Preface

Ubiquity: the property or ability to be present everywhere or at several places at the same time

"Ubiquitous computing" has become a highly popular topic of current developments in computer science and information technology. A special promise lies in environments that are not only "pervasively" populated by sensors and distributed, mobile and embedded devices, but also utilize the collected data "intelligently". "Ubiquitous" is also one of the most characteristic traits of the Internet, referring to its (potentially) global access and use. The Internet, and in particular the Web, and their being "everywhere", continue to pose key research and application challenges for computing in general and knowledge discovery in particular. This makes *Ubiquitous Knowledge Discovery* a key research area for the coming years.

But knowledge discovery is more than the application of algorithms – it encompasses the whole process of turning data into knowledge: business / application understanding, data understanding, data preparation, modelling, evaluation, and deployment. Users play a pivotal role in this process: they create data, data are related to them, and they are the ultimate beneficiaries of the discovered knowledge. Data-creating activities include the authoring of documents and of references and links between documents, explicit reactions to questions such as the input of registration data, and the behaviour that leaves traces, biometric measurements, etc. in log files. These data are often transformed into user models, i.e. into knowledge about users. These user models, in turn, are the foundation for general-audience or personalized improvements of services and devices, activities which directly benefit the end user or another human stakeholder (the owner of a Web site, a government agency, etc.). Because of this central role of the user, an understanding of users is required for application / business understanding, for data understanding, for the evaluation of discovered patterns, for the deployment of results, and for all other KD activities that depend on these steps. These considerations lead to the recognition that Knowledge Discovery should always be *Knowledge Discovery for Users*.

The purpose of this workshop is to explore the intersection and confluence of the two research areas Ubiquitous Knowledge Discovery and Knowledge Discovery for Users. The workshop results from the activities of the working group "HCI and Cognitive Modelling" of the EU FP6 Coordination Action "KD^{ubiq} – A Blueprint for Ubiquitous Knowledge Discovery Systems". One important goal of the working group is to fully clarify the implications, for users, of ubiquity in KD.Towards this goal, ubiquity needs to be understood in a broad sense, i.e. in all of its senses.

Therefore, in our selection of contributed papers and invited talks, we aimed not only at high quality, but also at a broad, multi-perspective view of ubiquity, and at a balance between more technical and more programmatical viewpoints. Each submission was evaluated by three reviewers from different backgrounds including KD, user modelling, IR, multimedia, and AI.

The contributions discuss the following *dimensions of ubiquity*:[1]

Ubiquity of devices and data *Schwartz, Heckmann, and Baus* consider the issue of sharing sensor data in intelligent environments. They describe the SUPIE architecture for doing that, and a positioning system with data fusion as example application. *Gürses, Berendt, and Santen* also refer to location-aware systems, which they use to describe the ubiquity-specific challenges for the more general problem of protecting users' privacy in ubiquitous knowledge discovery, and for considering real-life systems that typically have many, and many diverse, users/stakeholders. Based on this investigation, they propose a method for multilateral security requirements analysis. Location awareness is also a key feature of the case study presented by *Chongtay*, a medical/therapeutical application (see below).

Ubiquity (distributedness) of processing *Flasch, Kaspari, Morik, and Wurst* consider the distributed organization of data that is employed in collaborative-filtering systems which support users in searching and navigating media collections. They present Nemoz, a distributed media organizer based on tagging and distributed data mining. *Witschel* also focusses on the distributedness of end users in his information retrieval experiments. He shows that global term statistics can be replaced with estimates obtained from a representative reference corpus, which makes it possible to distributed compressed term lists onto mobile devices without taking up too much bandwidth or storage capacity. *Eibe, Hidalgo, and Menasalvas* focus on the distributedness of analysts. They investigate the ubiquitous, collaborative evaluation of KD patterns, and they present a visual evaluation framework to conceptualize visualization according to the diversity of users with different data mining expertise, locations, and devices.

Ubiquity of people and contexts *Kralisch and Berendt* propose the new concept of the "ubiquity of people" as a term to describe people's diversity and the call for the provision of global, equal access. They give an overview of empirical findings on the impact of cultural, linguistic, and other factors on users' interaction with IT, and they propose a first general framework for integrating the new concept into KD. *Anand* focusses on user and context models and their usefulness in collaborative filtering. He empirically evaluates the use of user models that differentiate between the user's long term and short term memory, and discusses the role of context for structuring long term memory. *Paliouras* combines questions of recommendation systems and diverse user groups in a survey of work on discovering user communities from Web usage data. He then investigates which community modelling features can help to handle the additional challenges arising in ubiquitous communication environments, such as additional information overload.

Information ubiquity *Baeza-Yates* discusses ubiquity as a property of information: to really be effectively accessible everywhere and by everyone. He combines this with other forms of ubiquity. He focusses on two forms of information ubiquity: personal data ubiquity (device independence) and Web ubiquity (effective visibility and findability of Web content).

[1] They can also be grouped by other criteria. One example is the grouping and sequence mirrored by the workshop schedule. Therefore, the proceedings are organised alphabetically.

Standards for ubiquity, or: Ubiquity of ubiquity *Chongtay* shows how the establishment of standards for ubiquity helps to guarantee more efficient and advanced functionality. To illustrate the general considerations, she presents an example application for Cognitive Behavioural Therapy that includes the use of the multimodal interaction standard which offers good means to interact with mobile devices, allows one-handed and hands-free operation, and combines standards such as XHTML and VoiceXML (x+v).

We would like to thank all authors and presenters, the reviewers, and our sponsors: KD^{ubiq} and Yahoo! Research.

To find out more about the workshop, visit http://vasarely.wiwi.hu-berlin.de/UKDU06; to learn about (and take part in!) current activities of our working group, come to http://vasarely.wiwi.hu-berlin.de/HCI-ubiq.

August 2006
Bettina Berendt and Ernestina Menasalvas

Organization

Workshop chairs

Bettina Berendt, Humboldt University Berlin, Germany.
Ernestina Menasalvas Universidad Politcnica de Madrid, Spain.

Program committee

Sarabjot Singh Anand, University of Warwick, UK.
Ricardo Baeza-Yates, Director of Yahoo! Research Barcelona, Spain
and Yahoo! Research Latin America at Santiago, Chile.
Jörg Baus, German Research Center for Artificial Intelligence, Saarland Univ.
Shlomo Berkovsky, University of Haifa, Israel.
Josep Blat, Universidad Pompeu Fabra, Spain.
Joanna Bryson, University of Bath, UK.
Marko Grobelnik, Jozef Stefan Institute, Ljubljana, Slovenia.
Dominik Heckman, German Research Center for Artificial Intelligence, Germany.
Anthony Jameson, German Research Center for Artificial Intelligence, Germany
Christian Kray,Informatics Research Institute. University of Newcastle, UK.
Antonio Krüger, Institute for Geoinformatics. University of Münster, Germany.
Dunja Mladenic, Jozef Stefan Institute, Ljubljana, Slovenia.
Mounir Mokhtari, National Institute of Telecommunications of Evry, Paris, France.
Katharina Morik, University of Dortmund, Germany.
Georgios Paliouras, National Centre for Scientific Research, Athens, Greece.
Francesco Ricci, ITC-irst Automated Reasoning Systems, Italy.
Myra Spiliopoulou, University of Magdeburg, Germany.
Stephan Weibelzahl,National College of Ireland, Ireland.
Panayiotis Zaphiris,City University London, UK.

Sponsoring institutions

EU Coordination Action KD^{ubiq}
Yahoo! Research

Table of Contents

Putting the User in Context

Sarabjot Singh Anand

Department of Computer Science
University of Warwick
United Kingdom
s.s.anand@warwick.ac.uk

Abstract. In this paper we discuss the central role played by context in providing adaptive interfaces to a user within a ubiquitous environment. In particular, we discuss approaches to building user models, that incorporate context, from behavioural data streams provided by sensors. To evaluate the efficacy of the user model we simulate a ubiquitous environment using the web and aim to personalize user interactions based on implicit user feedback. Specifically, we use implicit data collected by a web server to simulate data streams from sensors and show how context can be derived from such data and used within user based collaborative filtering. Central to this approach is the user model representation and the similarity metric used to generate the user neighbourhood. While numerous papers have addressed similarity measurement, few papers have discussed how context can play a role within modelling the user and how doing so can affect recommendation quality. In doing so, we also contribute to web personalization literature that has to date mostly ignored the concept of context.

1 Introduction

Human-computer interaction researchers have tirelessly stressed on the need for user-centred design. Their focus has generally been on a static interface design rather than the dynamic adaptation of interfaces to changing user circumstances. On the web, researchers pursuing the goal of delivering personalized content to its users have shifted the focus of user-centred design to adapting to the user's dynamics, highlighting the need for building, using and maintaining user profiles that reflect user preferences and needs [1]. A typical scenario is that of a user searching for an item[1] of interest from a large collection, for example, an interesting news story, a movie or music. Rather than requiring the user to perform an exhaustive search of all items, recommender systems have been developed that recommend, to the user, items that are deemed to be potentially interesting to the user. To achieve this, recommenders depend on user profiles, generally in the form of ratings by the user of other items. These ratings can be provided explicitly or may be derived from implicit indicators of interest such as

[1] We use the term item in this context to represent a service, product or indeed any piece of content on the web with a unique identifier

the time spent browsing the item, the amount of scrolling carried out, the extent of doodling with the mouse [2] etc.

In parallel to the research in recommender systems and personalization, research into context-aware computing has been progressing. The goal being to build systems that adapt to the user's context, with the ultimate aim of addressing what Erickson terms as the "awkward relationship between technology, users and context" [3]. This of course leads to a need to define what we mean by context. Lieberman and Selker define context as consisting of the state of the user, state of the physical environment, state of the computational environment and history of user-computer-environment interaction [4]. In fact they suggest that all non-explicit input to a system can be considered as context and the output of the system itself can be divided into explicit output and changes to the context.

The question is, how can these implicit factors be input to the system? Do we make the user explicitly input this information or do we somehow provide sensors that can detect these implicit factors? Explicit input has disadvantages that the user is distracted from his main train of thought/focus and that transcription errors may occur. If fact, Lieberman and Selker suggest that it is the possibility of eliminating transcription errors by using sensor networks that make the ubiquitous computing paradigm attractive. Of course some aspects of context, for example, state of the environment may be easily collected using sensors while aspects such as the mood of the user cannot be directly sensed, though other factors, that can be used to predict the state, may be collected. This suggests that in addition to collecting data from the environment, predictive models may be required to arrive at a context description of the user.

For the purposes of this paper, we focus on one specific aspect of ubiquitous computing, that is, the ability of sensor networks to collect information from the environment and share this information with other information appliances [5] with the goal of personalizing the user's "interface" to their environment.

We suggest that due to the lack of truly ubiquitous test environments, we can use the web to simulate such an environment and learn from successes and failures of approaches to personalizing the web. More specifically, we can treat the data loggers on the client or server, used to collect implicit interest indicators for recommender systems, as simulating sensors within a ubiquitous environment. While the recommendation scenario described earlier is typical of someone browsing web sites such as Amazon, it is equally applicable to the a user within a ubiquitous environment, replacing the e-tailer with a bricks and mortar store or indeed the Sunday newspaper. As we navigate through such an environment, we are typically faced with scenarios where choices need to be made. Automated tools that help us navigate these choices are just as welcome in such an environment as they are on the web.

One of the problems faced in ubiquitous environments is the fact that sensor networks can quite easily overwhelm applications that aim to utilise the implicit inputs received to adapt their interactions with users. Methods are therefore required that process these data streams from the sensors, distil features from the

streams that can then be used to predict the context and personalize interactions. This too rings true within the web environment as web servers and client-side data collectors have the potential to collect vast amounts of data. It is only after the data is pre-processed and factors such as linger time, extent of scrolling and latent factors [6–8] are extracted that user-modelling can be contemplated.

Lieberman and Selker's definition of context, while providing a useful starting point isn't readily deployable. For example, the history of the user's interactions with their environment can grow very quickly and it is unclear as to what part of that history defines the context of the current interaction, what parts define explicit interaction and what parts are irrelevant. Clearly not all of the data collected on the user can be used to define context, for example, it is highly unlikely that the fact that I ate breakfast this morning will affect what movie I watch, while it may affect the time and type of lunch I eat. We attempt to address these issues within this paper by taking the view that context defines the "situation" within which a user interacts with the system and hence it forms part of the history stored for the user. From the behaviour of the user we must extract/ predict the context of the user's interaction and use only relevant parts of past experience to personalize future interactions. This view of context recognition and prediction has been made previously however, we differ in our approach to context recognition and prediction as well as the type of sensors uses [9].

Surprisingly, little research as been conducted within web personalization with regard to the identification and prediction of context. In this paper we attempt to begin to address this shortfall so as to make research into personalization more applicable to the ubiquitous computing paradigm.

In Section 2 we describe our user model, that is built around the central concept of user context. We also describe some possible memory models that dictate how the user profile can be used for neighbourhood generation. This is followed, in Section 3 by an evaluation of various algorithms that implement the memory models. Section 4 provides a summary of related work while Section 5 concludes the paper with a discussion on future work planned to extend the work presented in this paper.

2 User Modelling

The focus of this paper is on how context can be identified from user behaviour data and incorporated within the user-based collaborative filtering approach to recommendation. We assume that we have a set of m users, $U = \{u_k : 1 \leq k \leq m\}$, and a set of n items, $I = \{i_j : 1 \leq j \leq n\}$. Let $u_a \in U$, referred to as the *active user*, represent the user whose navigation through I needs to be personalized. In previous interactions, u_a will have either explicitly or implicitly rated a set of items $I_a \subset I$. Let r_a represent the rating function for the active users defined as $r_a : I \rightarrow [0, 1] \cup \perp$ where $r_a(i_j) = \perp \forall i_j \notin I_a$.

We refer to the set $Q_a = I - I_a$ as the *candidate item set* for the user u_a. The goal of the recommendation engine is to select a set of items, $R_a \subseteq Q_a$ consisting of items of interest to the active user.

Social or collaborative filtering is traditionally a memory based approach to recommendation generation, though model based approaches have also been developed. The user model is generally in the form of an n-dimensional vector representing the ratings of the n items in I by the user. Hence in contrast to content-based approaches, social filtering does not traditionally use any item content descriptions. The recommendation process consists of discovering the active user's neighbourhood, N_a, i.e. other users in U that have a similar rating vector to that of u_a, and predicting the ratings of items in Q_a based on ratings of these items by $u_k \in N_a$ [10].

As can be seen from the discussion above, two key aspects of user-based collaborative filtering are the user model and the similarity metric used for defining the neighbourhood N_a. Numerous papers have described novel similarity metrics [11], [6], [12] however little research seems to have concentrated on the user model, that is assumed to be a single rating vector (see Section 4 for a discussion on exceptions). We now discuss our approach to modelling the user that goes beyond a single rating vector.

Our model is inspired by Atkinson and Shriffin's model of human memory [13], which is still the basis of our current understanding of the structure of human memory. The model proposed by Atkinson and Shriffin consisted of three structural components of human memory: the sensory register, the short term store and the long term store. According to their model, when stimulus is presented, it is immediately registered within the sensory register. A scan of the information within the sensory register leads to a search of the long term store for relevant information, leading to the transfer of information to the short term store from the sensory register and long term store. According to the model, the data within the sensory register and short term store decay with time, generally within a very short time period, whereas the long term store is more permanent. In addition to these three structural components, the model also identifies control processes such as transfer between short term and long term stores, storage, search and retrieval within short and long term storage.

This model fits in quite well with our needs. From the perspective of a recommender system, a user interacts with the system through implicit and explicit input provided through sensors. These inputs constitute the current activity of the user and can be thought of as being stored in the short term store (the sensory register being implicit in this interaction). However, there are past activities that the user may have had with the system, stored within a long term store, and some of these may be relevant to the current activity. Hence these relevant interactions from the long term store must be retrieved and transferred to the short term store for processing. Finally, the data concerning the current interaction will be transferred to the long term store for use in the future.

Hence, in modelling the user, u_k, we distinguish between two types of memories: Short Term (STM) and Long Term memory (LTM). The only other recommendation system that we are aware of that makes this distinction within the user model is AIS [14]. Our user model differs in a number of ways from that used in AIS in both, its representation and interaction models (see Section 4).

Key processes involved within this model are:

- identification of relevant items in LTM
- indexing of LTM memory
- transfer/storage of item ratings from STM to LTM

We model STM using a rating vector as described above. The user STM consists of items rated in the current active interaction of the user with the recommendation engine.

The user's long term memory is modelled as a set of two-tuples $\langle c_{k_i}, r_{k_i} \rangle$ where, r_{k_i} is a model of a previous interaction between the user and the recommender system and c_{k_i} is the context within the which the interaction modelled by r_{k_i} took place. In doing so we subscribe to Suchman's notion of users interacting with their environment through "situated actions" [15] with context playing the important role of defining the situation within which the actions took place.

2.1 Memory Interaction Models

In this section we identify three distinct models of how the short and long term memories of a user interact during recommendation generation. The models differ in terms of how LTM and STM are used to generate the active user's neighbourhood.

The first model coincides with current approaches to user-based collaborative filtering, where the neighbourhood is defined using all the item ratings from short and long term memory, essentially disregarding the context within which the ratings were made. We refer to this model as the *Inclusive Memory Model*.

The second model is based on the assumption that a user's rating of items gradually evolves, akin to concept drift in learning of classification functions [16] as opposed to being determined by the user's context. Hence the short term memory is "enriched" by using items most recently rated by the user. We refer to this as the *Temporal Memory Model* as its behaviour is based more on the recency of the ratings as opposed to the user context.

The third model uses only those r_{k_i} from the user model to enrich the short term memory that model interactions that took place within the same context as the context of the current interaction. We refer to this as the *Contextual Memory Model*.

2.2 Modelling Context

While context has been defined in numerous ways [17], we use the definition by Day [18] as the basis for our work, that context is "any information that can be used to characterise the situation of an entity". While the observed user behaviour is induced by this underlying context model, the context itself is not observable.

This assumption frees us from limiting our definition of context to a fixed set of attributes such as location, time or identities of nearby individuals/objects as

is commonly done in ubiquitous computing [19] nor do we assume the existence of a concept hierarchy [20] or pieces of text [21] as is often assumed in information retrieval applications. In fact, we suggest that the precise nature/ representation of the context is not as important as recognizing the existence of, and accurately predicting the user context from, a set of known contexts (states).

Let $C = \{c_1, c_2, ..., c_d\}$ be a set of d states corresponding to the set of d contexts within which a user u_k interacts with the system. We model context as a stochastic process that is in one of the d states defined by C. The state that the process is in is controlled by a probability distribution. Over a period of time, the user u_k will have interacted with the system, rating the items, I_k. Each of these ratings will have been made within a particular context. Hence, rather than a single rating function, r_k as described previously, we propose the existence of d rating functions, r_{k_i}, one for each context. Thus item ratings are specific to the user and the context within which the item was rated.

Clearly an important design consideration is the value of d, i.e. the number of distinct contexts within which a user may interact with the recommendation system. Deciding on the value of d is similar to the long standing problem in clustering, of choosing the number of clusters within a given dataset. Various methods have been proposed as a solution to this problem based on maximum likelihood and cross validation [22]. Thus we could choose d, in a similar way, so as to maximize the likelihood of the observed user behaviour.

The task of generating recommendations can now be split into

- context identification: As stated earlier, we take the approach that context is implicit within the observable user behaviour. Hence context identification can be viewed as a mapping $f : U \rightarrow C$. Various approaches can be explored for learning this mapping depending on whether or not the process modelling context exhibits Markov dependence.
- rating prediction: This stage corresponds to the current approach to user-based collaborative filtering where a neighbourhood of the active user must be generated and an ordered candidate item list created for recommendation. The key distinction here is that only that rating function, r_{a_i}, is used for neighbourhood formulation that corresponds to the user, u_a's predicted context.

