being able to both analogously perform the actions depicted therein and – in this project – also acoustically and technically verify the correct performance of each step.

The DVDconnector and the IntelligentPad Technology have already been successfully

applied in earlier projects, thereby ensuring that the technical prerequisites for this project can be met [2]. According to the authors, three aspects were elementary in preparing this project. First, a high-resolution video should depict the fastening of the strings and the tuning of the instrument. This video should then be enriched with additional information. Second, the environment of the application should give the user the possibility to interact with the different levels, in particular with the video. Third, the user should be provided with the possibility to verify his success in reality.

2.1. Design of the application environment

Since the instruction video constitutes the main component of this manual, the video and its place in the instruction environment had to be designed first. Detailed information on the design and the creation of the video can be found in chapter 2.2.

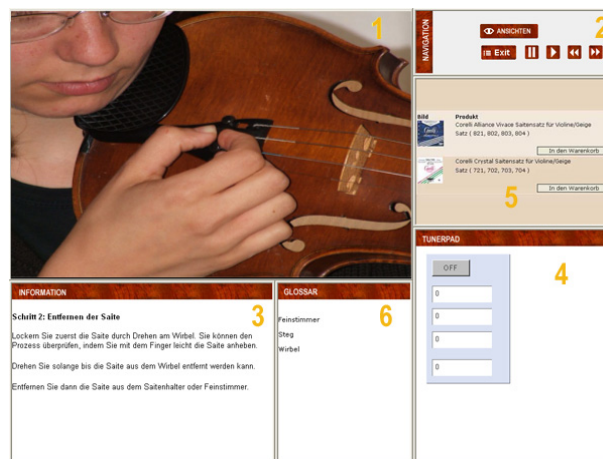


Figure 2: Instruction environment

As can be seen in figure 2, the video window (1) has a central position in the instruction environment. It is supplemented by a navigation window (2), as well as a window for additional information (3), a TunerPad-window (4), which serves as work station for transferring the sounds for tuning the violin, an optional window for e-business possibilities (5) and a glossary window (6), which shows specific terminology while the video is running. The DVDconnector by micromonics comes into operation when creating

this instruction environment. It offers to seamlessly link different media and applications and represents a basis for the essential hyper-video component (see also [1]).

2.2. Creating the instruction video

Generally, a manual follows sequential steps, which create a final product. Consequently, the video is designed likewise, i.e., the single steps from fastening the strings to tuning the instrument are sequentially documented.

The video is subdivided into 2 units. The first part comprises the fastening of the strings, while the second part consists of the tuning of the instrument. Both units constitute modules which can be used independently from each other in other projects.

In addition to the audio commentary, which comments the different steps, the sound is a significant component for tuning the instrument. The authors therefore deem the high quality of the audiovisual media to be essential. We hence picked a DVD as medium, in order to be able to provide for sound and vision in best quality. To generate the medium, we used a Mini DV camera and converted the film data files into DVD format with the help of a WinAVI video converter.

3. Applied technologies

In order to allow for the interaction of the user with the manual, the Meme Media Technology and the DVDconnector Technology was applied. The DVDconnector constitutes the framework and the basis for the hypervideo function, while the IntelligentPad Technology allows for linking different applications with the DVDconnector.

3.1. Meme Media Technology

The Meme Media Technology [6] allows for editing, distributing and linking different program resources with one another.

The IntelligentPad Technology [6], [7] is part of the Meme Media application. IntelligentPad presents every functional component as a two-dimensional object, which

is termed Pad. The user can combine two Pads to form one new Pad and thus equip it with complex functions. Each Pad has an interface, which displays its status data and is termed Slot. If one Pad is inserted into another Pad, the user can define a "Parent - Child Relation" between both Pads and link their Slots, thereby establishing a functional connection between the Pads. In this way, Pads can be connected to one another to form different multimedia documents and applications. Unless otherwise defined, combined Pads can always be separated and changed again. Recently, the Meme Media Technology was enhanced for the reuse of web-based resources, such as web documents and web applications. Web resources can thus be modified and their functions can be combined by "Copy and Paste".