3 Evaluation

3.1 The Data

Server log files from an movie retailer were processed using standard techniques described in [23] for page view, visit and visitor identification. The processed data consisted of 4,417,950 visits and 2,196,773 unique visitors.

For the experiments presented in this paper, we selected visits that had a minimum of 10 rated items and a maximum of 50 rated items[2]. The total num-

[2] The minimum and maximum number of items within a visit were set so as to ensure that atleast 5 items are available for each visit within the test and hidden data sets and to reduce the effect of any rogue spiders, respectively

ber of unique visitors and visits meeting this criteria were 46,280 and 66,511 respectively. A hold out sample of 30% of all visits was selected randomly to form the test data. Of the visits selected as test data, five items were randomly selected from each visit and hidden from the recommender system. The remaining data was used to model the short term memory of the active user. Items within the data set were movie, actor or director pages rated implicitly by the user defined as the log of linger time of the user on the page.

3.2 Recommendation Algorithms

The algorithms used in this paper differ only in their definition of the active user's neighbourhood based on the three different memory interaction models presented in Section 2.1. In addition to the three algorithms based on these memory models, we also used two baseline algorithms, STO and SRLT, that ignore the user's long term memory and randomly select a context of the current user interaction, respectively.

All algorithms use a weighted sum approach to generating a predicted rating for candidate items [11]. The top 100 items, based on their predicted rating, are recommended to the user. The algorithms are briefly described below:

- STO: The STO algorithm ignored the user's long term memory, defining the neighbourhood only based on the item ratings within his short term memory.
- SRLT: The SRLT algorithm is a baseline algorithm that uses the user's item ratings in short term and a random visit from the user's long term memory as the basis for neighbourhood generation.
- SLLT: The SLLT algorithm implements the temporal memory model, defining the user's neighbourhood based on the short term memory and ratings from the previous interaction by the user.
- SFLT: The SFLT algorithm implements the inclusive memory model.
- SCLT: The SCLT algorithm implements the contextual memory model. As the definition of the memory models and context are intentionally broad within this paper, the behaviour of this algorithm needs to be explained further. Rather than explicitly defining a value for d (the number of contexts in the model), SCLT stores a separate rating function for each user visit within the user's long term memory. The algorithm then uses a similarity threshold, sim_t, to associate visits from long term memory with the context of the user's current interaction. Visits in the user's long term memory that have a similarity greater than the threshold are deemed to have taken place within the same context as the current context. Similarity between the current interaction, v_{a_t}, of the user with other visits within their long term memory, v_{a_l}, is defined using the *Generalized Cosine Max (GCM)* metric [6], defined as

$$v_{a_l} \cdot v_{a_t} = \sum_{(i_j, i_f) \in S} r_{a_l}(i_j) \times r_{a_t}(i_f) \times sim(i_j, i_f) \tag{1}$$

  where, each item i_j is defined as an m-dimensional vector consisting of ratings by the m users in U, $sim(i_j, i_f)$ is an item similarity function, which for the

purposes of this evaluation is calculated as the cosine similarity between items i_j and i_f $(i_j \cdot i_f)$, and $S \subset I_a \times I_b$ is computed as shown in the algorithm in Table 1. An alternative to using cosine similarity to calculate item similarity, when an item knowledge base is available has previously been presented in [6].

Algorithm CalculateS(I_a,I_b,sim(.,.))

Let I_c refer to the smaller of the two sets I_a and I_b, and I_d to the larger of the two.
For $1 \leq i \leq |I_c|$ and $1 \leq j \leq |I_d|$ create the triple $(i, j, sim(i, j))$.

Sort the resulting set of triples in descending order by similarity, creating a list of triples $x_t = (i_t, j_t, sim_t)$ where $1 \leq t \leq N$ and $N = |I_c||I_d|$
S = ϕ; $S_c = \phi$; $S_d = \phi$
For $1 \leq t \leq N$
    If i_t not in S_c and j_t not in S_d
        Add i_t to S_c and j_t to S_d
        Add (i_t,j_t) to S
return S

Table 1. Algorithm for determining S

3.3 Evaluation Metrics

For each algorithm we limited the neighbourhood size to be 50 and the number of recommendations generated to 100. For measuring the accuracy of the recommendation engines we use precision, recall and F1 [24]. Precision and Recall are traditionally used in information retrieval. To use these metrics we categorised items into those "liked" and "disliked" by the user based on a threshold rating value of 3. Hence true positives were those items that had a predicted as well as actual rating of greater than 3. True Negative, False Negative and False Positives were similarly defined. F1 is the harmonic mean of Precision and Recall.

3.4 Results

In addition to the algorithms described in Section 3.2, we also evaluated variants of these algorithms that ignore the ratings in short term memory using only the long term memory. These represent the ability of these algorithms to generate useful recommendations at the start of a user interaction, i.e. when the short term memory is empty. These algorithms are referred to in the evaluation section using the notation $\langle algorithm \rangle 0$. Note that we also include an evaluation of SCLT0

even though, the short term memory is used to derive the user context in this case, i.e. choose visits from LTM with similarity to STM greater than sim_t.

Table 2 shows the effect of the similarity threshold on the precision, recall and F1 when the short term memory itself is large. Table 3 shows the same results when the short term memory is small[3]. In the both tables, the results shown in parentheses are when only long term memory is employed for recommendation generation.

A number of conclusions can be drawn from these tables:

- short term memory has a positive effect on the recall. However it has a negative effect on precision.
- when the short term memory is larger, long term memory can add noise to the recommendation process and hence a larger similarity threshold (0.9, in this case) provides best results[4] as compared with smaller short term memory (0.2, in this case).

Table 4 and 5 show the results of the evaluation of the recommendations generated by the algorithms described in Section 3.2, in the presence of the two short term memory sizes. These results highlight the following:

- in the absence of short term memory, SFLT provides recommendations with high precision but low recall[5]. Further, the F1 value of SFLT0 is comparable to that of SCLT0 in the case when the short term memory has lower cardinality. This can be attributed to the lack of accuracy in calculating the similarity of the current interaction with past interactions.
- SLRT0, that use a random past interaction for neighbourhood formulation, outperforms SLLT0, which uses a basis similar to that proposed in [14]. Hence recency of interaction in itself does not appear to be a useful basis for selecting past ratings by a user.
- For both sizes of short term memory, the STO and SCLT algorithms present similar accuracies, with SCLT doing marginally better. This could imply one of two things.
  - the usefulness of long term memory is overrated for recommendation generation, especially when the short term memory is large enough.
  - better models for identifying context are required. Note that the SCLT algorithm is only one and that too a rather simplistic approach at implementing the contextual memory model described in Section 2.1.
  - the SFLT algorithm, that corresponds to the current wisdom in recommender systems of using all available user ratings, performs worse than SCLT and STO.
- all the algorithms improve on the Random Neighbourhood baseline.

[3] For generating these results we switched the hidden and test data sets generated as described in Section 3.1

[4] defined using the F1 measure

[5] Note that we are not including SCLT0 in this comparison as it does not correspond to a situation where the short term memory is empty

Sim. Thres.	Precision	Recall	F1
0	79.92% (83.79%)	7.87% (6.89%)	0.143 (0.127)
0.1	81.4% (84.3%)	9.1% (7.1%)	0.164 (0.131)
0.2	82.2% (84.9%)	9.3% (6.8%)	0.164 (0.126)
0.3	82.1% (84.9%)	9.3% (6.4%)	0.167 (0.119)
0.4	82.0% (85.28%)	9.26% (6.19%)	0.167 (0.115)
0.5	82.0% (85.4%)	9.25% (5.9%)	0.166 (0.111)
0.6	81.9% (85.5%)	9.22% (5.8%)	0.166 (0.108)
0.7	82.0% (85.6%)	9.17% (5.7%)	0.165 (0.106)
0.8	82.0% (85.6%)	9.14% (5.6%)	0.164 (0.105)
0.9	82.0% (85.6%)	9.1% (5.59%)	0.163 (0.105)
1.0	82.1%	9.09%	0.163

Table 2. Effect of Similarity Threshold (Small Short Term Memory)

Sim. Thres.	Precision	Recall	F1
0	80.45% (84.2%)	8.22% (5.94%)	0.149 (0.11)
0.1	81.4% (84.2%)	9.15% (6.36%)	0.164 (0.118)
0.2	81.87% (85.34%)	9.69% (6.68%)	0.173 (0.124)
0.3	82.48% (85.05%)	9.96% (6.71%)	0.177 (0.124)
0.4	83.3% (85.5%)	10% (6.71%)	0.179 (0.124)
0.5	83.9% (85.9%)	10.28% (6.65%)	0.183 (0.123)
0.6	83.8% (85.78%)	10.29% (6.59%)	0.183 (0.122)
0.7	83.8% (86%)	10.34% (6.53%)	0.184 (0.121)
0.8	83.8% (86.4%)	10.37% (6.53%)	0.184 (0.121)
0.9	83.8% (86.2%)	10.38% (6.51%)	0.184 (0.121)
1.0	83.8%	10.38%	0.184

Table 3. Effect of Similarity Threshold (Large Short Term Memory)

Algorithm	Number of Unique Visitors	Number of Visits	Number of Recommendations	Precision	Recall	F1
RandomNeighbour	561	681	1376	78.3%	0.98%	0.019
STO	2389	3997	9654	82.1%	9.09%	0.163
SFLT0	2036	3320	8424	83.79%	6.89%	0.127
SFLT	2406	3860	9188	79.92%	7.87%	0.143
SLLT0	1334	2017	4015	77.26%	3.34%	0.064
SLLT	2362	3773	7926	81.35%	7%	0.128
SRLT0	1832	2621	5496	78.03%	4.64%	0.087
SRLT	2352	3688	8098	80.97%	7.14%	0.131
SCLT0	1690	3002	8584	84.3%	7.1%	0.131
SCLT	2562	4341	10477	82.2%	9.3%	0.167

Table 4. Evaluation Results (Small Short Term Memory)

Algorithm	Number of Unique Visitors	Number of Visits	Number of Recommendations	Precision	Recall	F1
RandomNeighbour	546	630	848	80.4%	1.7%	0.033
STO	1821	2735	4277	83.7%	10.3%	0.183
SFLT0	1120	1630	2608	84.2%	5.93%	0.11
SFLT	1613	2234	3579	80.45%	8.22%	0.149
SLLT0	669	858	1245	77%	2.7%	0.052
SLLT	1729	2405	3716	83.1%	8.98%	0.162
SRLT0	907	1125	1658	78.35%	3.73%	0.071
SRLT	1728	2428	3763	83.39%	9.13%	0.164
SCLT0	975	1545	2874	85.5%	6.71%	0.124
SCLT	1819	2738	4283	83.8%	10.38%	0.184

Table 5. Evaluation Results (Large Short Term Memory)

4 Related Work

Research on recommender systems has largely ignored the issue of user interest dynamics. One exception to this is the Adaptive Information Server (AIS) [14]. In AIS, Billsus and Pazzani, distinguished between long term and short term interests of a user. They used *tfidf* scores for words appearing in the last 100 documents accessed by the user for modelling short term interests while long term interests were modelled using *tfidf* scores for words appearing in all documents accessed by the user. Our user model, while consisting of a long and short term memory is very different from that introduced by Billsus and Pazzani 2. Firstly, the long term memory incorporates the notion of context and hence is not simply a single vector describing the user's long term interests as is the case in AIS. Incorporating context within long term memory allows us to go beyond the notion of concept drift over time to include cyclic interests of users. Secondly, in AIS, long term memory is only invoked if the short term memory is not able to suitably classify a news story as being of interest or not to the user. In our case, portions of long term memory deemed to have originated in previous user actions within a similar context to the current interaction always augment the short term memory. It is worth noting that the modelling of long term and short term memory and their interactions during an activity is still an open question actively researched in psychology literature [25].

Koychev et al. extended the work by Billsus by suggesting a continuous weighting function that weighted previous interactions on a continuous scale with larger weights being assigned to more recent interactions [16], rather than using a fixed number of recent interactions to form part of the short term memory.

Sheath proposed an evolutionary approach where a population of profiles for each user was maintained [26]. Each of the profiles represented user interests in previous interactions. As these interests changed the population was expected to converge towards a profile that reflected current user interest.

O'Connor and Herlocker investigated the use of clustering items in an attempt to partition the item space to reduce sparseness [27]. Recommendations

are generated by calculating neighbourhoods for the active user and generating recommendations from each of the item partitions. They found that while this approach improved scalability, it had mixed results on accuracy. Each of the item clusters can be viewed as being defined by a different context and hence you would expect that depending on the active user's context, one item space partition may be more relevant to the user than other.

5 Conclusions and Future Work

In this paper, we explored the relationship between context-aware computing, web personalization, cognitive science and ubiquitous computing. Specifically, we explored how ideas from context-aware computing and cognitive science can be brought to bear on current research in web personalization. We also explored how the web can provide a useful simulation for testing out techniques applicable to adapting a user's interface to his environment within a ubiquitous computing setting.

Specifically, we revisited user model representation for recommender systems. We proposed a model that differentiates between the short and long term memory of a user. Further we defined the role of context of previous user interactions in indexing/ structuring the users long term memory. We also proposed and provided one possible implementation for three memory interaction models. Evaluation of these models presented interesting results that suggested that the current wisdom in recommender systems, that suggests that the more user ratings, the better, may not be optimal. The results also suggest a need for further research into the role of context within such a user model.

While we presented our view on context as a hidden stochastic process, within this papers, we did not model it explicitly. In the future we would intend to investigate the use of Hidden Markov Models and other related techniques to explicitly model context and incorporate it into the recommendation process.

Additionally we have only evaluated the different memory models based on the resulting accuracy of the recommendations. However, there are a number of other dimensions along which recommender systems can be evaluated. One such dimension is the diversity of the recommendations. Intuitively we would expect the recommendations generated using the Inclusive memory model to be most diverse as the model includes item ratings collected from the user under varying contexts. We intend to explore this issue in our future work.

Finally, we would like to extend the system presented in this paper by introducing data streams from additional sensors, characterizing the different types of data that can be collected in a ubiquitous environment including navigational and preferential data and generalizing user profiles to span across multiple domains.

References

1. Anand, S.S., Mobasher, B.: Intelligent techniques in web personalization. In Mobasher, B., Anand, S.S., eds.: Intelligent Techniques in Web Personalization.

LNAI 3169. Springer-Verlag (2005) 1–37
2. Claypool, M., Le, P., Waseda, M., Brown, D.: Implicit interest indicators. In: Proceedings of the 6th International Conference on Intelligent User Interfaces. (2001) 33–40
3. Erickson, T.: Some problems with the notion of context-aware computing. Communications of the ACM **45**(2) (2002) 102–104
4. Lieberman, H., Selker, T.: Out of context: Computer systems that adapt to, and learn from, context. IBM Systems Journal **39**(3 & 4) (2000)
5. Norman, D.A.: The Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution. The MIT Press, Cambridge, Massachusetts (1998)
6. Anand, S.S., Kearney, P., Shapcott, M.: Generating semantically enriched user profiles for web personalization. ACM Transactions on Internet Technologies **7**(4) (2007)
7. B. Mehta, T. Hofmann, P.F.: Cross system personalization by factor analysis. In: 4th Workshop on Intelligent Techniques in Web Personalization, (Technical Report, WS-06-10, AAAI Press) 10–18
8. P. Symeonidis, A. Nanopoulos, A.P.Y.M.: Scalable collaborative filtering based on latent semantic indexing. In: 4th Workshop on Intelligent Techniques in Web Personalization, (Technical Report, WS-06-10, AAAI Press) 1–9
9. R. Mayrhofer, H.R., Ferscha, A.: Recognizing and predicting context by learning from user behavior. Radiomatics: Journal of Communication Engineering, special issue on Advances in Mobile Multimedia **1**(1) (2004) 30–42
10. Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., , Riedl, J.: Grouplens: An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 Computer Supported Collaborative Work Conference. (1994)
11. Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 1999 Conference on Research and Development in Information Retrieval. (1999)
12. Ziegler, C., Lausen, G., Schmidt-Thieme, L.: Taxonomy-driven computation of product recommendations. In: Proceedings of the ACM Confernece on Information and Knowledge Management. (2004) 406–415
13. Atkinson, R.C., Shiffrin, R.M.: Human memory: A proposed system and its control processes. Psychology of Learning and Motivation **2** (1968) 89–195
14. Billsus, D., Pazzani, M.J.: User modeling for adaptive news access. User Modelling and User-Adapted Interaction **10** (2000) 147–180
15. Suchman, L.: Plans and Situtated Actions. Cambridge University Press, Cambridge, UK (1987)
16. Koychev, I., Schwab, I.: Adapting to drifting user's interests. In: Proceedings of ECML2000/MLNet workshop on Machine Learning in the New Information Age. (2000)
17. Dourish, P.: What do we talk about when we talk about context. Personal and Ubiquitous Computing **8**(1) (2004) 19–30
18. Day, A.K.: Understanding and using context. Personal and Ubiquitous Computing **5**(1) (2001) 4–7
19. Schilit, B., Theimer, M.: Disseminating active map information to mobile hosts. IEEE Network **8** (1994) 22–32
20. Parent, S., Mobasher, B., Lytinen, S.: An adaptive agent for web exploration based on concept hierarchies. In: Proceedings of the 9th International Conference on Human Computer Interaction. (2001)

21. Kraft, R., Maghoul, F., Chang, C.C.: Y!q: Context search at the point of inspiration. In: Proceedings of the ACM Conference on Information and Knowledge Management. (2005) 816–823
22. Smyth, P.: Clustering using monte carlo cross-validation. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. (1996) 126–133
23. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems **1**(1) (1999)
24. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems **22**(1) (2004) 5–53
25. Burgess, N., Hitch, G.: Computational models of working memory:putting long-term memory into context. Trends in Cognitive Science **9**(11) (2005) 535–541
26. Sheth, B.: A learning approach to personlized information filtering. Masters Thesis, Massachusetts Institute of Technology (1994)
27. O'Connor, M., Herlocker, J.: Clustering items for collaborative filtering. In: Proceedings of ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation. (1999)

Information ubiquity
(Invited Talk – Abstract)

Ricardo Baeza-Yates

Director of Yahoo! Research Barcelona, Spain
and Yahoo! Research Latin America at Santiago, Chile
http://www.dcc.uchile.cl/˜rbaeza/

Current trends emphasize the ubiquity of devices which imply the ubiquity of people. However, device ubiquity, using for example wireless networks, does not solve the problem. We also need data and information ubiquity. This kind of ubiquity has two main facets that we explore in this talk:

– One facet is personal data ubiquity, which should be device independent. A first attempt of this case is Yahoo! Go. In the future we will need even a more seamless ubiquity.
– Another facet is the information or service provider ubiquity. Currently the main problem in this case is Web ubiquity which relates to the interaction of web site design and search engines. (Information is Web-ubiquitous if it not only exists on the Web, but is effectively visible and findable.) In the future this can be improved by more cooperation between all the parts involved.

Standards for ubiquity
(Invited Talk – Abstract)

Rocio A. Chongtay

Mobile People, Copenhagen, Denmark
http://www.mobilepeople.dk/

The fast growing development of communication technologies as well as the widespread and still increasing use of mobile devices for both voice and data communication have motivated the creation of new applications to access information any time anywhere. The development of mobile applications face challenges such as interoperability between the high diversity of devices, limited memory capabilities, small screens, reduced keypads, etc.. The establishment of standards for ubiquity helps to guarantee more efficient and advanced functionality.

Some reflections and possibilities of standards used for ubiquity will be presented in this talk. Furthermore, an example of a location-based application that includes inherited standards re-used from the non-ubiquitous standards, will be presented.

The location-based application was designed to aid the treatment of a phobic condition known as fear of heights (acrophobia). A phobia is an irrational fear to some situations or things that interferes with the functioning of the individual that suffers from it. The location-based system has been designed to aid in some of the training steps of one of the most common and successfully used treatments for phobic conditions known as Cognitive Behavioral Therapy (CBT), by helping people to learn to detect thinking patterns which trigger the irrational fear and replace them with more realistic ideas. One of the main components of CBT is gradual exposure, in which the patient learns to confront the stimuli that provoke the anxiety. The system is designed to serve as an interactive guide during the steps of the gradual exposure.

The applications design includes the use of the multimodal interaction standard which offers good means to interact with mobile devices, allows one-handed and hands-free operation, and combines standards such as XHTML and VoiceXML (x+v). The advantages and disadvantages of this design will be discussed in the context of the current state of the standards for ubiquity.

Ubiquitous evaluation of KDD results: a visual framework

Santiago Eibe[1], Miguel Hidalgo[1], and Ernestina Menasalvas*[1]

Facultad de Informatica, Universidad Politecnica, Madrid, Spain
seibe@fi.upm.es, mahidalgo@alumnos.upm.es, emenasalvas@fi.upm.es

Abstract. The diversity of users, devices, contexts and ubiquity in general raises new challenges for knowledge discovery research. In particular, ubiquitous evaluation of knowledge patterns would be needed. Data is being generated in a continuous fashion and methods to on line extract knowledge and make it available for deployment is a need. In this scenario, it is of uttermost importance collaborative evaluation and interpretation of results as data mining experts being available to interpret and evaluate results in a continuous way are not any longer a reality. HCI plays an important role in the definition of requirements of a visual environment for ubiquitous evaluation to capture differences in user experience when analyzing results. Consequently in this paper we present a visual evaluation framework to conceptualize visualization according to diversity of users interpreting results in the broadest meaning: different data mining expertise, location, devices. The framework will provide users with the possibility to built personalized visualizations. As a by-product of the approach, the framework makes it possible to learn from the users usage of the environment so to adapt to users needs in the future.

1 Introduction

Data generated by ubiquitous environments such as sensor networks (car, satellite, ...), peer-to-peer and wireless mobile networks, mobile/embedded computing devices, web mining or grid computing platforms is a new challenge for KDD research (KD-ubiq). In particular, evaluation and deployment of patterns has to be revisited.

According to CRISP-DM [14] in the Assessment task: "The data mining engineer interprets the models according to his domain knowledge, the data mining success criteria and the desired test design". In ubiquitous knowledge discovery applications the data miner would also to be "ubiquitous" in the broadest scope: expertise, domain knowledge, device, ... what raises challenges and opportunities to be taken into account when designing tools for evaluation of patterns in this scenario:

- Diversity of users, devices and context in general. Data and knowledge presentation must respect the limits imposed by the combination of ubiquitous devices and general human perceptual and cognitive limitations (e.g., display resolution), and the specific requirements on accessibility posed by people's diversity [7].

* This work has been partially supported by Ministerio de Educacion y Ciencia (Spain) under project TIN2004-05873

– Location transparency. Data mining in mobile/embedded devices faces various challenges due limited resource availability and software and hardware heterogeneity to name but a few.
– Data streams. Data Mining methods must be able to keep track of models and patterns that may change over time, due to external factors or through internal processes.
– Feedback, iteration and cooperation support. Ubiquitous evaluation of patterns would require a collaborative platform for data miners engineers and expert domain collaboration to asses results.

Before proceeding to define such an environment, the definition and conceptualization of people diversity in this environment together with context has to be defined. New metaphors and formats for data/knowledge presentation must be found where choices are presented to the user of the visualization (either user or service). Consequently, we propose in this paper to define a framework for visual data mining evaluation in ubiquitous environment. The reason for a framework is to be able to conceptualize and abstract needs, some still undefined, and requirements to be implemented in data mining environments that will help from the HCI perspective to improve ubiquitous knowledge discovery processes and be a first step towards autonomous data mining components.

Such an environment will make it possible for users (different expertise, contexts, ...) to interpret results regardless of location, context and expertise in a collaborative way bridging the gap between data mining tools results and consumers of this knowledge. In the proposed approach visualization acts as a catalyst in order to reach two basic goals:

– fill the gap between data mining tools and the user to interpret results
– learn from the experience of the usage and interactions of users in this environment towards future autonomous data mining components

In order to achieve these two goals semantics of the underlying process (mining and business) has to be captured and integrated in the framework. Therefore, in the following sections we present a framework for visual evaluation (VEF, Visual Evaluation Framework) that supports user personalized and device adapted visual analysis of patterns. Besides user and device descriptions the framework also includes task descriptions to minimize the number and importance of user (analyst of business expert) mistakes in the evaluation process.

To show a first deployment of the proposed approach in this paper we also present the implementation of the framework for evaluating association rules obtained from analyzing web logs of the search engine of the department web site.

The rest of the document is organized as follows. In section 2 we show relevant approaches to ubiquitous computing, visualization, visual data mining and other related works. In section 3 the visual framework (VEF) is presented in which components of the model (scenes, actors and channels) are deeply explained. In section 4 the deployment of the visual approach for a particular real case is shown. To end with, section 5 presents the conclusion and outlook.

2 Related Work

The field of ubiquitous computing [1, 35, 21] sets new challenges for knowledge discovery [31] and in particular for visual representation of discovered knowledge patterns. The demand for visual and interactive analysis tools is particularly pressing in ubiquitous data mining scenarios. In general, the abundance of data available fosters the need of developing tools and methodologies to help user in extracting significant information. In ubiquitous environments data are available from many devices, locations and they are changing continuously. Research on data mining on data stream deals with the problem of concept shift [2, 19, 18, 3, 32] but little attention is paid to visualization.

Information Visualization [13, 4] is a rapidly growing research area which aims to provide visual depictions of very large information spaces. It covers interdisciplinary areas such as Information Retrieval [5, 9] or Human Computer Interaction [17]. An active and paradigmatic application area for visualization is related to the World Wide Web. Visualizing the World Wide Web requires new front-end tools to aid navigation and search interfaces that follows the structure of information. In [4] tools to navigate through a set of documents that are clustered for different user needs and where each cluster is labeled with related words are presented. More recently and joined to the advent of the Semantic Web some approaches such as [11] and MoSeNa [6], are looking for semantic navigation structures modeling.

In the recent years, Visual Data Mining (VDM) [29, 28, 27, 30] is exploiting data mining algorithms together with information visualization techniques, mainly applied to structures such as hierarchies and networks. As a consequence, a variety of powerful methods and tools such as [24, 22, 16] that visualize data in an interactive way has been developed. The tools come with a large set of components and control panels to configure the visualizations. These components are integrated into a coherent framework that enable users to do an interactive visual analysis across complex visual structures as scatter plots, parallel coordinates, tree maps, graphs, etc.

However, a more general and flexible [8] solution where tasks and users are included is needed. Moreover, in the new Distributed Data Mining [25, 33, 34] or Device Mining [26] environments, visualization plays an integral role to integrate seamlessly different components. In such scenarios, teams of people even geographically separated develop different parts of the processing and tools for collaborative evaluation are a need.

Concerning association rules visualization, the work presented in [36] covers the association rules model, in which rules are graphically displayed with colored bars where color and length of bars varies to represent different values of confidence and support. This way to display rules is not as intuitive for a user not familiar with association rules. On the other hand, in order to explore specific rules, the user must change the support and confidence values without the possibility of observing the full set of rules.

In [10] a way to assist the user in the analysis of rules by using graphs is proposed. Graph visualization is an excellent approach as it allows the user to observe the overall set of rules and by means of interaction techniques the exploration is assisted. This way of representation makes it easy to understand relationships among antecedent and consequent rather than by using textual representation.

With relation to stream mining and visualization in [12] an approach to represent data from sensor data and network intrusion is presented. Components of the system like stream mining visualizer display association rules in 3-D bar chart where support and confidence are the height and color of bars respectively.

Nevertheless, evaluation on patterns in ubiquitous knowledge discovery demands distance between tools and users to be minimized. This is to say, non expert data miners would have to be able to evaluate results in collaborative environments where the tool would assist them to create customizable visualizations and learns from their experience.

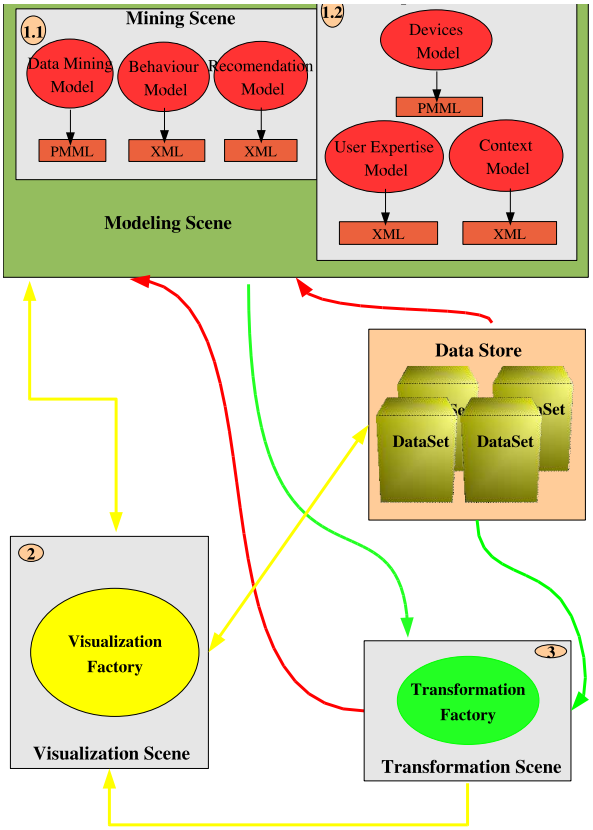


Fig. 1. Visual Evaluation Framework

3 Visual Evaluation Framework

Ubiquitous data mining reopens the not fully solved problem of data mining patterns evaluation adding the whole set of parameters that play their role when describing ubiq-

uity: multiple and different sources of data, diversity of users, context and devices. Therefore, the challenge is how to integrate them in a global model where collaborative and autonomous mining could be possible. In this setting we believe that customizable visual environment can be the solution to integrate the different elements and roles participating in the evaluation process. In order to accomplish this integration we propose to design a general framework that supports a systematic approach for visual mining evaluation. Systematic means that the model satisfies the following requirements:

- Adaptable to different application areas
- Adaptable to users with different backgrounds, settings and expertise level (both domain and mining)
- Able to plug-in modules for visualization, mining and integration

Next, we present the visual evaluation framework (see figure 1) and a case of study where association rules are evaluated (figure 2).

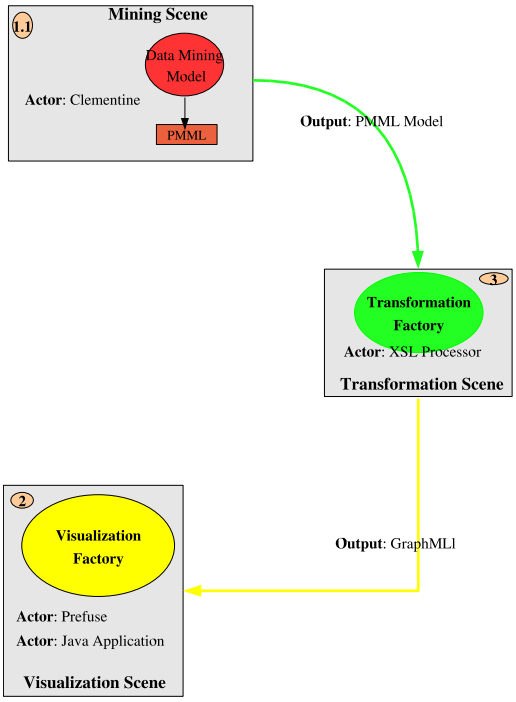


Fig. 2. Visual Evaluation Process

3.1 Basic Abstractions

The framework is compound of three basic abstractions: 1) scenes with 2) actors and 3) channels connecting scenes. The idea behind this abstract model is that visual analysis of patterns is an activity that can be viewed in different planes. Thus, scenes represent working areas of the tasks of the evaluation process either performed by the user with some tools support: actors in the abstraction model or in a fully automatic way without user participation. Each scene is dedicated to some specific set of tasks, visualization tasks, understanding behavior, modeling behavior, ...

According to the tasks being performed, constraints represented in the model as channels will be applied.

It is important to point out that the framework defines a conceptualization problem about the visual analysis of patterns in ubiquitous environments independently of the implementation paradigm used to its deployment.

3.2 Scenes and Actors of the model

Four scenes have been defined in the framework:

- **Mining Scene**: is subdivided in two sub-scenes to represent both the modeling of data mining results (decision trees, association rules, rules to represent the domain, ontologies) and the modeling of ubiquitousness of the environment (miner expertise, contextual limitations, device descriptions).
  Consequently, the Data Mining Model sub-scene represents tasks related to mining model extraction and representation while the Behavior Model and Recommendation Model sub-scene represents the tasks related to domain description and changes.
- **Ubiquitous Scene** represents the models to describe devices, knowledge patterns consumers (miners, business experts, services) and context definitions.
- **Visualization Scene** houses the visualizations producers. For such goal, a visualization factory that generates the diverse visual components (icons, graphs) has been defined.
- **Transformation Scene** hosts the actors dedicated to perform the transformations required to make it possible to visualize results according to the diversities and heterogeneities in the system (user, devices, context, ...).

3.3 Channels

The channels support the semantics underlying the transitions between the scenes of the model. The main goal of channels is to avoid mistakes in the flow of processes to be developed for a proper evaluation. Transitions from one scene to another are defined by channels in such a way that not connected scenes do not allow transitions among the tasks they represent. We have depicted some channels in the figure 1 to represent the most usual transitions between scenes.

Visual evaluation of data mining requires doing tasks in the four scenarios. For example, while the analyst/user of the framework is modeling, the actors/agents support

their activity. Next, when this task has properly been completed, some transformation is needed. Actors in transformation scene support it while results are been visualized in the visualization scene. Cycles of the model represents the iterations needed if going back to either transformation or modeling scene is required in order to complete the evaluation process.

4 Association rules visual evaluation

As a part of the framework development, we present a case of study in which association rules are evaluated. In the first prototype without loss of generality, the following simplifications are adopted:

- Mining scene: In the present approach we assume the mining model computes association rules so recommendation and behavior models will represent the factors of evolution or criteria of the business expert to choose the best rule to recommend depending on the user. However in the first prototype we just concentrate on the mining model. We will also assume that we deal both with the rules and with the frequent item sets together with validity criteria expressed in PMML format [20, 15]. In what actors of the model concerns, there is one to represent in our case Clementine tool that calculates the association model.
- The ubiquitous scene only contemplates ubiquity related with the expertise of the evaluator considering in this case only two roles: expert data miner and domain expert.
- In the transformation model only transformations due to different visualizations of graphs are considered not dealing with problems related to diversity of devices and/or users. In fact, we are going to use a simple transformation mechanism to produce graph-based presentations (see section 4.2).

```
$<$AssociationRule support="11.0" confidence="6.1" antecedent="5"
consequent="2"$>$
$<$!------$>$
$<$Extension name="generation" value="0"/$>$
$<$Extension name="lift" value="1.66"/$>$
$<$Extension name="leverage" value="0.98"/$>$
$<$/AssociationRule$>$
```

Fig. 3. PMML extension mechanism: new attributes aggregated per rule

Therefore, for the purpose of the paper we focus on the visualization scene describing the process in detail.

4.1 Visual Evaluation Process

This process of visual evaluation can be divided in the steps that are showed in the figure 2 showed in page 2:

1. To start with association rules data mining model is produced using Clementine.
2. The data mining model in PMML format (figure 3) is converted to GraphML in the transformation scene. This conversion respond to the requirement of graph structures visualization that has been the metaphor chosen for the experiment. In this case, a XSL style sheet is used to define the transformation process (more detail in section 4.2). The solution in this example is simple but flexible and show us one possible implementation of transformation actors.
3. Visual analysis of the graph that represents the rules model. This includes the design of the visualization in an iterative and interactive process. When an user construct the visualization it could be constrained with some axioms in order to achieve correctness in the analysis task. Channels will act as available templates for visualizations.

4.2 Extension Mechanism

From a real perspective, there could be certain domains that require more information than the one considered by a PMML producer like Clementine. In order to set a complete model we present a PMML extension mechanism introducing an extension element that is used in this case for extending its content (see figure 3).

```
$<$xsl:template match="pmml:AssociationRule"$>$
$<$xsl:element name="edge"$>$
$<$!-- translation from association rules to edges--$>$
$<$xsl:attribute name="source"$>$
$<$xsl:value-of
select="key('num{\_}child',@antecedent)/@id"/$><$/xsl:attribute$>$
$<$!-- when adding the generation value through extension mechanism--$>$
$<$xsl:attribute name="generation"$><$xsl:value-of select="@generation"/$>$
$<$/xsl:attribute$>$
$<$!-- where metrics are created--$>$
$<$data key="supp"$><$xsl:value-of select="@support"/$><$/data$>$
$<$data key="conf"$><$xsl:value-of select="@confidence"/$><$/data$>$
$<$!-- final appearance when adding new metrics --$>$
$<$data key="lift"$> \quad <$xsl:value-of select="@lift"/$>$ $<$/data$>$
$<$data key="leverage"$> \quad <$xsl:value-of select="@leverage"/$>$ $<$/data$>$
```

Fig. 4. XSLT code extract

As main properties, extension mechanism has a *name* and *value* attributes. These specify respectively the name of the extension and the value. For example, to track

changes in association rules a new extension named *generation* could be created. Its value could be determined by the recommendation model.

If different metrics besides support and confidence such as the ones proposed in [23] are desired, they are susceptible to be incorporated with the extension element described above. For instance, figure 4 shows the transformation needed if lift and leverage metrics want to be considered.

4.3 Components in the Visualization Scene

As mentioned above, in this scene the components to build a dynamic and customizable environment for visualizing association rules are represented. Thus, in this case so to assist the user, the graphical interface will provide a rich set of controls such as:

1. WheelZoom control, that zooms in or out the visualization display
2. drag control to move specific nodes as far or close the user wants
3. pan control to move the full visualization and locate it at either closer to the middle or at any corner
4. ZoomToFit Control which adjusts the visualization on the screen bounds
5. ...

Labels have been the other kind of component used to aid the user in the evaluation process. In the case of study labels have been used so to inform the user about the range colors available for the validity metrics in use. They have also been used to inform about initial statistics, number of nodes participating, number of rules represented and so on.

Finally, dynamic controls (clicks, filters, ...) have also been added for interaction for example by clicking into a node the user can see the number of times the attributes represented bu the node appears as antecedent or consequent in a rule.

5 Conclusions

Ubiquitous knowledge discovery is related to knowledge discovery in ubiquitous environments. This is a new challenge for KDD researchers, in particular in this paper we have analyzed the problem of ubiquitous evaluation of data mining patterns stressing over the lack of a collaborative ubiquitous environment to be provided for data miners and domain experts to analyze patterns obtained in KDD processes.

In the new scenario of sensors gathering data from different locations, not only collaborative environment would be needed but autonomous data mining components. Nevertheless, for data mining process to be autonomous the user experience and behavior when analyzing patterns has to be further investigated.

For all these reasons we have proposed in this paper a visual evaluation framework whose aim is twofold: on the one hand, the framework represents a conceptualization of the problem of visualizing patterns in ubiquitous scenario in the broadest sense (users, devices, contexts). On the other hand, components of the framework make it possible to collect information of the knowledge consumer behavior.

The framework is based on 3 basic abstractions: scenes, actors and channels that altogether capture the semantics and context of the user in a ubiquitous evaluation scenario.

In the paper we have also presented the materialization of the framework for the case of association rules evaluation using graphs.