3.2. Tuning function

For tuning the violin, two Pads were implemented, the SoundPad and the TunerPad. The SoundPad represents a sound bearing a certain frequency, which is determined as a value in the Frequency Slot. If the user now clicks on a SoundPad, a defined sound is played-back (e.g. the sound of the A-string). The TunerPad recognizes the pitch recorded by the microphone and visualizes information on the sound, such as volume, basic frequency and the deviation between the original frequency and the desired frequency. A TunerPad transfers these values to the Input, Frequency and Difference Slots. If the user combines a SoundPad with a TunerPad, the SoundPad transfers its frequency to the TargetFrequency Slot of the TunerPad and thus defines the desired frequency for the tuning procedure. If the user clicks on a string of the violin in the video, a SoundPad appears, which contains as information the sound frequency of the tuned violin. The User can now drag this Pad to his work station in the TunerPad window. Upon clicking on it, this SoundPad produces the sound of the tuned string and the user can tune the respective violin string with the help of this sound. In order to tune the string precisely, he can

now copy the SoundPad into the TunerPad. There, the difference in frequency between the predefined sound of the SoundPad and the sound recorded via microphone is displayed. In this way, the user can adjust the frequency of the sound played by him/her to the preset sound of the SoundPad.

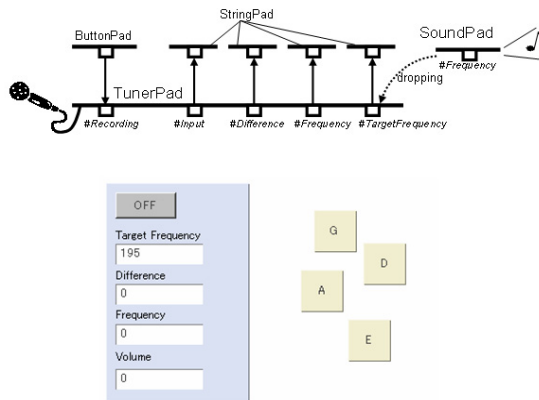


Figure 3: TunerPad and SoundPad

3.3. DVDconnector Technology

The DVDconnector Technology by micro-nomics enables seamless linking of web applications and media. An elementary part of this project is the possibility to link IntelligentPad Technology in the video via Hotspots. Further, offline and online content can be combined and thus enable the user to access current information without experiencing a loss of quality of the supplied media due to compression during data transfer. A detailed account of the applied DVDconnector functions is given in: "Highlights of Algorithmic Learning: Technischer Report." [1].

4. Project implementation

This project shows the fastening of a violin with new strings with the help of a high-resolution video on a DVD medium. The user can thus exactly follow the necessary steps and, if need be, immediately reproduce them. Furthermore, the order of the different steps is supplied in textual form. The user can therefore gain a fast overview of the steps to be taken and select a video sequence analogous to the textual instruction.

Additional information on the instrument can be selected by the user at any time, either via the glossary promptly displayed or via so-called Hotspots in the video. Hotspots are links in the film, which make it possible to design the film as a hyper video (for a detailed description see [1]). Via an online-component, the user can always be supplied with new information.

Further, the producer can integrate the portal for customer support, as well as a dynamic and continually updated FAQ-page for the customer into the manual (see (5) in Fig.2). Further, while appearing in the video, a glossary window displays terminology, which can be selected in the form of a popup-window explaining the respective terms in more detail.

The user can choose between different view options for the instruction video, so the textual information can also be hidden and the video displayed in full view. If the user prefers to have the instructions in textual form and the video as background, the view options allows to exchange the video window with the information window. If the strings are fastened on the violin, the video demonstrates the correct tuning of the violin. The user now has the possibility to draw Pads from the played strings and place them onto his "work station" in the TunerPad window. These Pads play-back the precise sound of the string tuned in the video.

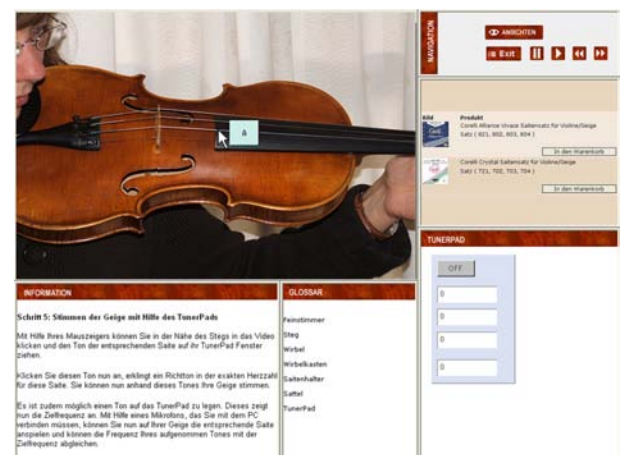


Figure 4: Drawing SoundPad from a string

Thus, the user has all four basic sounds of the violin on his work station (g, d', a' und e'') and can tune the strings of his own violin with the help of a microphone and a pad-interface, as well as the four tune Pads in the TunerPad window. Aside from just receiving information, the user can interact with the manual and, what is more, verify whether he has performed the described steps correctly.