In the near future we will extend the materialization not only to work with association rules and graphs but with all data mining results and more visualization. This will make it possible to analyze lacks of the proposed framework so we can improve it. It is also under research the mechanism to learn from the user experience.

References

1. Gregory D. Abowd and Elizabeth D. Mynatt. Charting past, present, and future research in ubiquitous computing. *ACM Trans. Comput.-Hum. Interact.*, 7(1):29–58, 2000.
2. Charu C. Aggarwal. On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):587–600, 2005. Senior Member-Charu C. Aggarwal.
3. Jess S. Aguilar-Ruiz and Francisco J. Ferrer-Troyano. Visual data mining. *J. UCS*, 11(11):1749–1751, 2005.
4. P. Au, M. Carey, S. Sewraz, Y. Guo, and S.M. Rger. New paradigms in information visualization. In *Research and Development in Information Retrieval*, 2000.
5. R. Baeza-Yates and B. Ribeiro. *Modern Information Retrieval*. ACM Press, 1999.
6. Jorg Becker, Christian Brelage, Karsten Klose, and Michael Thygs. Conceptual modeling of semantic navigation structures: the MoSeNa-approach. In *Proceedings of the fifth ACM international workshop on Web information and data management*, pages 118–125. ACM Press, 2003.
7. Bettina Berendt. Hci and cognitive modelling in ubiquitous knowledge discovery wg6 of kdubiq. http://vasareli.wiwi.hu-berlin.de/HCI-ubiq.
8. K.W. Brodlie, J. Wood, D.A. Duce, J.R. Gallop, D. Gavaghan, M. Giles, S. Hague, J. Walton, M. Rudgyard, B. Collins, J. Ibbotson, and A. Knox. XML for Visualization. In *EuroWeb 2002*, 2002.
9. Peter Brusilovsky and Carlo Tasso. Preface to special issue on user modeling for web information retrieval. *User Model. User-Adapt. Interact.*, 14(2-3):147–157, 2004.
10. Dario Bruzzese and Paolo Buono. Combining visual techniques for association rules exploration. In *AVI*, pages 381–384, 2004.
11. Cristina Cachero and Nora Koch. Conceptual Navigation Analysis: a Device and Platform Independent Navigation Specification. In *In Second International Workshop on Web-oriented Software Technology (IWWOST02), CYTED, D. Schwabe, O. Pastor, G. Rossi, and L. Olsina, editors*, 2002.
12. Y. Dora Cai, David Clutter, Greg Pape, Jiawei Han, Michael Welge, and Loretta Auvil. Maids: Mining alarming incidents from data streams. In *SIGMOD Conference*, pages 919–920, 2004.
13. S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision To Think*. Morgan Kaufmann, 1999.
14. P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth. The CRISP-DM process model.
15. Xml-Based Multimedia Content. Pervasive multimedia markup language (pmml): an.

16. John Ellson, Emden R. Gansner, Eleftherios Koutsofios, Stephen C. North, and Gordon Woodhull. Graphviz - open source graph drawing tools. In *Graph Drawing*, pages 483–484, 2001.
17. A. Dix et al. *Human-Computer Interaction, 2nd Edition*. Prentice Hall, 1998.
18. Francisco J. Ferrer-Troyano, Jess S. Aguilar-Ruiz, and Jos Cristbal Riquelme Santos. Connecting segments for visual data exploration and interactive mining of decision rules. *J. UCS*, 11(11):1835–1848, 2005.
19. Joao Gama, Pedro Medas, and Pedro Rodrigues. Learning decision trees from dynamic data streams. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 573–577, New York, NY, USA, 2005. ACM Press.
20. R. Grossman, S. Bailey, A. Ramu, B. Malhi, P. Hallstrom, I. Pulleyn, and X. Qin. The management and mining of multiple predictive models using the predictive modelling markup language, 1999.
21. Jonathan Grudin. Group dynamics and ubiquitous computing. *Commun. ACM*, 45(12):74–78, 2002.
22. gViz: Visualization Middleware for e-Science.
See website at `http://www.visualization.leeds.ac.uk/gViz/`.
23. Maria Halkidi and Michalis Vazirgiannis. Quality assessment approaches in data mining. In Oded Maimon and Lior Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 661–696. Springer, 2005.
24. Jeffrey Heer, Stuart K. Card, and James A. Landay. Prefuse: a toolkit for interactive information visualization. In *CHI*, pages 421–430, 2005.
25. H. Kargupta, B. Park, D. Hershberger, and E. Johnson. Collective data mining: A new perspective towards distributed data mining. In Hillol Kargupta and Philip Chan, editors, *Advances in Distributed and Parallel Knowledge Discovery*, pages 133–184. MIT/AAAI Press, 2000.
26. Hillol Kargupta, Ruchita Bhargava, Kun Liu, Michael Powers, Patrick Blair, Samuel Bushra, James Dull, Kakali Sarkar, Martin Klein, Mitesh Vasa, and David Handy. Vedas: A mobile and distributed data stream mining system for real-time vehicle monitoring. In *SDM*, 2004.
27. Daniel A. Keim. Visual exploration of large data sets. *Commun. ACM*, 44(8):38–44, 2001.
28. Daniel A. Keim. Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph.*, 8(1):1–8, 2002.
29. Daniel A. Keim, Christian Panse, and Mike Sips. Visual data mining of large spatial data sets. In *DNIS*, pages 201–215, 2003.
30. Haim Levkowitz Maria Cristina Ferreira de Oliveira. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 14(9-3):378–394, 2003.
31. Michael May. Kdubiq (knowledge discovery in ubiquitous environments). http://www.kdubiq.org.
32. Carlos Rojas Olfa Nasraoui, Cesar Cardona. Mining evolving web clickstreams with explicit retrieval similarity measures.
33. A. Prodromidis and P. Chan. Meta-learning in Distributed Data Mining Systems: Issues and Approaches. In Hillol Kargupta and Philip Chan, editors, *Advances of Distributed Data Mining*. MIT/AAAI Press, 2000.
34. F. Provost. Distributed Data Mining: Scaling Up and Beyond. In Hillol Kargupta and Philip Chan, editors, *Advances in Distributed Data Mining*. MIT/AAAI Press, 2000.
35. Daniel P. Siewiorek. New frontiers of application design. *Commun. ACM*, 45(12):79–82, 2002.
36. Dietrich Wettschereck. A kddse-independent pmml visualizer. In Marko Bohanec, Dunja Mladenic, and Nada Lavrac, editors, *2nd International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, August 2002.

Aspect-based Tagging for Collaborative Media Organization

Oliver Flasch, Andreas Kaspari, Katharina Morik, Michael Wurst

University of Dortmund, AI Unit
{flasch,kaspari,morik,wurst}@ls8.cs.uni-dortmund

Abstract. Organizing multimedia data is very challenging. One of the most important approaches to support users in searching and navigating media collections is collaborative filtering. Recently, systems as flickr or last.fm have become popular. They allow users to not only rate but also tag items with arbitrary labels. Such systems replace the concept of a global common ontology, as envisioned by the Semantic Web, with a paradigm of heterogeneous, local "folksonomies". The problem of such tagging systems is, however, that resulting taggings carry only little semantics. In this paper, we present an extension to the tagging approach. We allow tags to be grouped into aspects. We show that introducing aspects does not only help the user to manage large numbers of tags, but also facilitates data mining in various ways. We exemplify our approach on Nemoz , a distributed media organizer based on tagging and distributed data mining.

1 Introduction

Networks allow users to access information in a transparent, location-independent way. The ability to exchange any kind of data and information, independent of ones current geographical location, paves the way for patterns of cooperation that have not been possible before. The collaborative Wikipedia contains more articles than the Encyclopedia Britannica. On the Usenet, each day 3 Terabyte of information is exchanged in more than 60,000 discussion groups. However, in order to make full use of these possibilities, search engines are not enough, but *intelligent mediation* is needed. Mediation refers to the task of making information provided by one user accessible and beneficial for other users. One option is to enrich resources of the Internet with more semantics. The most ambitious project towards this goal is the Semantic Web. While this approach would allow for a better semantic description, it faces considerable acceptance problems. Assuming a global semantic model (a global view of the world) does not reflect the very subjective way in which users handle information. Annotating resources using Semantic Web formalisms requires too much time and is rewarded only indirectly and after quite a while. Hence, an alternative option has recently been proposed, the Web 2.0, where tagging is performed individually at the distributed local sites without reference to a global ontology. Examples are tagging systems such as flickr, del.icio.us, or last.fm. They allow users to assign arbitrary tags

to content. Formal ontologies are replaced by so-called "folksonomies", that do not depend on common global concepts and terminology. The ability to freely choose concepts used for annotation is the key that led to the high acceptance of these systems by users. One drawback of these approaches is, however, that automatic analysis of the resulting structures is difficult because they are usually extremely ambiguous. Web 2.0 tagging trades in semantic information for the ease of use.

Intelligent mediation can be investigated looking at recommender systems [1,2,3]. They are a cornerstone of the current Internet and can be found in a broad range of applications ranging from online stores as Amazon, over music organizers to web site recommenders. Recommendations may be based on knowledge about the domain or on similarity of users. The knowledge-based approach can be illustrated by the Pandora system of Tim Westergren's Music Genome Project. Pandora exploits careful expert annotations of music which are based on a music-theoretic ontology and matches them with user feedback for the recommendation of music, presented as on-line Internet radio. The collaborative approach can be illustrated by Amazon, where the shopping behavior of customers is observed and the overlapping items in the baskets of several users is used as a similarity measure of users. The associated items which are not yet overlapping are recommended.

Currently large amounts of multimedia content are stored at personal computers, MP3 devices, and other devices. The organization of these collections is cumbersome. First, the users have a different view of their collection depending on the occasions in which they want to use it. For instance, for a party you retrieve rather different music than for a candle-light dinner. Also places are related to views of the collection. In the car you might prefer rather different music than than at the working place. Moreover, a good host might play music for a special guest which he himself doesn't like usually. Hence, there is not one general preference structure per person, but several preference structures of the same person. Of course, some people share their views. However, it might well happen that one person's favorite songs for car-driving are best liked by another person while cleaning the house, where for car-driving this person wants different music. We generalize these observations to *aspects* of structuring collections. Second, flat structures like the one used by collaborative filtering are not sufficient. There, the correlation of items is taken into account, but no further structures. This leads to the presentation of a collection in terms of a table (like in iTunes) or just lists. Instead, we would like to organize a collection in terms of hierarchies so that browsing through the collection becomes easy, and the user gets a nice overview of the collection.

In this paper we present ongoing work on the project Nemoz , a collaborative music organizer based on distributed data and multimedia mining techniques. Distributed organization of music is an ideal test case for ubiquitous data mining. First, the management of multi-media data is a very hard task because corresponding semantic descriptions depend on highly social and personal factors. Second, media collections are usually inherently distributed. Third, multi-media

data is stored and managed on a large variety of different devices with very different capabilities concerning network connection and computational power. In this paper, we show that some of the problems connected with the automatic analysis of user-created tags can be solved by allowing to group tags into aspects. We investigate, how to tailor the representation such that it supports the personalized services provided to users as well as distributed data mining techniques.

The rest of this work is structured as follows: Section 2 introduces the Nemoz platform. Section 3 introduces the idea of multi-aspect tagging and discusses several possible restrictions, their utility and their implications. In section 4 we analyze how aspect-enriched tagging structures can be exploited by distributed data mining methods. In section 5 we give a conclusion.

2 Nemoz - Networked Media Organizer

Together with a group of students we have developed Nemoz [1] as a framework for studying collaborative music organization. Nemoz is made for experimenting with intelligent functionality of distributed media organization systems. Of course, the basic functions of media systems are implemented: download and import of songs, playing music, retrieving music from a collection based on given meta data, and creating play lists. The data model is an implementation of the knowledge representation formalism described in section 3. It has been enhanced to cover not only standard music meta data (performer, composer, album, year, duration of the song, genre, and comment) and a reference to the location of the song, but also features which are extracted from the raw sample data.

Communication among (W)LAN nodes via TCP and UDP is supported by a network service. Intelligent functions are based on the principles described in section 4. A collection can be organized using several aspects in parallel. These aspects can be interpreted as taxonomies. At each tag node, an extended function can be stored, which decides whether a new song belongs to this node, or not. This classifier is the learned model, where the learning task has been performed by one of the methods supplied by the machine learning environment YALE [4][2].

Based on this data structure Nemoz already implements several (intelligent) functions:

- New tags can be created.
- Existing tags can be assigned to arbitrary sets of items.
- Tags can be grouped into arbitrary aspects.
- Tags can be automatically assigned to new items.
- Users can search for similar aspects and tags in the network.
- Users can search for music similar to a selected song or group of songs.
- Tag structures can be automatically enhanced through the tags of other users.

[1] See also http://nemoz.sf.net.
[2] Available at http://yale.cs.uni-dortmund.de.

Applying Machine Learning methods to the field of personal music management offers many new opportunities. Typical applications include the classification of music items according to predefined schemes like genres [5,6], automatic clustering and visualization of music plays [7,8], recommendations of songs [9], as well as the automatic creation of playlists based on audio similarity and user feedback [10,11].

A key issue in all these approaches is the representation of the underlying music plays. A large variety of approaches has been proposed for extracting features from audio data [12,13]. However, it turns out that optimal audio features strongly depend on the task at hand [14] and the current subset of items [8]. One possibility to cope with this problem is to learn an adapted set of features for each learning task separately [15,16]. These approaches achieve a high accuracy, but are computationally very demanding and not well suited for real time processing. Furthermore, it is often important to capture cultural aspects not inferable from the audio data. Such aspects have a significant influence on how people structure and perceive their music [17]. Cultural aspects are usually incorporated by co-occurrences of songs in playlists [9] or by textual data [18,19].

Regarding local patterns is also crucial for the process of feature extraction, since a feature set which is valid for the overall collection is hard to find [14]. It is not very likely that a feature set delivering excellent performance on the separation of classical and popular music works well also for the separation of music structured according to occasions. This problem already arises for high-level structures like musical genres and is even aggregated due to the locality induced by personal structures.

If there would exist one complete set of audio features, from which each learning task selects its proper part, the feature problem could be reduced to feature selection. However, there is no tractable feature set to select from. The number of possible feature extractions is so large – virtually infinite – that it would be intractable to enumerate it. A unified framework for extraction methods has been developed which allows for automatic learning the optimal feature extractors for a given learning task [15]. The result is a (nested) sequence of data transformations which calculates the optimal feature set. Learning feature extraction delivers good results, but training the feature extraction is time-consuming and demands a sufficient set of examples.

In Nemoz , each user may create arbitrary, personal classification schemes to organize her music. For instance, some users structure their collection according to mood and situations, others according to genres, etc. Some of these structures may overlap, e.g., the blues genre may cover songs which are also covered by a personal concept "melancholic" of a structure describing moods. Nemoz supports the users in structuring their media objects while not forcing them to use the same set of concepts or annotations. If an ad hoc network has been established, peers support each other in structuring.

By recommending tags and structures to other users, we establish emerging views on the underlying space of objects. This approach naturally leads to a social filtering of such views. If someone creates a (partial) tag structure found

useful by many other users, it is often copied. If several tag structures equally fit a query, a well-distributed tag structure is recommended with higher probability. This pushes high quality tag structures and allows to filter random or non-sense ones. While the collaborative approach offers many opportunities, audio features can still be very helpful in several ways. The most important is that they allow to replace exact matches by similarity matches. This is essential when dealing with sparse data, i.e. when the number of objects in the tag structure is rather small.

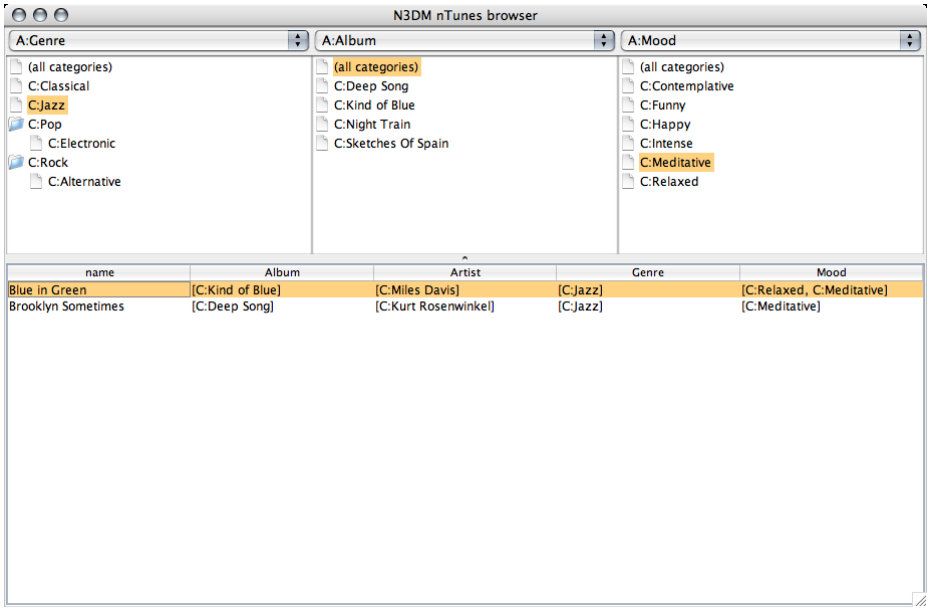


Fig. 1. The Nemoz filtering browser gives the user a filtered view of all the items in her media library, which can be successively refined by applying up to three aspects filters.

3 Multi-Aspect Tagging

One of the major challenges in enabling distributed, collaborative media organization is to find an appropriate representation mechanism. In the following we will derive requirements that such a mechanism must fulfill. We will show that neither current Semantic Web approaches nor popular Web 2.0 tagging approaches fulfill these requirements. This is the point of departure for our aspect-based tagging approach.

While developing the Nemoz system, we found a set of requirements that a representation mechanism must fulfill to be well-suited for distributed, collaborative media organization:

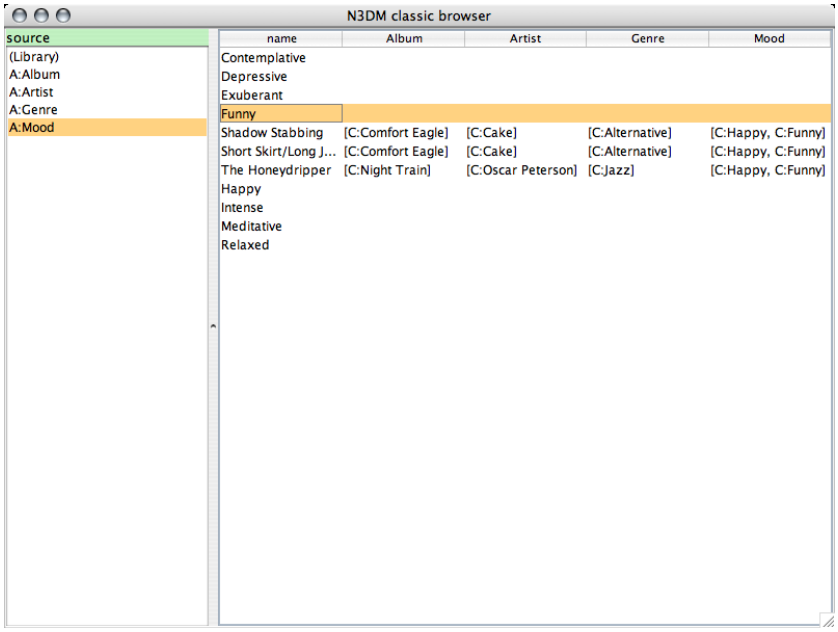


Fig. 2. Aspects in our formalism form trees. The Nemoz taxonomy browser shows these hierarchical structures in a concise manner.

1. No explicit coordination
   The growth of the Internet can be attributed largely to its loosely coupled character. If, for instance, every owner of a site would have to agree, that someone links to her site, the Internet would probably not have grown as fast as it did, nor would approaches as link analysis be as powerful as they are. We therefore require that a representation mechanism must not depend on explicit coordination among users.
2. Co-existence of different views
   Often, users do not agree on how to structure a certain set of items. It is therefore essential, that different representations of the same items may co-exist. In the extreme, each user should be allowed to create views completely independently of all other users. This allows for bottom-up innovation, as each user is capable of creating novel views. Which views become popular should emerge automatically, just like popular web-pages emerge automatically, as many other pages link to them.
3. Support for data mining and mediation
   While using loosely coupled representations is very attractive, the question remains how to derive useful information from such heterogeneous views and to allow users to profit from what other users did. A representation mechanism should therefore allow for the successful application of data mining and mediation methods.

4. Efficiency
   Relevant operations must be executable efficiently. For media management, the most important operations are the retrieval of items, basic consistency checks and the application of data mining methods, such as automatic classification.
5. Manageability
   The representation mechanism should be such, that it is easy for the user to overview and maintain the knowledge structures she created.
6. Ubiquitous environments
   The mechanism must be applicable in highly distributed environments. It must not expect, that all nodes are connected to the network all the time. Also, distributed data mining and retrieval methods must be applicable, such that the overall effort in communication time and cost is low, as media data is often organized on computationally poor devices connected by a loosely coupled network (such as p2p or ad hoc networks).

On the other hand, we think that other properties of knowledge representation mechanisms, especially as developed by the AI community, are not overly relevant for media organization. First, the representation of complex relationships is not of essential importance. Regular users are often not capable of dealing with such complex relationships (the large majority of Google users never even applied simple logical operators in their search requests). Also, complex relations are only seldom contained in the domain in question. Most properties can be simply expressed by pairs of attribute and value (artist, year of publication, ...). Furthermore, logical inference is often not really useful, as most users express their knowledge rather ad hoc and do not even accept logical entailment of what they expressed. We do not claim however, that these properties are irrelevant in general, we only claim that they are not relevant for media organization.

The most important representation mechanism for Internet resources is the Semantic Web. It is based on first order logic based representation mechanism. Given the above requirements, the Semantic Web is not well suited as representation mechanism for media organization. It is quite complex and requires explicit coordination among users. The co-existence of views and emerging views are not directly supported. Also, as the representation mechanism is quite powerful, operations may become inefficient (example OWL). It is based on logical entailment and is often not comprehensible for regular users. Finally, as it is usually based on explicit coordination, it is very hard to implement in a ubiquitous environment.

Recently, new applications emerged under the Web 2.0 paradigm. Systems as flickr or del.icio.us allow users to annotate items with arbitrary chosen tags. Such tags complement global properties, e.g. artist, album, genre, etc. for music collections used by traditional media organizers. In contrast to these global properties, many user-assigned tags are *local*, i.e. they represent the personal views of a certain user not aiming at a global structure or semantic. These systems allow for multiple and emerging views, do not require any coordination and are very easy to implement in an ubiquitous environment. A major drawback is,

that tag structures tend to be chaotic and hard to manage. Also they are not really well suited for data mining, which is a prerequisite for collaborative media organization.

In the following we show, that we can weaken these problems by introducing a knowledge representation formalism designed to support the concept of aspects.

Folksonomies emerging from popular social content services like last.fm or flickr constitute a large source of information. By virtue of compatibility, our formalism makes this information available for ontology-based knowledge discovery. Representing information from existing services consistently in one formalism enables us to create "mash-ups" of these services, i.e. to join data from multiple sources. This possibility is a defining trait of Web 2.0 applications. By integrating into the existing Web 2.0, new applications avoid the dilemma of a "cold start". BibSonomy [3] (see also [20]), a collaborative bookmark and publication sharing system, includes DBLP data in this fashion.

3.1 Basic entities and concepts

The basic entities in our formalism are *users*, *items* (songs), *categories*, and *aspects*:

Definition 1. *(Domain sets)*

$$\begin{array}{ll} U = \{u_1, \ldots, u_l\} & \textit{(User Identifiers)} \\ I = \{i_1, \ldots, i_n\} & \textit{(Item Identifiers)} \\ C = \{c_1, \ldots, c_m\} & \textit{(Category Identifiers)} \\ A = \{a_1, \ldots, a_k\} & \textit{(Aspect Identifiers)} \end{array}$$

Instead of storing these entities directly, we distinguish between abstract, opaque entity identifiers and entity representations. This distinction is motivated by the "Representational State Transfer" [21] paradigm of the World Wide Web, to which our formalism adheres to. In the rest of this work, we will only deal with abstract entity identifiers in the form of URNs.

In the following paragraphs, we describe the concepts of our formalism as a series of extensions to the Web 2.0 tagging model. In this model, users annotate a common set of items with tags. We represent tags by category identifiers. Links between items and tags are called $\mathcal{IC}$-Links:

Definition 2. *($\mathcal{IC}$-Link Relation)*

$$\rhd_{IC} \subseteq I \times C.$$

Our concept of a category extends the Web 2.0 tagging model by explicitly allowing "categories of categories", thereby enabling the representation of hierarchical structures akin to first order logic and description logics [22]:

[3] Online at http://www.bibsonomy.org.

Definition 3. *(CC-Link partial order)*
The CC-Link partial order is a relation

$$\preceq_{CC} :\subseteq C \times C$$

which satisfies the following axiom:

$$c \preceq_{CC} c' \Rightarrow ext(c) \subseteq ext(c') \text{ where } c, c' \in C, \tag{1}$$

where ext(c) is the item extension of a category c.

Note that in our formalism, the fact that $\text{ext}(c) \subseteq \text{ext}(c')$ does *not* imply that $c \preceq_{CC} c'$. We will motivate this design decision by an example: Consider a user whose music library contains very little jazz, all by Miles Davis. Our formalism would not force this user to accept the rather nonsensical identification of jazz and Miles Davis implied by the identity of the extension sets. If this identification actually reflects the user's opinion, she is still free to declare it explicitly.

Our formalism allows the user to organize categories further by grouping them into aspects:

Definition 4. *(CA-Link Relation)*

$$\blacktriangleright_{CA} :\subseteq C \times A$$

Typical examples for aspects from the music domain are "genre", "mood", "artist" and "tempo". The addition of aspects enables, among other things, the extraction of corresponding taxonomies, as described in section 4.

The usefulness of aspects has several facets. First, hierarchical category structures tend to become unmanageable when growing in size. Aspects enable the user to create complex structures to organize her items and simultaneously maintain clarity. Consider a user, who uses del.icio.us to organize her hyperlinks. With a great number of tags, retrieving one such link becomes more and more complicated. Grouping tags/categories into aspects eases this task considerably. Second, aspects can be used for filtering large category structures. Filtering means restricting the visible fraction of these structures to a specific topic. A limited variant of this notion is implemented in the iTunes media organizer, where the user can select a genre or an artist she wants to browse. Our framework enables the user to browse her items by arbitrary aspects. Third, aspects implicitly define a similarity measure on items that can be used to realize aspect-based clustering and visualization.

All links are considered as first class objects, facilitating the implementation of the formalism in a distributed environment.

3.2 Users and Ownership

In our formalism, entities are "ownerless", only links are owned by users:

Definition 5. *(Link-ownership relation)*

$$\triangleright_O :\subseteq (\triangleright_{IC} \cup \preceq_{CC} \cup \blacktriangleright_{CA}) \times (\mathcal{P}(U) \backslash \emptyset)$$

Each link must have at least one owner. It may have multiple owners, if it has been added independently by multiple users. This is why we wrote the power set of users, $\mathcal{P}$. Item ownership is not a first class concept in our formalism. Nonetheless, our prototypic implementation (Nemoz) provides a notion of item ownership: An item (i.e., a song) is said to be "owned" by a User, if this User possesses a representation of this item (i.e., an audio file of this song) stored on her local machine.

Our user concept comprises human users as well as intelligent agents. An agent acts on behalf of a human user, but has an identity of its own. For example, the "intelligent" operations (i.e. clustering and classification) of Nemoz (see section 2) have been modeled using such agents. Each time an intelligent operation is triggered by a user, an agent user is created that performs the operation and adds the resulting links to the knowledge base. Our design gives the user control over the effects of these operations by clearly distinguishing between automatically generated and manually entered knowledge. An automatically generated link may be promoted to a user-approved link be changing the link ownership from an agent to its client user. The effects of an intelligent operation may be canceled by deleting the responsible agent. By keeping automatically generated knowledge in an ephemeral state until it has been approved by the user, we hope to tame the sometimes frustrating effects of a poor performing intelligent operation.

3.3 Nemoz Knowledge Bases

With the preliminaries in place, we are now able to define our notion of an aspect-enriched tagging structure:

Definition 6. *(Nemoz Knowledge Base) A* Nemoz Knowledge Base KB_{Nemoz} *is defined as an 8-tuple:*

$$KB_{Nemoz} := (I, C, A, U, \rhd_{IC}, \preceq_{CC}, \blacktriangleright_{CA}, \rhd_O),$$

which satisfies the following axioms:

$$\forall c \in C. \exists i. (i, c) \in \rhd_{IC} \tag{2}$$

$$\forall a \in A. \exists c. (c, a) \in \blacktriangleright_{CA} . \tag{3}$$

These axioms ensure that all categories and aspects in a Nemoz Knowledge Base are not empty, a property we will refer to as *supportedness*. Supportedness implies that all categories and aspects have "extensional support", which is favorable from a machine learning perspective as well as from a user perspective.

Constraining the definition of a Nemoz Knowledge Base, we can describe tagging systems as well as some description logics-based formalisms.

An obvious restriction leads to *flat Nemoz Knowledge Bases*, that disallow hierarchically structured categories:

Definition 7. *(flat Nemoz Knowledge Base) A flat Nemoz Knowledge Base $KB_{Nemoz/flat}$ is defined as a Nemoz Knowledge Base without $\mathcal{CC}$-Links ($\preceq_{CC} = \emptyset$), described as a 7-tuple:*

$$KB_{Nemoz/flat} := (I, C, A, U, \triangleright_{IC}, \blacktriangleright_{CA}, \triangleright_{O}).$$

A flat Nemoz Knowledge Bases is an aspect-enriched tagging system. These systems offer the benefits of aspects without the complexity of hierarchical category structures.

A further restriction leads to simple tagging systems:

Definition 8. *(Tag Knowledge Base) A Tag Knowledge Base KB_{tag} is defined as a flat Nemoz Knowledge Base without aspect identifiers ($A = \emptyset$) which implies an empty $\mathcal{CA}$-Link Relation ($\blacktriangleright_{CA} = \emptyset$). Thus, a Tag Knowledge Base can be described as a 5-tuple:*

$$KB_{tag} := (I, C, U, \triangleright_{IC}, \triangleright_{O}).$$

A Tag Knowledge Base is a special case of a Nemoz Knowledge Base and may be seamlessly enriched by hierarchical categories or aspects. At the same time, each Nemoz Knowledge Base may be stripped down to a Tag Knowledge Base in a trivial manner. This flexibility enables simple inter-operation with existing knowledge bases of the Web 2.0.

A direct advantage of aspects is that the user is not confronted with a large number of tags, but with only some aspects that can be used to select subsets of tags. This essentially eases the visualization and maintenance of tag structures.

In the next section, we show that users profit from aspects yet in another way. The resulting structures are much better suited for data mining, which is the basis for the collaborative functionality of the system.

4 Aspect-based Multimedia Mining

Based on user taggings, several data mining techniques can be applied. The aim of the techniques is to support users in organizing and navigating their media collections. In the following we will focus on two tasks: first, the question of how audio files can be automatically annotated given a set of tagged examples and second the question of how to structure a set of audio files exploiting taggings of others. For both tasks, distributed data mining methods are provided.

4.1 Automatic Tagging

Often it is important to assign tags to audio files automatically. First, users can tag only a small amount of audio files manually, while the remainder of the audio files is tagged automatically. Second, tagging audio files automatically allows to browse and visualize music collections of other users with ones own terms.

To tag audio files automatically, we use the aspect representation presented in the last section together with hierarchical classification. As seen above, all

concepts belonging to an aspect must represent a tree. Thus for each aspect, there is exactly one most general concept. We tag audio files by training a classifier for each node in a concept hierarchy. Then, for each audio file these classifiers are applied in a top down manner until a leaf node is reached. This procedure is applied for each aspect and each audio file. Thus each audio file receives one most specific tag per aspect, which is connected to several super concepts according to the concept relation.

Aspects serve two purposes here. They firstly define the entry points for hierarchical classification and secondly group concepts explicitly into a set of trees. Without aspects, the algorithm would face a huge number of binary classification problems (one for each tag) for which furthermore negative examples are not explicitly given.

Furthermore, we use tags by other users as features to train classifiers. This approach is described in [23]. The idea is, that even though tags of different users may not be identical, they can still be highly utile for classification. For example background music is not identical to jazz, could however together with an additional audio feature, allow to perform classification with higher accuracy. Therefore, nodes can query other nodes for tags, which are then used as features.

4.2 Unsupervised Tagging

If a user has not yet assigned tags, these can be inferred by clustering. Traditional feature based clustering does not produce satisfying results, as it is very hard to infer labels for clusters, which is the basis for tagging.

In [24] we propose the LACE method which combines tags of other users to tag the own music collection. The idea is to cover the items to be tagged with a set of concept structures obtained from other users. This allows to cluster even heterogeneous sets of items, as different subset may be covered with different clusterings. Items that cannot be covered with clusterings obtained from other users are simply assigned tags using classification, as described above.

5 Discussion and Conclusion

In this paper we have introduced aspects as a means to group tagged items. This allows us to handle several hierarchies for one user. Whereas personalization approaches identify one user with one aspect, we take into account that the same user plays different roles depending on occasions. A result of this more finely grained structuring is that users have more opportunities to share items. A user may tag items under one aspect quite similar to the tags of another user (even under an aspect with a different name), where the overall taggings of the two users differ a lot. Hence, the introduction of aspects serves the collaboration in the distributed Web 2.0 setting.

The knowledge representation which we have formally described organizes tags into hierarchical structures. In particular, to each aspect, there is a hierarchy

of tags. This enables us to provide better services to users who organize their multimedia data. Retrieving, browsing, filtering becomes easy and accommodated to the user's personal aspects. Beyond enhanced human computer interfaces, the representation also allows more intelligent services. Automatic tagging using machine learning techniques for classification and unsupervised tagging using collaborative clustering reduces the burden of tagging.

The concepts and algorithms for aspect-based tagging are general, independent of the particular media which are to be structured. We have exemplified our approach by the Nemoz system which organizes music collections. Music collections are particularly hard to handle. For the user, a song must be listened to before she can tag it. In contrast, texts can more easily be skimmed through. For computation, music is given in a representation which must be converted into features. In contrast, texts already carry their primary ingredients of features, namely words. We have shown, how Nemoz deals with music on the one hand using simply identifiers, on the other hand using feature extraction. The hierarchical structuring works on the basis of identifiers, the collaborative clustering works on the basis of extracted features.

We believe that the automatic support by machine learning techniques as well as the distributed setting with its collaborative approaches open the floor for new ways of user collaboration and better services for users.

References

1. Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating "word of mouth". In: Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems. Volume 1. (1995) 210–217
2. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: Applying collaborative filtering to Usenet news. Communications of the ACM **40**(3) (1997) 77–87
3. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. Internet Computing, IEEE **7**(1) (2003) 76–80
4. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), ACM Press (2006)
5. Guoyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhytmic descriptors for musical genre classification. In: Proceedings of the International AES Conference. (2004)
6. Lidy, T., Rauber, A.: Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: Proceedings of the International Conference on Music Information Retrieval. (2005) 34–41
7. Schedl, M., Pampalk, E., Widmer, G.: Intelligent structuring and exploration of digital music collections. e und i – Elektrotechnik und Informationstechnik **7/8** (2005)
8. Moerchen, F., Ultsch, A., Thies, M., Loehken, I., Noecker, M.and Stamm, C., Efthymiou, N., Kuemmerer, M.: Musicminer: Visualizing perceptual distances of music as topograpical maps. Technical report, Dept. of Mathematics and Computer Science, University of Marburg, Germany (2004)

9. Stenzel, R., Kamps, T.: Improving content-based similarity measures by training a collaborative model. In: Proceedings of the International Conference on Music Information Retrieval. (2005)
10. Pampalk, E., Widmer, G., Chan, A.: A new approach to hierarchical clustering and structuring of data with self-organizing maps. Intelligent Data Analysis **8(2)** (2005)
11. Logan, B.: Content-based playlist generation: Exploratory experiments. In: Proceedings of the International Symposium on Music Information Retrieval. (2002)
12. Guo, G., Li, S.Z.: Content-Based Audio Classification and Retrieval by Support Vector Machines. IEEE Transaction on Neural Networks **14**(1) (2003) 209–215
13. Tzanetakis, G.: Manipulation, Analysis and Retrieval Systems for Audio Signals. PhD thesis, Computer Science Department, Princeton University (2002)
14. Pohle, T., Pampalk, E., Widmer, G.: Evaluation of frequently used audio features for classification of music into perceptual categories. In: Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05). (2005)
15. Mierswa, I., Morik, K.: Automatic feature extraction for classifying audio data. Machine Learning Journal **58** (2005) 127–149
16. Zils, A., Pachet, F.: Automatic extraction of music descriptors from acoustic signals using eds. In: Proceedings of the 116th Convention of the AES. (2004)
17. Baumann, S., Hummel, O.: Using cultural metadata for artist recommendations. In: Proceedings of the International Conference on WEB Delivering of Music. (2003)
18. Knees, P., Pampalk, E., Widmer, G.: Artist classification with web-based data. In: Proceedings of the International Conference on Music Information Retrieval. (2004)
19. Schedl, M., Knees, P., Widmer, G.: Discovering and visualizing prototypical artists by web-based co-occurrence analysis. In: Proceedings of the International Conference on Music Information Retrieval. (2005)
20. Haase, P., Ehrig, M., Hotho, A., Schnizler, B.: Personalized information access in a bibliographic peer-to-peer system. In Stuckenschmidt, H., Staab, S., eds.: Semantic Web and Peer-to-Peer. Springer (2005) 141–156
21. Fielding, R.T.: Architectural styles and the design of network-based software architectures. PhD thesis, University of California (2000)
22. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook. Cambridge University Press, Cambridge (UK) (2003)
23. Wurst, M., Morik, K.: Distributed feature extraction in a p2p setting - a case study. Future Generation Computer Systems, Special Issue on Data Mining (2006) (to appear).
24. Wurst, M., Morik, K., Mierswa, I.: Localized alternative cluster ensembles for collaborative structuring. In: Proceedings of the European Conference on Machine Learning. (2006)

Multilateral Security Requirements Analysis for Preserving Privacy in Ubiquitous Environments

Seda F. Gürses[1], Bettina Berendt[1], and Thomas Santen[2]

[1] Institute of Information Systems, Humboldt University Berlin, Berlin, Germany
[2] Department of Computer Science, Technical University Berlin, Berlin, Germany

Abstract. Privacy is of great concern in ubiquitous environments in which various technologies collect vast amounts of information about ubiquitous users with differing privacy and security interests. This concern also holds for knowledge-discovery systems in which data mining technologies infer substantial new knowledge from these data. Various methods have been proposed to preserve privacy in such environments, including privacy preserving data mining, mixes etc. However, it is not clear which of these methods provide the kind of privacy that users are looking for. We take a step back and look at what privacy means, and how the notion of privacy-related data has changed in ubiquitous environments. Further, we look at how we can state privacy requirements for ubiquitous environments through *security goals*, and we suggest a method for analysing multilateral security requirements which takes into account users' varying privacy interests. We believe that such a requirements analysis method will help determine whether and when privacy-preserving methods make sense for fulfilling privacy and security requirements of diverse sets of users in ubiquitous environments.