5. Outlook

This case study in whole or parts of it can easily be placed in other projects or E-learning modules due to the modular design of the content. The topic of this project allows wide reuse of the content, since neither technology nor material of the described objects will change. This also guarantees the sustainability of the manual. The authors are convinced that this type of manual motivates the user and guarantees greater success when applied. Since the approach is generic, it can be transferred to other areas of application of operation manuals.

This small case study also serves as a first step to researching sensor input devices in E-learning scenarios using IntelligentPad components. In this project the source for inputting information from the environment into the system was a well defined single source. With the use of a microphone it was possible to represent the frequency of the played violin strings by the user in the system, thus providing the possibility to verify if the instrument has been tuned-up correctly according to the instructions given in the E-Learning application. In further projects the authors may concentrate on implementing further means to input environmental information into a learning system and in doing so taking a step to creating a basis for a virtual laboratory using the DVDconnector and IntelligentPad technologies.

Acknowledgement

The authors want to thank Anne-Kathrin Feld, violin teacher and player for many years, whose expert knowledge in the field of violins was authoritative for providing needed information for the creation of this E-learning scenario about violins.

References

- [1] Jochen F. Böhm. Highlights of algorithmic learning: Technischer report. In [www.http://www.fit-leipzig.de/ Publikationen/Berichte_neueSerie/Report_FIT_Leipzig_2005-05.pdf](http://www.fit-leipzig.de/Publikationen/Berichte_neueSerie/Report_FIT_Leipzig_2005-05.pdf), 2005.
- [2] Jochen F. Böhm, Jun Fujima, Klaus P. Jantke, Aran Lunzer, and Yuzuru Tanaka. Novel technology integration for learning by imitation. In *SITE 2006, Orlando, FL, USA, March 20-24, 2006*.
- [3] P. Grieser, G. und Griegoriev. Erstellung und integration eines hypervideos in das e-learning system damit ein erfahrungsbericht. In *In Jantke, K.P. Wittig, W. und Herrmann, J. (Hrsg.) Von E-learning bis e-Payment 2003. Das Internet als sicherer Marktplatz. Tagungsband LIT03, 24.26. Leipzig. Akademische Verlagsgesellschaft AKA, S.269277, 2003*.
- [4] Kriz S. Cate C. Hegarty, M. The role of mental animations and external animations in understanding mechanical systems. In *Cognition and Instruction 21(4). S.325360, 2003*.
- [5] N. Miyake. Constructive interaction and the iterative process of understanding. In *Cognitive Science 10. S.151177, 1986*.
- [6] Y. Tanaka. *Meme Media and Meme Market Architectures*. IEEE Press and Wiley-Interscience., 2003.
- [7] T. Tanaka Y. und Imataki. Intelligentpad: A hypermedia system allowing functional composition of active media objects through direct manipulations. In *Proc. IFIP89, San Francisco, USA, S.541546, 1989*.