1 Introduction: Towards Privacy-Preserving Knowledge Discovery

It is a truism that a user-centric approach to Information Technology design requires that user wishes, needs, concerns, etc. be taken into account. It is less obvious that to find out about these user requirements, the designer needs to know how and what to ask. To elicit and analyse requirements, one also needs to know whom to ask. Who are the users who interact directly with the functional aspects of a system? Which differences between users need to be taken into account? Does a particular requirement also affect people outside the traditional group of users, i.e. further stakeholders? Is a specific requirement monolithic, or can there be dynamic or conflicting requirements in different contexts?

Among user-centric requirements, *privacy* has received a lot of attention in the past years, and it is by now accepted that privacy concerns need to be addressed while developing data intensive technologies. Privacy is a particularly pressing concern for knowledge-discovering systems: Data mining by definition requires (lots of) data as input and creates knowledge (or makes implicit knowledge explicit) about the data subjects. In addition, in real-life applications, knowledge from different IT systems can be and is combined. Privacy becomes an even more pressing concern in *ubiquitous* knowledge-discovery systems because of the huge increase in the amount and coverage of collected data, and thus in the possibilities of inferring further knowledge.

In IT systems, privacy can be protected by *security* mechanisms. In a *multilaterally secure* system, the security interests of all system stakeholders are considered, interest conflicts are identified, and methods for negotiating these conflicts are proposed.

But what is privacy? In many laws as well as in the data mining community, the concept is generally translated into data protection, or more specifically, the protection of personal data. Personal data is "any information relating to an identified or identifiable natural person [...]; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" (EU Directive 95/46/EC [6], Art. 2 (a)). This definition has two effects: first, it focusses the attention on data (as opposed to people), and second, it focusses the attention on the identification of a natural person as the problem. Thus, it implicitly declares data-processing activities to be "privacy-preserving" if and only if they do not (better, if they provably cannot) involve the identification of the natural person who is the "carrier" of a record of attributes. This notion is reflected in the data mining literature by regarding goals such as k-anonymity as the defining properties of a privacy-preserving algorithm.

However, the privacy literature suggests that a people view of privacy involves not one but many identities, that there are concerns over profiles independently of identification, and that context is all-important. A true knowledge discovery perspective takes this into account when it transcends the algorithm-centric view described above: It must involve business / application understanding, and it must involve the evaluation of deployment and the lessons learned from previous cycles of the KD process. This means that the complexities of software development and software use must be considered: different users interact with an IT system in different contexts, and each particular system exists in an environment populated by many systems.

Further, different parties may have diverging or even conflicting understandings of privacy. It may be difficult to use privacy-enhancing technologies such that these conflicting interests are articulated and resolved. Developing and implementing privacy enhancing technologies without an understanding of the larger context (= the way a system is used, the way a system is surrounded by other systems, etc.) may also fall short of attending to the privacy concerns of different users in ubiquitous environments. Such shortcomings may result in privacy breaches that may cause a loss of trust in such systems and affect technology acceptance.

The purpose of this paper is to stimulate discussion about what it means to do privacy-preserving knowledge discovery in general, and privacy-preserving knowledge discovery in ubiquitous-computing environments in particular. In order to do this, we describe different views of privacy (Section 2), emphasise the specifics of ubiquitous-computing environments by focusing on location-aware systems (Section 3), argue that privacy goals must be transformed into security goals in order to be integrated into system design (Section 4), and describe a method for doing the requirements analysis that must form the beginning of such system design (Section 5). The scope of the analysis extends beyond knowledge-discovery systems; as we have argued above, it is particularly relevant in knowledge-discovery systems because they are data intensive and generate new knowledge about people

The paper argues at the high level of privacy goals, security goals, and requirements analysis in order to ask and find out *what* is wanted, in terms of privacy preservation, in ubiquitous knowledge discovery for users. Details of *how* privacy is endangered by data mining, and how privacy preservation can be implemented into its algorithms, are not the topic of this paper (although an integration is a goal of our future research). We refer the reader to [4] for a tutorial introduction to the issues of privacy, security, and data mining, to [20] for the classic description of the inference or triangulation problem, and to [24] for an overview of privacy-preserving approaches in data mining algorithms.

2 Definitions of Privacy

Privacy is a contested term. Arguments range over the exact meaning of the term, over its placement into more fundamental categories with implications on the extent to which personal data can be traded, and over which principles governing data collection and handling can best ensure it.

Phillips [18] takes a sociological approach and distinguishes four **kinds of privacy**:

1. **Freedom from intrusion** : This is most popularly known as the "right to be let alone" [25]. This kind of privacy presumes an intrinsic self and protects a sphere of autonomy in which the individual is free to express and inhabit that self and enjoy intimate relationships. The home as an intimate and untouchable space is an example of this kind of privacy.
2. **Construction of the public/private divide** : This distinction concerns the social negotiation of what remains private (i.e. silent and out of the public discourse) and what becomes public. For example, in the case of voting, individuals may want to keep their choice private, and in the case of domestic violence, interest groups may have an interest in defining the "domestic" as a public issue.
3. **Separation of identities** : This allows individuals to selectively employ revelation and concealment to facilitate their social performances and relationships. It gives individuals the right to control, edit, manage and delete information about themselves. An example is the possibility to maintain multiple identities in Internet chat rooms, games, etc.
4. **Protection from surveillance** : Surveillance refers to the creation and managing of social knowledge about population groups. This kind of privacy can easily be violated if individual observations are collated and used for statistical classification, which applied to individuals makes statements about their (non)compliance with norms, their belonging to groups with given properties and values, etc. Market segmentation is an example of the classification of population groups which may constitute a breach of this kind of privacy. Conversely, the prohibition to combine certain data (as in Germany, where the combination of certain online and offline data is disallowed) can protect this kind of privacy to a certain extent.

These four kinds of privacy are neither exhaustive nor mutually exclusive, but they help in understanding the kinds of privacy that may be protected or threatened by emerging technologies. Many of the existing concepts of privacy-preserving technologies can also be better understood in the light of these definitions. For example, anonymizers

protect all four kinds of privacy, while identity management systems clearly attend to the "separation of identities".

In addition to these different foci on identity (3), profiles (4), or contexts (1 and 2),[3] privacy can be regarded as a basic and inalienable human right, or as a personal right or possession. When regarded as a basic and inalienable human right, privacy has to be regulated through laws, political decisions, and ethics. when regarded as a personal right or possession over which the individual has free discretion, it becomes a matter of utility, taste and freedom of choice.

Regardless of the underlying notion of individual and society, there are cases in which one person guarding or giving up her own privacy may breach or strengthen the privacy of others. This means that regulations may put restrictions on individual wishes to give up privacy, or at least forbid that consequences are taken that might make sense from a purely economic point of view.

For example, if a majority of health-insurance clients disclose their individual risk profiles to their insurance companies, it will invariably be concluded that the others have something to hide, constitute bad risks, and should therefore be charged higher premiums. This may make chronically ill patients effectively uninsurable and create deep rifts through society. As a political consequence of these considerations, German public health insurance companies are not allowed to act on risk information received from clients(although the clients are free to disclose this information).

Depending on how basic the right is considered to be, the need may also arise for the state to "protect people against their own wishes". For example, the German law treats human dignity as an absolute human right in the sense that the state *must* protect people against being treated like an object (even if they consent to it). Recent examples include consensual cannibalism (treated in court as a murder) and the sport dwarf-tossing. The case is less clear for privacy, as the uproar (that soon sank into oblivion) over TV shows such as "Big Brother" in various European countries has shown.

Given the complexity and diversity of current technological developments, privacy probably needs to be seen as both. The preservation of privacy as a social good requires expert understanding and top-down regulations. At the same time, given that global approaches can not respond to the specifics of many technological applications and situations, privacy as a personal right should be guaranteed such that it allows individuals to negotiate and exert control over how much information they reveal about themselves.

Further, notions of privacy have often been challenged by technological developments. Technologies applied in ubiquitous environments are expected to collect more data about individuals, at finer granularity, closer to the physical reality, which can be analysed using sophisticated tools. As an example, technology for continuously locating a person and its privacy implications may have been a schizophrenic nightmare in the 1950s (see the account of John Nash's life in "A Beautiful Mind"), but RFID-chip implants have made them a reality today, at least for animals.

In any given setting of legal conceptualisation and technological possibilities, the question is how data handling can be regulated such that the desired forms of privacy are safeguarded. The first two of Phillipp's kinds of privacy imply that data, or certain data,

[3] See [1] for empirical evidence that people differ with regard to whether they consider identity reconstruction or profiling as major privacy threats.

should not be generated, or collected, at all if it can be avoided ("data parsimony"). The fourth also calls for data parsimony and in addition implies strong restrictions on the unforeseen re-use of data for new purposes. The third has, in principle, no implications on the amount of data, but calls for effective measures of control and forbids re-uses that might compromise the separation of identities.

Guidelines for data protection (or for data protection laws) such as the OECD's Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [15], the Fair Information Practices (FIP) *notice, choice, access, and security* that were set down as a recommendation in the US [23] and updated by the Federal Trade Commission [8], or the principles of the EU Privacy Directives [6, 7], are in line with this in the sense that they define privacy not only as a matter of concealment of personal information, but also as the ability to control what happens with it.

In the following, we will work with the principles set up in the **OECD Guidelines**. These are sometimes also referred to as (the Eight Principles of) **"Fair Information Practices"** (e.g., [9]). To avoid confusion, we will refer to them as the *FIP 8*:

Collection limitation : Data collectors should only collect information that is necessary, and should do so by lawful and fair means, i.e., with the knowledge or consent of the data subject.

Data quality : The collected data should be kept up-to-date and stored only as long as it is relevant.

Purpose specification : The purpose for which data is collected should be specified (and announced) ahead of the data collection.

Use limitation : Personal data should only be used for the stated purpose, except with the data subject's consent or as required by law.

Security safeguards : Reasonable security safeguards should protect collected data from unauthorised access, use, modification, or disclosure.

Openness : It should be possible for data subjects to learn about the data controller's identity, and how to get in touch with him.

Individual participation : A data subject should be able to obtain from a data controller confirmation of whether or not the controller has data relating to him, to obtain such data, to challenge data relating to him and, if the challenge is successful, to have the data erased, rectified, completed or amended.

Accountability : Data controllers should be accountable for complying with these principles.

Although these principles are desirable, relying only on them when building systems can be misleading. Collection limitation in one system does not protect against the aggregation of that data in many systems. Openness may be overwhelming in ubiquitous environments, where the number of data controllers may be inflationary. A user may be overwhelmed by the difficulties of individual participation. Further, even if all these principles were implemented, it would be very difficult to identify violations. Even in the case of trusted parties, system security violations (i.e. hacked systems) or design failures (i.e. information leakages) or linking of different sources of safely released data may cause unwanted release of information.

Therefore, an ideal system in terms of different privacy definitions from Phillips and the "collection limitation" in the FIP 8, is a system that collects the minimum

amount of personal data. If possible, data collection should be avoided, since inference on non-personal data may still reveal a lot of information about individuals or groups of individuals. The FIP 8 need to be observed. This can be done by incorporating these principles already during the requirements analysis phase of systems.

Example (part I): To illustrate the concepts used in this paper, we use a fictitious running example from an application domain that is currently popular in the literature on ubiquitous environments: a presence support system that gives passengers directions at an airport. Before we start the example, we first have to introduce the special implications of location awareness to privacy.

3 Location Awareness and New Threats to Privacy

Ubiquitous computing (aka pervasive computing) is about smart environments enriched with sensors, actuators and other invisible embedded devices. These environments interact with individuals, with small devices that may or may not be carried by the user (anywhere from hand-held computers to passive RFID tags). Context awareness is also an important factor in ubiquitous computing. Context awareness is possible through the collection of environmental and user data which are then analysed with respect to time (i.e. history) and other environmental characteristics. These technologies collect immense amounts of data, which are often personal data coupled with location data.

Location is most easily defined as the triple (id, position, time). Location is considered a central attribute of context awareness and thus desired functionality in ubiquitous environments. At the same time, location data are particularly sensitive information because of the specific characteristics of space: One cannot not be in a location at any time (one cannot "opt-out of being somewhere"), so the impression of lacking self-control and comprehensiveness of surveillance is particularly pronounced. In addition, spatial and spatio-temporal data allow many inferences because of rich social background knowledge. One reason is that because of physical constraints, the number of interpretations of being at one location is often bounded (e.g., visits to a doctor specialising in the treatment of AIDS patients). Another reason is that there are typical spatio-temporal behaviour patterns (a location where a person habitually spends the night without moving is most likely that person's home).

The preservation of location privacy in ubiquitous environments is difficult, if not impossible. Multiple channels collect different kinds of physical (e.g. radio frequency fingerprints) and digital information, which can be used to precisely locate a given user. Trajectories made up of a temporally ordered series of location-triple observations with the same id, coupled with a-priori knowledge on places and social contexts, can be analysed to infer individual identities (violating freedom from intrusion), used to link profiles of individuals (violating separation of identities), or classify persons into previously defined groups (surveillance).

There are two phases in which privacy can be accounted for in location-aware systems. The first phase is during the collection of data in a privacy-preserving manner. We already suggested that although more desirable, this is very difficult to guarantee because as a rule, multiple channels exist. Several solutions have been suggested for preserving location privacy in location-aware systems (via the security goals anonymity and unlinkability, see Section 4). They include temporal and spatial cloaking which

guarantee location k-anonymity [11] and location mixes (the location of a user before entering and after leaving a defined zone – the mix zone – is not recognisable as belonging to the same person) [2]. The second phase in which privacy can be accounted for is in the dissemination of data in a privacy-preserving manner by trusted parties. Users of location-aware systems may want to make use of personalised services, for which they may be willing to give away some of their identity information. Dissemination solutions make use of location privacy-preserving data mining methods as well as legal frameworks. Historical location k-anonymity [3] and the Platform for Privacy Preferences Protocol for location data [5] are examples of such technologies. Further, the trusted party that disseminates the data is required to carefully analyse existing data mining techniques that can be used to breach privacy by linking data. These analyses may not be 100% complete because there may always be a new technique that breaches privacy through partial release of data. Nevertheless, understanding the strength of different location privacy-preserving methods as well as the strength of methods to breach location privacy through linking data are interesting research questions.[4]

Given privacy's complexity and vulnerability, considering it in future technologies requires a systematic approach starting early in system development. Further, developments in location awareness and data mining require thinking about the interaction of these technologies with other existing systems and the information available through them. Therefore, we suggest that these concerns be addressed already during the requirements analysis phase of systems development. In the following sections, we introduce our method for doing this.

Example (part II): Consider a traveller (T) who wants to get to her gate at the airport quickly. She wants to receive directions on her PDA that depend on her current location. However, she does not want to be tracked, and she does not want to receive further messages unless she gets lost. On her way, she visits the Duty Free shop.

Her privacy goals are based on her perception that her drug habits are private, such that her behaviour and purchases at the Duty Free shop should not be linked to her identity (she wants this public/private divide to be protected). [We assume a setting in which she can show her boarding card but hide her name.] She also wants to be let alone unless at those times where she gets lost. She wants to be absolutely left alone in the bathroom (freedom from intrusion, public/private divide). Also, she wants to maintain her identity as a traveller at the current airport separate from her identity as a traveller elsewhere. I.e., she does not want any knowledge transfer from the direction-giving system at this airport to other systems (separation of identities). Finally, she objects to being observed and being profiled as a traveller (protection from surveillance).

[4] Hardware can also be designed to limit data collection or dissemination. Examples include out-of-focus / blurred cameras to preclude face recognition, and selective decisions concerning whether a device can send, receive, or do both. (User-owned devices may guard privacy by only receiving but not sending information, whereas devices like RFID readers may guard privacy by sending "I exist and I am reading" messages but endanger it by only receiving.) An important question is how and to what extent devices can actually be turned off. For example, mobile phones emit signals even when switched off, as long as their battery is operating.

4 Articulating Privacy in Terms of Security Goals

Privacy in ubiquitous environments can be protected using security mechanisms. This is not to say that privacy is only a technical issue, but rather that its protection in technical artefacts depends on security mechanisms as well as legal frameworks and regulations. Mechanisms such as access control and authentication can safeguard against direct disclosures and may be useful in implementing some privacy by controlling "who sees which data". Nevertheless, these mechanisms do not address disclosures based on inferences made from partially released data using data mining techniques [21, 16].

Therefore, it is helpful to execute an analysis of privacy in ubiquitous environments at a higher abstraction level using what are called security goals. Security goals can be used to identify the security and privacy properties that an environment in which a computer system is embedded has to fulfil at all times [19].

Pfitzmann and Wolf [26, 27] give a detailed classification of security goals, which can be used to write down security requirements. These security goals are based on the well-known classification of confidentiality, integrity and availability that was developed for use in Internet based communication systems. Confidentiality-related goals comprise confidentiality, anonymity and unobservability. Integrity-related goals comprise integrity and accountability. Availability-related goals comprise availability and reachability. The **security goals** are:

Confidentiality : No one other than the communication partners can recognise the content of the communication.

Anonymity : A user can use services or resources without being recognised by anyone. Even the communication partner does not know the real identity of the user.

Pseudonymity : A user does not identify herself but is still accountable for her actions.

Unobservability : No one other than the communication partners are aware that communication is taking place.

Integrity : Any modifications made to the content of communication can be recognised by the communication partners.

Accountability : Communication partners can prove the existence of a communication to a third party.

Availability : Communicated messages or the resources are available when the user wants to use them.

Reachability : An entity (i.e., a user, machine etc.) either can or cannot be contacted depending on user interests.

Ubiquitous computing demands a rethinking of these requirements. The security goals should not be limited to communication properties, but should be applicable to all interactions in intelligent environments in which the users are active. In their analysis of identity-management terminology, Hansen and Pfitzmann [17] scrutinise the definitions of the security goals anonymity and unobservability, and then follow with a definition of unlinkability for a communication network.

We adapt this terminology to trajectories, so that privacy requirements that individuals may have with respect to collection or dissemination of trajectory data can be

articulated through the same three security goals. In this adoption, we refer to the common concept of "attackers" that is used to identify a security mechanism's power. These definitions help in identifying the **security goals for location privacy**:

Anonymity is the state of being not identifiable within a set of individuals, also called the anonymity set. For the sake of privacy with respect to trajectories, it is not enough to anonymise an individual's identity information, but it is also necessary to "anonymise" her trajectory data. This means that a link between a trajectory and the identity of that person can not be made.

Unobservability of a trajectory means that it cannot be recognised whether a trajectory exists or not. Unobservability implies the anonymity of trajectory owners.

Also, depending on these definitions and an attacker that has been characterised:

Unlinkability of two or more trajectories means that within the system from the attacker's perspective, these trajectories are no more and no less related after his observation than they are related concerning his a-priori knowledge.

We doubt that unobservability and anonymity, and hence unlinkability of trajectory data can be guaranteed if trajectory data is going to be disseminated. We assume this will be the case despite the use of location privacy-preserving data mining methods (see the discussion in Section 3). If this is the case, then **user requirements for identity management** in line with FIP 8 need to be considered in building systems. Terveen et.al. [22] list these requirements as follows:

Personalised disclosure : Users are able to reveal their personal information at different levels of resolution to others users at different times.

Transparency : The user is able to understand what other users will be able to see in specified situations as a result of their selected privacy policies.

Ambiguity : The system preserves doubt about whether a user visited certain locations.

Example (part III): Traveller T's privacy goals translate into the following security goals: She would like to be anonymous in the Duty Free shop with respect to purchase, and she requests unobservability of her behaviour during shopping. (She cannot request unobservability of purchasing, because retail laws require that all purchases are registered.)

She requests the unlinkability of the records of her behaviour (trajectories) in the current airport with the records of her behaviour in other places. Since anonymity is not really possible in an airport context, and since it would preclude that T can be contacted when she gets lost, she rather wants to use a relation pseudonym, where the relation with the airport is valid only as long as her stay at the airport.

She requests the unlinkability of the records of her behaviour with records of other people's behaviour (profile construction). Her wish to be let alone unless she gets lost does not translate into a security goal, but into the way the system is designed to output messages. Her wish to be let alone in the bathroom translates into the goal to be unreachable in bathroom locations.

5 A method for Multilateral Security Requirements Analysis

We now describe our method for requirements analysis which incorporates the security goals for the different users and stakeholders in a system-to-be. Further details of the method, related approaches, together with examples, are given in [13, 14, 12].

We build on the definition of requirements by Jackson and Zave [28] since they specifically emphasise domain analysis and provide a precise vocabulary to analyse environments, instead of reducing requirements to an understanding of the actors' interactions with the machine. The authors distinguish between the *environment* and the *system*. The system denotes the machine that we want to develop. The purpose of the system is to enhance the behaviour of the environment. *Requirements* are statements about how we want the environment to behave once the system is embedded in it. A *specification* describes the system. Further, *domain knowledge* of the environment is necessary to transform requirements into specifications. In this domain knowledge, we may have *facts*. These are always true in the environment under all circumstances. Further, we may have to make *assumptions* about the environment: these are properties that cannot be guaranteed but are necessary to make the requirements satisfiable.

The *stakeholders* of the system are all persons involved in the conception, production, use and maintenance of the system. We assume that stakeholders have security and privacy interests (*security preferences*) in the system. *Users* are the stakeholders who will potentially interact with the system. This distinction is helpful in recognising situations in which data relevant for stakeholders are contained in a system, where these stakeholders do not interact with the functionalities of the system, therefore are not users of the system.

In an ideal multilaterally secure system the security preferences of all stakeholders are considered and are negotiable. The objective is to provide each individual the chance to decide on his security preferences and on the mechanisms to implement these preferences within the system. In case of a conflict in security preferences, the system should provide mechanisms to negotiate differences, if negotiation is possible.

The security preferences of the stakeholders of a multilaterally secure system are expressed in terms of security goals. With our method, we propose a way to elicit these security goals and later make suggestions on how these can be composed to finalise the security requirements of a multilaterally secure and privacy-preserving system. The security preferences of the stakeholders of a system are not limited to security goals against malicious attackers but also towards all other stakeholders of the system. We assume that stakeholders may have common as well as conflicting security preferences towards the system and towards other stakeholders, depending on their social context.

In order to distinguish the diverging functional and security interests of the different stakeholders we work with the concepts of roles and subgroups. Each stakeholder has relationships with the system and with other stakeholders of the system. These relationships (*roles*) determine the security preferences of the stakeholders in the system.

We distinguish between social roles, functional roles and security roles. A *social role* describes the set of behaviours and preferences a person may have as a consequence of belonging to a social group, also known as the stakeholders. A *functional role* describes the set of resources a person may use and the actions a person may take within a system. A *security role* defines the security goals a person has or the security goals attributed to a stakeholder for a specific functionality of the system. A role may be attributed to many stakeholders and each stakeholder may have many roles. A set of stakeholders with common roles and goals are defined as a *subgroup* of the system.

The method consists of seven steps which are illustrated in Figure 1. Since we focus on multilateral security, we are interested in distinguishing between the heterogeneous privacy, security and functional requirements of the various stakeholders. Further, we are interested in distinguishing the different functionalities of the system for which the security requirements vary.

The first step of the requirements analysis method identifies the stakeholders and hence the social roles of the system. Here we also identify if there are any constraints as to the social roles an individual may have simultaneously or sequentially. Further, we perform a typical functional analysis of the system. During this step, we identify the resources and functionalities of the system-to-be. The functional analysis is also used to identify the functional roles of the system. Again, any constraints on multiple-roles attributed to a single user are identified.

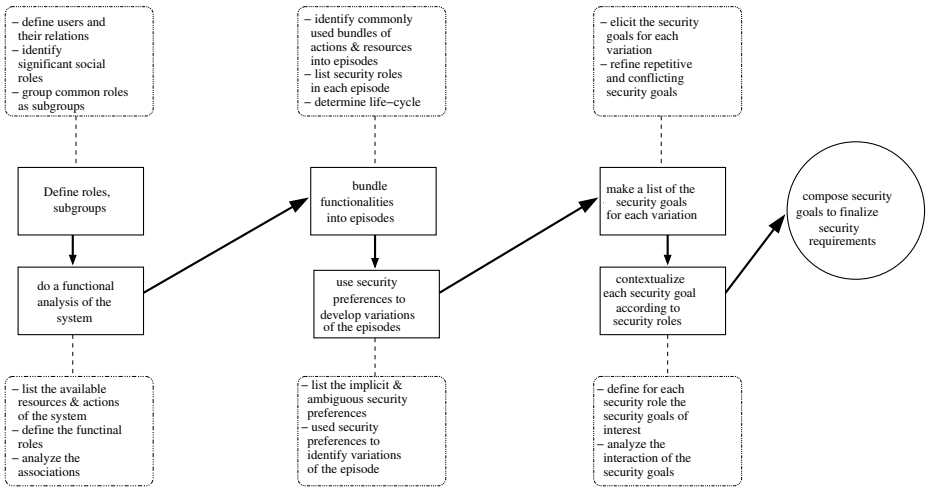


Fig. 1. Seven steps from security goals to security requirements.

The method distinguishes between the preferences that users have towards different parts of the system. To realize this we make use of abstractions of system functionalities called *episodes*. Episodes consist of actions and resources within a time frame for which common security preferences for a set of stakeholders can be defined. Episodes are different from "use cases" in that they differentiate between the same "use" in different contexts, and in that they take into account the roles of stakeholders who may or may not be actors of the use cases in the episodes.

Hence, in the third step of the method, episodes of the system are identified. Since episodes are determined by common security preferences, here we also identify security roles. Security roles are especially important for being able to talk about groups of individuals towards which stakeholders have security preferences. Security roles may also be attributed to those who are not part of the system (i.e. external malicious attackers). These are called the *counter stakeholders* of a security goal in an episode.

Given that stakeholders of a system may have different security preferences, in the fourth step of the method we consider these differences to identify what variations of functionalities a system should offer. For example, there may be some users who want to make use of an episode anonymously vs. a group of users who are willing to be identified in return for personalised services in that episode. In this case, two variations of that episode need to be considered, for which different privacy and security requirements are articulated by different groups of users.

In the fifth and sixth steps we analyse the security goals of the stakeholders towards episodes. We ***contextualise*** **each security goal**. That means that for each security goal, the method identifies whose security goal it is (stakeholder), against whom (counter-stakeholder), and for which functionality or resource of the system (episode). Last but not least, suggestions are made on how to compose the different security goals into valid and consistent security requirements.

Example (part IV): Traveller T wants to use different services that may be linked with different use cases. For example, she wants to ask the system for directions, or she wants to be alerted to a necessary change in her current path. However, the use case "receiving an alert" is contextualised into different episodes. These are: a) in a hallway – here, she acts in an open and "public" space in which many similar travellers are in her vicinity, but usually not in close physical proximity, and she is happy if details of her flight and the gate number appear on the display; b) in the Duty Free shop – here, she acts in close physical proximity to the clerk, and she does not want her flight data and gate number to be visible on the display of her PDA; c) in the bathroom. Here, she does not want to receive any messages even if this unreachability goal is in conflict with her interest in getting critical messages in a timely manner.

T's goals are in partial conflict with the stakeholder "airport administration". The administration wants to measure the usage of the hallways and bathrooms in order to plan capacities for the upcoming renovation of the building. In the case of hallways, the conflict can be resolved as follows: usage is only recorded as the total number of people passing a certain location in a certain time interval (say, 1 hour); trajectory connections between measurement points are not recorded, and no further information about the travellers is recorded. This compromise however requires transparency as a user requirement for identity management.

In the case of bathrooms, the conflict is difficult to resolve because people like T object to being counted in the same way as in the hallways. Measuring the number of people passing the hallway entry point to the bathroom might be a compromise, but it may also be regarded as just a proxy. In the case of Duty Free shops, the number of purchase receipts per time interval may be used as a proxy for the number of shoppers. This is an imperfect proxy because one shopper may make several purchases, and other shoppers may make no purchases. However, it is a privacy-preserving proxy in the sense that the data need to recorded (unobservability of purchases is precluded by law) and that they are anonymous [we assume cash-only purchases or adequate anonymization of credit-card purchases].

6 Conclusions and Future Research

We have argued that integrating privacy and security requirements into the functional requirements analysis of ubiquitous environments is necessary. We first gave definitions of the privacy that we would like to see protected in such environments. Next, we talked about the kinds of privacy threats possible through these new technologies. And last, we

described a method that deals with the complexities of articulating privacy and security requirements for ubiquitous systems.

We are currently working on an RFID-based case study and on analysing multilateral security and privacy requirements for a distributed location-aware system. One of our goals with this case study is to better understand privacy and security requirements with respect to ubiquitous technologies. Further, we are developing a notation for modelling these requirements such that conflicts and inconsistencies, as well as missing requirements can be better identified.

One of our future goals is to compare our method and notation to existing methods of security requirements elicitation. Specifically, we are interested in comparing our method to the tropos security requirements framework [10]. We plan to validate our method through case studies in which we show that without the method it is not possible to elicit the conflicting privacy and security requirements of the stakeholders systematically. Another topic of future research is a method for composing security goals into security requirements in which conflicts are resolved.

User/stakeholder tools for articulating and implementing privacy interests must not only exist, but also be usable. The realization of identity-management goals for location-aware systems as identified by Terveen et al. [22] will only make sense if the users understand what these identity-management possibilities are and can utilise them. Thus, along with more standard usability requirements, the following **usability requirements** need to be considered in the development of ubiquitous technologies: (a) the user is aware of the different security and privacy possibilities in the system, (b) the user is aware or understands what can happen with her data, (c) the user is able to put security and privacy mechanisms to use – without investing an unacceptable amount of effort to do so. In future work, we will integrate such usability requirements into our multilateral security requirements analysis method.

References

1. B. Berendt, O. Günther, and S. Spiekermann. Privacy in e-commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, 48(4):101–106, 2005.
2. A. Beresford and F. Stajano. Mix zones: User privacy in location-aware services. In *In Proc. IEEE Worksh. on Pervasive Computing and Communication Security*, pp. 127–131, 2004.
3. C. Bettini, X. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. In *Proc. of the 2nd Workshop on Secure Data Management at VLDB 2005*, pp. 185–199. Springer Verlag, Berlin, LNCS 3674, 2005.
4. C. Clifton. *Privacy, security and data mining.* Tutorial at ECML/PKDD'02, Helsinki, Finland, August 2002. `ecmlpkdd.cs.helsinki.fi/pdf/clifton_psdm.pdf`.
5. A. Escudero, T. Holleboom, and S. Fischer-Huebner. *Privacy for location data in mobile networks*, 2002.
6. EU. *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.* `http://europa.eu.int/eur-lex/en/consleg/main/1995/en_1995L0046_index.html`.
7. EU. *Directive 2002/58/ec of the european parliament and of the council concerning the processing of personal data and the protection of privacy in the electronic communications sector*, 2002. `http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf`.

8. Federal Trade Commission (FTC). *Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress*, May 2000. http://www.ftc.gov/reports/privacy2000/privacy2000.pdf.
9. C. Floerkemeier, R. Schneider, and M. Langheinrich. Scanning with a purpose - supporting the fair information principles in rfid protocols. In *Proc. Second International Symposium on Ubiquitous Computing Systems UCS, Tokyo, Japan*, 2004.
10. P. Giorgini, F. Massacci, and J. Mylopoulos. Requirement engineering meets security: A case study on modelling secure electronic transactions by visa and mastercard. In *Proc. ER'03, Chicago, Illinois, October 2003*, 2003.
11. M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proc. of MobiSys 2003*, 2003.
12. S. Gürses. *Security requirements elicitation in multi-laterally secure systems.* Master's thesis, Humboldt University, Berlin, 2004.
13. S. Gürses, J. H. Jahnke, C. Obry, A. Onabajo, T. Santen, and M. Price. Eliciting confidentiality requirements in practice. In *In Proc. 15th Annual International Conference hosted by the IBM Centers for Advanced Studies (CASCON 2005). IBM, Canada*, 2005.
14. S. Gürses and T. Santen. Contextualizing security goals: A method for multilateral security requirements. In *Proc. Sicherheit 2006 - Schutz und Zuverlässigkeit*, 2006.
15. OECD. *Guidelines on the protection of privacy and transborder flows of personal data*. http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html.
16. T. Owad. *Data mining 101: Funding subversives with amazon wishlists*, 2006. http://www.applefritter.com/bannedbooks.
17. A. Pfitzmann and M. Hansen. *Anonymity, unlinkability, unobservability, pseudonymity, and identity management - a consolidated proposal for terminology*, 2006. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.
18. D.J. Phillips. Privacy policy and PETs: The influence of policy regimes on the development and social implications of priv. enhanc. technol. *New Media Society*, 6(6):691–706, 2004.
19. H. Schmidt. Security engineering using problem frames. In *LNCS Proceedings of the International Conference on Emerging Trends in Communination Security*, 2006.
20. L. Sweeney. *Computational Disclosure Control: A Primer on Data Privacy Protection*. Ph.D. Thesis, MIT, Cambridge, MA, 2001. pdf.
21. L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
22. L. Terveen, R. Akolka, P. Ludford, C. Zhou, J. Murphy, J. Konstan, and J. Riedl. Location-aware community applications: Privacy issues and user interfaces. In *In Proceedings of the Location Privacy Workshop*, 2004.
23. U.S. Department of Health, Education and Welfare (HEW). *Secretary's advisory committee on automated personal data systems, records, computers, and the rights of citizens viii*, 1973. See http://www.epic.org/privacy/consumer/code_fair_info.html.
24. V. S. Verykios, E. Bertino, L. Parasiliti, I. N. Favino, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *Sigmod Record*, 33(1):685–699, 2004.
25. S. Warren and L. Brandeis. The right to privacy. *Harward Law Review*, 4:193–220, 1890.
26. G. Wolf and A. Pfitzmann. Empowering users to set their security goals. In *Multilateral Security for Global Communication Technology*, 1999.
27. G. Wolf and A. Pfitzmann. Properties of protection goals and their integration into a user interface. *Computer Networks*, 32:685–699, 2000.
28. P. Zave and M. Jackson. Four dark corners of requirements engineering. *ACM Transactions on Software Engineering and Methodology*, 6(1):1–30, 1997.

Ubiquity of People: Understanding People's Diversity for Effective Knowledge Discovery

Anett Kralisch and Bettina Berendt

Institute of Information Systems, Humboldt University Berlin
http://www.wiwi.hu-berlin.de/{kralisch|berendt}

User-centred knowledge discovery explores a perspective that hitherto has been under-represented in the field of knowledge discovery: the user. The shift from a technological focus towards an emphasis on human aspects asks for a multidisciplinary approach, integrating insight from behavioural, psychological, and linguistic sciences into the field of knowledge discovery. In this paper, we introduce the concept *ubiquity of people* to emphasize that data and knowledge is created and accessed globally, from users who differ in language, culture, and other factors. The paper investigates how knowledge discovery may benefit from an integration of the concept of ubiquity of people. We provide an overview of the impact of language and culture on how data and knowledge is accessed, shared, and evaluated. We conclude with a discussion of research questions that are raised by the integration of the ubiquity of people into knowledge discovery, in particular with regard to data collection, data processing, and data presentation

Keywords: user-centred knowledge discovery, ubiquitous knowledge discovery, ubiquity of people, culture, multilingualism

1 Introduction: Why Knowledge Discovery for Users (KDU)? Why Ubiquitous Knowledge Discovery for Users (UKDU)?

Knowledge discovery is more than the application of algorithms – it encompasses the whole process of turning data into knowledge: business / application understanding, data understanding, data preparation, modelling, evaluation, and deployment. Users play a pivotal role in this process: they create data, data and knowledge are about them, and they are the ultimate beneficiaries of the discovered knowledge. Data-creating activities include the authoring of documents and of references and links between documents, explicit reactions to questions such as the input of registration data, and the behaviour that leaves traces, biometric measurements, etc. in log files. These data are often transformed into user models, i.e. into knowledge about users. These user models, in turn, are the foundation for general-audience or personalized improvements of services and devices, activities which directly benefit the end user or another human stakeholder (the owner of a Web site, a government agency, etc.).

Because of this central role of the user, an understanding of users is required for application / business understanding, for data understanding, for the evaluation of discovered patterns, for the deployment of results, and for all other KD activities that depend on these steps. These considerations lead to the recognition that Knowledge Discovery should always be *Knowledge Discovery for Users.*
Understanding users involves understanding differences between users. While economic differences and some psychological differences (such as learning styles) have been recognised as a factor in KD and HCI for some time, most studies investigate groups of users living in one country and speaking one language and may find differences within these groups. This can become a problem both for understanding and catering to worldwide users. The purpose of the present article is to give an overview of empirical studies on the effects of differences between worldwide groups on behaviours and attitudes relevant to IT usage, and to argue that integrating these insights into KD is a key step towards making Knowledge Discovery for Users become *Ubiquitous Knowledge Discovery for Users.*

The present article builds on our investigations, summarised in (Kralisch, 2006), of the impact of culture, language, and related factors on the use of the Internet. For technical details of how these concepts can be employed in Web mining, we refer the reader to these earlier studies. The contributions of the present article are twofold: First, we introduce the concept of the ubiquity of people, and second, we propose a first general framework for integrating this concept into KD. KD in this sense is any kind of KD involving worldwide audiences – on the Web or in other "smart environments" – and thus *Ubiquitous Knowledge Discovery* in the various senses explored in this UKDU'06 workshop.

After introducing the new concept of ubiquity of people in Section 2, we will argue in Section 3 how KD and its applications can benefit from insights about it. In Section 4, language and culture will be investigated as prime aspects of the ubiquity of people, and other aspects will be mentioned. Implications for KD will be outlined in the concluding Section 5.

2 Ubiquity of People as a New Concept in User-centred Knowledge Discovery

"Ubiquity" is the property or ability to be present everywhere or at several places at the same time (Oxford English Dictionary). "Ubiquitous" is one of the most characteristic traits of the Internet, referring to its (potential) global access and use. The W3C consortium defines "ubiquitous web" as "web access for anyone, anywhere, anytime, using any device" (www.w3.org/UbiWeb/). The term "ubiquity" is established in Information Systems and in computing and often used in the context of services that are available anytime and anywhere (e.g. mobile commerce). With regard to "ubiquitous computing", "ubiquity" predominantly relates to computation embedded (everywhere) into our environment, with devices that are intelligent and interconnected. This notion of ubiquity focuses on technological aspects and only marginally interprets the word in geographical terms.

However, with increasing technological progress, human factors come into the foreground, replacing the erstwhile importance of technological challenges. With a shift from technological areas towards human aspects, the focus in ubiquity research shifts in equal measure to the ubiquity of people.

Data and knowledge sources such as the Internet are accessed and used by users that grew up and reside in different parts of the world. *Ubiquity of people* consequently encompasses diversities with regard to language and culture, but also divergences in economic and social status, technological skills and educational skills. Often, these factors are intertwined and cannot be clearly separated. A connection between language and technological issues is for example given by the fact that languages with a writing system based on Roman characters (e.g., English, German, French) face significantly fewer technical problems in publishing information than languages with different writing systems (e.g., Mandarin, Inuktitut).