- 01 Rüdiger Grimm, „Vertrauen im Internet – Wie sicher soll E-Commerce sein?“, April 2001, 22 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, ruediger.grimm@tu-ilmenau.de
- 02 Martin Löffelholz, „Von Weber zum Web – Journalismusforschung im 21. Jahrhundert: theoretische Konzepte und empirische Befunde im systematischen Überblick“, Juli 2001, 25 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, martin.loeffelholz@tu-ilmenau.de
- 03 Alfred Kirpal, „Beiträge zur Mediengeschichte – Basteln, Konstruieren und Erfinden in der Radioentwicklung“, Oktober 2001, 28 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, alfred.kirpal@tu-ilmenau.de
- 04 Gerhard Vowe, „Medienpolitik: Regulierung der medialen öffentlichen Kommunikation“, November 2001, 68 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, gerhard.vowe@tu-ilmenau.de
- 05 Christiane Hänseroth, Angelika Zobel, Rüdiger Grimm, „Sicheres Homebanking in Deutschland – Ein Vergleich mit 1998 aus organisatorisch-technischer Sicht“, November 2001, 54 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, ruediger.grimm@tu-ilmenau.de
- 06 Paul Klimsa, Anja Richter, „Psychologische und didaktische Grundlagen des Einsatzes von Bildungsmedien“, Dezember 2001, 53 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, paul.klimsa@tu-ilmenau.de
- 07 Martin Löffelholz, „Von ‚neuen Medien‘ zu ‚dynamischen Systemen‘, Eine Bestandsaufnahme zentraler Metaphern zur Beschreibung der Emergenz öffentlicher Kommunikation“, Juli 2002, 29 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, martin.loeffelholz@tu-ilmenau.de
- 08 Gerhard Vowe, „Politische Kommunikation. Ein historischer und systematischer Überblick der Forschung“, September 2002, 43 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, gerhard.vowe@tu-ilmenau.de
- 09 Rüdiger Grimm (Ed.), „E-Learning: Beherrschbarkeit und Sicherheit“, November 2003, 90 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, ruediger.grimm@tu-ilmenau.de
- 10 Gerhard Vowe, „Der Informationsbegriff in der Politikwissenschaft“, Januar 2004, 25 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, gerhard.vowe@tu-ilmenau.de
- 11 Martin Löffelholz, David H. Weaver, Thorsten Quandt, Thomas Hanitzsch, Klaus-Dieter Altmeyen, „American and German online journalists at the beginning of the 21st century: A bi-national survey“, Januar 2004, 15 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, martin.loeffelholz@tu-ilmenau.de
- 12 Rüdiger Grimm, Barbara Schulz-Brünken, Konrad Herrmann, „Integration elektronischer Zahlung und Zugangskontrolle in ein elektronisches Lernsystem“, Mai 2004, 23 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, ruediger.grimm@tu-ilmenau.de

- 13 Alfred Kirpal, Andreas Ilsmann, „Die DDR als Wissenschaftsland? Themen und Inhalte von Wissenschaftsmagazinen im DDR-Fernsehen“, August 2004, 21 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, alfred.kirpal@tu-ilmenau.de
- 14 Paul Klimsa, Torsten Konnopasch, „Der Einfluss von XML auf die Redaktionsarbeit von Tageszeitungen“, September 2004, 30 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, paul.klimsa@tu-ilmenau.de
- 15 Rüdiger Grimm, „Shannon verstehen. Eine Erläuterung von C. Shannons mathematischer Theorie der Kommunikation“, Dezember 2004, 51 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, ruediger.grimm@tu-ilmenau.de
- 16 Gerhard Vowe, „Mehr als öffentlicher Druck und politischer Einfluss: Das Spannungsfeld von Verbänden und Medien“, Februar 2005, 51 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, gerhard.vowe@tu-ilmenau.de
- 17 Alfred Kirpal, Marcel Norbey, „Technikkommunikation bei Hochtechnologien: Situationsbeschreibung und inhaltsanalytische Untersuchung zu den Anfängen der Transistorelektronik unter besonderer Berücksichtigung der deutschen Fachzeitschriften“, September 2005, 121 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, alfred.kirpal@tu-ilmenau.de
- 18 Sven Jöckel, „Digitale Spiele und Event-Movie im Phänomen *Star Wars*. Deskriptive Ergebnisse zur cross-medialen Verwertung von Filmen und digitalen Spielen der *Star Wars* Reihe“, November 2005, 31 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, sven.joeckel@tu-ilmenau.de
- 19 Sven Jöckel, Andreas Will, „Die Bedeutung von Marketing und Zuschauerbewertungen für den Erfolg von Kinospielefilmen. Eine empirische Untersuchung der Auswertung erfolgreicher Kinospielefilme“, Januar 2006, 29 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, sven.joeckel@tu-ilmenau.de
- 20 Paul Klimsa, Carla Colona G., Lukas Ispandriarno, Teresa Sasinska-Klas, Nicola Döring, Katharina Hellwig, „Generation „SMS“. An empirical, 4-country study carried out in Germany, Poland, Peru, and Indonesia“, Februar 2006, 21 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, paul.klimsa@tu-ilmenau.de
- 21 Klaus P. Jantke & Gunther Kreuzberger (eds.), „Knowledge Media Technologies. First International Core-to-Core Workshop“, July 2006, 204 S.
TU Ilmenau, Institut für Medien- und Kommunikationswissenschaft, klaus-peter.jantke@tu-ilmenau.de