Why do we introduce a new concept, and what are its relations to other meanings of "ubiquitous"? The concept "ubiquity of people" can serve as one bridge[1] from the technology "ubiquitous computing" (in the sense of mobile, embedded, etc.) to the goal "global access" (in the sense of equal access). Operationally, the concept calls for extensions to user and context modelling, and for extensions to system design. To what extent these extensions only concern content (e.g., further attributes in user models, offers of different layout options) or also formal aspects of modelling and implementation details, is a question for further research. We will discuss both the "why" and "how" of the concept in more detail in the following section.

We focus in our paper on linguistic and cultural differences since diversity with regard to these aspects is the most apparent and important in a global context. Global access to data and knowledge eliminates the constraints that were long imposed by geographical location. At the same time globalisation raises the challenge of meeting individuals' divergent abilities, perceptions, and preferences.

3 How Does Knowledge Discovery Benefit from Insight about Ubiquity of People?

The increasing speed of technological progress makes it possible to regard technological barriers as obstacles that can usually be overcome within short periods of time. In contrast, diversities that are inherent to the users, such as their linguistic and cultural backgrounds, are much more stable over time and should be regarded as an established fact for which alternative solutions need to be found.

Such heterogeneity among users leads to heterogeneity in data sets. Data sets are also heterogeneous because data is generated in different contexts. For example, if the market share of a service within a group of native speakers is used as an indicator of a service's success, a comparison between different services and different groups of native speakers is more accurate if the number of potential alternative services is considered as well. Hence, in order to extract meaningful information and knowledge

[1] It is only one bridge because technical, economical, political, etc. factors also play important roles; it is also a bridge from computing that is ubiquitous in other senses (see Section 1).

from a global data set, the diversity of contexts needs to be taken into account. To recognize patterns, ubiquitous data processing must incorporate human factors of accessing and using data and knowledge.

The integration of human aspects into the field of ubiquitous knowledge discovery calls for an interdisciplinary approach where insights from psychological, behavioural, and linguistic sciences provide the necessary background for adequate data and knowledge gathering, processing, and presentation. Psychological, behavioural, and linguistic sciences provide insight into existing divergences between users and potential barriers for knowledge acquisition and generation. These disciplines investigate which variables may determine whether or not and to which extent people are able to access data and knowledge, how they evaluate information, and to which degree they are willing and able to share information.

This type of background knowledge is a prerequisite for efficient and accurate data collection as well as for meaningful and correct data processing and interpretation. To give an example: as shown later in this article, an individual's cultural background is a major determinant of his/her attitude towards privacy and data disclosure which in turn might affect the ease of data gathering and the correctness of the data provided.

Finally, one can derive guidelines on appropriate data and knowledge presentation: either directly from insight about the impact of the users' cultural and linguistic background or indirectly from results obtained from data processing that takes cultural and linguistic factors into account. This raises the question which design choices diminish the "digital divide", i.e. which design choices help to provide equal access and encourage participation in knowledge acquisition and generation. Design choices regard technologies that are able to bridge the gap between linguistically and culturally divergent users as well as technologies that are adapted to the different needs of different users.

Thus, a focus on the ubiquity of people contributes to equal *access* to data and knowledge. An accommodation of the ubiquity of people can be understood as the antonym of digital divide: it aims to assure that people independently of their locations, linguistic, cultural, or social backgrounds are able to use the Internet or other global services as information sources. In a second step, this also includes a successful knowledge exchange and knowledge generation across linguistic and cultural borders. This is, for example, particularly important for international work groups or distance-learning groups.

Why not just find out that (say) German Web users like layout type X and US Web users like layout type Y, provide two versions of the software, and regard this as sufficient for globalisation/localisation? While simple solutions like this one may be applicable in certain circumstances, our examples below illustrate that the empirical findings often call for a more differentiated view. We will argue that the ubiquity of people consists of a number of (relative stable) "traits" and (relatively dynamic) "states" that often arise from the interaction of a user's traits with the environmental context. In user and context modelling (cf. Heckmann, 2005; Jameson, 2001), traits and states are often assembled in the user model (e.g., demographics, personality variables, skills, emotional and physical states), while properties of the environment are assembled in the context model (e.g., weather conditions), and interactions are modelled as influence relations. The concepts we investigate show the importance of

dynamics and of the interaction between user and context. For example, a person's "culture" is often equated with her country of origin, but it may also be the country where she has spent most of their recent life, and it may extend to, e.g., the professional culture a person inhabits or from which she accesses a given piece of information. Both may shift over time. Similarly "linguistic background" generally refers to a person's native language. On the other hand, whether a person operates in her native language or in a non-native language is a function also of the environment (here, the language of the information accessed). The same holds for domain knowledge.[2]

4 Understanding User Diversity: The Role of Language and Culture

In this section, we provide an overview about cultural and linguistic studies that analyse the impact of culture and language. Their results provide a first outline of background knowledge necessary for successful ubiquitous data collection and processing. We focus mainly on user behaviour regarding the Internet since it is today the primary example of a global information source. We emphasise how culture and language affect (1) access to data and knowledge, (2) people's willingness and ability to share it, and (3) their evaluation of information. These three aspects interact in several ways, so a clear line cannot always be drawn. If, for example, access to data and knowledge is difficult or impossible, sharing of information is restricted as well.

4.1 Language

When a user accesses linguistic content in a service, she may access content presented in her first (native) language and thus find herself in an "L1 situation". She may also access content in a second (non-native) language and thus find herself in an "L2 situation". Since this distinction is the most basic and best-researched variation in language, we will operationalize "language" as L1 vs. L2.

Access to Data and Knowledge. Availability of content differs between languages to a major extent. A myriad of articles (e.g. Danet & Herring, in press) describe the original dominance of English language content on the Internet. Over the years the monolingual dominance has been increasingly counterbalanced by other, widely spoken languages (see statistics by Global Reach, 2006). Nevertheless, the majority of languages is still underrepresented on the Internet and will probably never attain an adequate and proportional representation. This can be attributed to various factors such as the number of native speakers, their economic importance, or the distribution of the Internet in certain areas. These factors diminish the importance of a group of native speakers as a target group and limit at the same time the number of potential native language website creators.

[2] An interesting extension would be a differentiation between "acting in a familiar culture" and "acting in an unfamiliar culture".

If native language (L1) content is not available or limited to a few topics, users are forced to access information in a non-native language (L2).[3] The L2 is usually English, and sometimes the area's lingua franca, such as Russian for communication within the former Russian republics (Wei & Kolko, 2005). Depending on a user's L2 proficiency levels, access to data and knowledge might not be possible or reduced to a minimum amount. Palfreyman and Al-Khalil (2003) find that even in a diglossic[4] situation, such as commonly found in Arab countries, the use of the high dichotomy (= standard language) often constitutes a major barrier for users with lower education.

Kralisch and Mandl (2006) and Halavais (2000) show by means of logfile analyses that websites are indeed favoured by native speakers, even if the respective percentages of native speakers and content alternatives are taken into account.

The ease of reading information is only one aspect of the accessibility of data and knowledge. In fact, accessibility is also determined by the ease of rendering available information. From a technological point of view, restrictions in the usage of characters present a significant inconvenience for certain language groups. The ASCII Code, originally based on the English language, favours writing systems that are derived from the Latin alphabet ("typographical imperialism" – Phillipson, 1992; Phillipson & Skutnabb-Kangas, 2001). Speakers of languages that are based on different writing systems (e.g. Cyrillic alphabet, Chinese signs) are disadvantaged by the ASCII code (Pargman & Palme, 2004) and forced to find work-arounds such as visual numbers (Palfreyman & Al-Khalil, 2003; Tseliga, in press). Also, due to the still common use of ASCII code, access to less widespread writing systems (e.g Inuktitut) usually requires downloading the specific character set (Herring & Estrada, 2004). Covering almost all writing systems in current use today, the introduction of Unicode is an essential step towards multilingual content generation and multilingual computer processing and hence towards equal accessibility.

Sharing Data and Knowledge. Language may constitute a central barrier for sharing data and knowledge. With increasing distribution of the Internet, the role of the English language as the Internet's lingua franca is significantly reduced: the percentage of English native speakers and highly proficient L2 users decreases, whereas the number of different native languages grows (Global Reach 2006). Consequently, communication barriers due to the lack of a common language increasingly arise.

Herring and her colleagues (2006) discuss, in a study of language networks on LiveJournal, the role of language as a determinant of network generations and in consequence as a determinant of data and knowledge exchange which occurs mainly *within* each network. It should be noted that the networks' sizes and densities differ between languages. Herring et al. explain these divergences by differences in the numbers of users per language and in the degree of bilingualism and point out that a critical mass is necessary to create a robust language network. Similar phenomena

[3] The shift to an L2 situation may also involve other causes, including the possibility to access a larger repository, the wish to compare different opinions, the need to retrieve information in English because it is better exploitable in other contexts.

[4] In simplified terms, diglossia describes a linguistic situation in which a standard form of a language with high prestige (e.g. classic Arabic) and a dialect form with lower prestige (e.g. the regional forms of Arabic) are spoken in a society. For details, see (Ferguson, 1959).

are reported about the use of the multilingual European discussion forum Futurum. Despite the fact that participation in the language of choice is encouraged (a battery of translators ensure translation between languages), communication in English clearly dominates. Moreover, discussion threads that are introduced in other languages tend to be shorter (Wodak & Wright, 2004).

Since multilingualism will increase rather than diminish with the growing distribution of global technology[5], data gathering and data processing will more and more be required to take these barriers into consideration as well as consider the use of multilingual technologies.

Evaluation of Information. In contrast to culture (see below), the impact of language on information evaluation is rather indirect. Language predominantly determines how easy it is to access data and knowledge. According to Davis' (1989) model of technology acceptance (TAM), ease of use is a determinant of usefulness, attitude towards an information system, and satisfaction. It can therefore be assumed that information in a user's native language leads to a better evaluation (Kralisch & Köppen, 2005). This effect might be strengthened if the native language is perceived as an identity-constituting factor.

4.2 Culture

"Culture" is a many-faceted and controversial term. In its most general meaning , it denotes attitudes and behaviours of a group that are relatively stable over time, and the term is also used to denote the group united by these commonalities. Many studies that are relevant for IT understanding and design have operationalized culture as a country or a collection of countries. While we are aware of the problems induced by this reading of "culture", we adopt it as a working definition (for a detailed discussion, see Kralisch, 2006). The reasons are the predominance in the literature and the relevance for applications that often aim at understanding and opening a new market that is defined by the same boundaries: countries or country groups.

Access to Data and Knowledge. In contrast to language, culture represents a less visible obstacle to data and knowledge access. However, the Internet itself "[…] is not a culturally neutral or value-free space in which culturally diverse individuals communicate with equal ease" (Reeder, Macfadyen, Chase, & Roche, 2004). It is a cultural product that reflects the cultural values of its Anglo-American producers. "Their … cultures value aggressive/competitive individualistic behaviours. … These cultural value communications are characterized by speech, reach, openness, quick response, questions/debate and informality" (Reeder et al., 2004). A number of studies investigate the impact of the Internet's cultural values on accessing data and knowledge. Their results provide a differentiated picture with sometimes contradictory outcomes.

Reeder et al. (2004) state that due to the implicit cultural values embodied by the Internet, English speaking, academic cultures have the least difficulty in communicating over the Internet (see also De Mooij, 2000).

[5] This expectation is based on past and current developments, see (Global Reach, 2006)

Hongladarom (2004) describes the efforts carried out by the Thai government to encourage Internet access among all classes of the Thai society. The author points out that, despite the government's technological and economic efforts (such as hardware and software support), the success of this initiative is rather limited, which the author considers a result of the Internet's implicit cultural values which contradict traditional Thai values. These results are in line with Warschauer's (2003) argumentation that the digital divide is interrelated with socio-cultural factors. Hongladarom's and Warschauer's findings are contradicted by Dyson's study (2004). Dyson finds in his analysis of adoption of Internet communication technologies among indigenous Australians that the cultural values of the Internet do not represent an obstacle for accessing data and knowledge. Other factors are pointed out by the author as the cause of the low adoption rate.

Beside the negative or neutral impact of the Internet's cultural values, the characteristics of the Internet have also been shown to positively affect the participation of users from societies with divergent cultural values. Several studies point out that the impersonality of Internet communication encourages women from traditional, high Power Distance countries (e.g. Thailand) to participate more freely in Internet use and communication (Panyametheekul & Herring, 2003; Wheeler, 2001). A similar effect was shown for Asian students. Asian teaching styles are characterized as authoritative with strong hierarchies between students and teachers. The impersonal communication style on the Internet encourages Asian students to participate more during lessons (e.g. Bauer, Chin, & Chang, 2000).

Sharing Data and Knowledge. Whereas language predominantly affects the *capacities* of sharing data and knowledge, culture has a major impact on users' *willingness* to share information. Within this regard, we consider the aspects of self-conception, their attitude towards privacy, and a cultural groups' hierarchical organisation as the most important factors.

A group's self-conception and cultural identity play a major role with regard to the conception of ingroups and outgroups and its impact. The strength of ingroup and outgroup perception is strongly correlated with the cultural dimension of individualism and collectivism[6]: collectivistic cultures – in contrast to individualistic societies – tend to strongly differentiate between ingroup and outgroups. Various studies indicate that, as a consequence, traditional and collectivistic cultures fear fraud and moral damage through the information that is provided on the Internet.

Privacy issues have also been shown to be affected by a culture's degree of individualism. Members of individualistic cultures tend to be less willing to provide sensitive information than members of collectivistic cultures. This can be explained by the observation that individualistic cultures value private space more than collectivistic cultures (e.g. Milberg, Smith, & Burke, 2000).

Individualistic and collectivistic cultures also differ in the type of information they provide when negotiating identity (Burk, 2004; Reeder et al., 2004). "It is likely that in some cultural settings, information considered highly personal by Western stan-

[6] Individualism and Collectivism are cultural dimensions developed by Hofstede (1991). Individualism implies loose ties between the members of a society; collectivism implies that people are integrated into strong, cohesive groups (Marcus & West Gould, 2000).

dards, such as wealth or spending habits, may be deemed open and public, whereas information considered relatively innocuous in Western settings, such as a nickname, might be considered extremely private" (Burk, 2004). Debnath and Bhal (2004) point out that ethical issues related to privacy differ among Indian citizens depending on their acquired norms of ethical and moral conduct. Burk (2004) emphasizes that "privacy as a matter of individual autonomy may be relatively unimportant in cultural settings where communal information is unlikely to be accommodated within the data protection models now sweeping across the globe".

In addition, power distance[7] has a similar impact on users' willingness to disclose data, with members of high power distant countries being more willing to provide data than members of low power distant countries (Kralisch, 2006). However, within high power distant societies, knowledge sharing from high hierarchies to low hierarchies is difficult since it would transfer decision making authorities to subordinates. Heier and Borgman (2002) describe how this effect challenged the international HRbase of Deutsche Bank: usage rates were about 20% in Germany and the UK but only about 4% in Asian countries.

Evaluation of Information. Differences in information evaluation are strongly related to differences / compliances in communication styles. In particular, the preference for face-to-face (personal) communication over impersonal communication is pointed out as an important cultural factor of Internet use and information evaluation. High preference for personal communication usually leads to negative evaluation of information from outsider groups; this information is deemed less reliable. Reliability of information is attributed in these cultures to the reliability of its carrier. Technologies are not seen as an equivalent of interpersonal communication and are therefore not as trustworthy. "Technologies that facilitate interpersonal connections among people who want to stay in touch (e-mail, cell-phones) would be [therefore] adopted much faster than impersonal devices (…, web-info sites, e-commerce, call-centers, automated messages) (Markova, 2004). In line with these findings, results from a study conducted by Siala and his colleagues (2004) reveal that collective cultures buy mainly from within-group members. Similarly, Jarvenpaa and Tractinsky (1999) found that trust in e-commerce is culturally sensitive.

Furthermore, members of different cultural groups have different approaches towards contradicting information and its evaluation. Here, again the level of power distance appears to have an important impact with members of high power distant countries more accepting provided information unquestioningly. Markova (2004) describes the cultural concept of information in central Asia: information is not searched or evaluated, but it is memorized the way it is "taught". The cultural belief in objective truth is supported by government-controlled accessed to information that inhibits access to conflicting information. This finding is in line with the teaching styles in high power distant countries (see "knowledge sharing across hierarchical levels" above).

[7] One of Hofstede's cultural dimensions that describes the extent to which less powerful members of institutions and organisations accept that power is distributed unequally (Hofstede, 1991).

Finally, culture also shapes assumptions about which knowledge is important (DeLong & Fahey, 2000). De la Cruz et al. (2005) show that members of different cultures assign different importance to the same website criteria. As a consequence the quality of websites is interpreted differently. More detailed research is however required in order to specify the relationship between cultural values and importance of information elements.

Last but not least, depending on their cultural background users differ in their way they express their evaluation. Evers and Day (1999) have therefore developed recommendations that help normalize evaluations from users with different cultural backgrounds. For example, they recommend 6-point Likert scales to avoid neutral positions that are often adopted by members of collectivistic cultures.

4.3 Factors beyond Language and Culture

We introduced ubiquity of people as a term that describes people's diversity and the call for the provision of equal access. In view of increasing globalisation we focused on linguistic and cultural aspects. Nevertheless, people's diversity covers more than these traits. We mentioned in the introductory part that differences in technological progress, social and economic status as well as levels of education are further factors that affect access to data and knowledge, sharing and evaluation. These factors often mediate or reinforce the impact of culture and language on user behaviour. An individual's level of education (or domain knowledge) appears to be particularly related to his/her linguistic abilities. Economic/social status and technological progress often strengthen the impact of cultural values but also linguistic abilities.

The results of Kralisch and Berendt (2005) indicate that L2 users with high domain knowledge manifest the same preferences in their search for information as L1 users. In contrast, L2 users with low domain knowledge show divergent preferences. In other words, domain knowledge can compensate for language deficiencies. More generally, a user's education plays a major role in areas with lower internet distribution and large educational divergences within a society. Education affects access to information since usually only elites have access to the Internet and possess the necessary computer skills (e.g. Mafu, 2004) and/or the literacy to take full advantage of complex content. As described above, in situations of diglossia, higher education usually assures a higher language proficiency level in the standard language. This is particularly true when it comes to productive mastering of written language.

Linguistically and culturally determined difficulties in accessing data and knowledge are often complicated by lack of access to technologies in remote areas or high access fees. Dyson (2004), for example, attributes the lower adoption rate among indigenous Australians to limited access to Internet communication technologies, to high costs, poor telecommunication infrastructure, and low computer skills.

5 Conclusions: Research Questions for User-centred Knowledge Discovery in a Global Context

We presented aspects of how language and culture may influence the way people access data and knowledge, share them, and evaluate them. These aspects were pointed out as necessary background knowledge for user-centred knowledge discovery that deals with the ubiquity of people. We close by discussing research questions in user-centred knowledge discovery that are raised by people's ubiquity. We focus on three aspects of knowledge discovery: data gathering, data processing, and data presentation.

Data Collection. Data collection from ubiquitous users must cope with two major problems: challenges of obtaining data and challenges of their representativeness.

Data collection efforts that rely on users' self-reports need to consider that users differ in their ability to provide information as well as in their willingness to share it. Language, for example, may constitute a major barrier to access questionnaires or websites and may limit users in their abilities to answers questions. Users' willingness to disclose data is highly culturally determined, as shown by studies of individualistic and collectivistic cultures. A low willingness to share information leads to the question of how reliable the information provided by the user is. Culturally determined differences in privacy issues and willingness to share information ask for a detailed examination of the extent to which data gathering would constitute an intrusion into the private space. As shown above, differences may regard users' general attitudes as well as specific types of information. Privacy research has shown that presenting reasons for collecting data creates confidence among users and augments the amount of data provided by them (Kobsa & Teltzrow, 2005). In a global context, the presented reasons might need to be reviewed and adapted to the local needs and preferences.

Given the differences in accessibility of data and knowledge sources, the question is raised how representative the investigated group is. An English-language Internet questionnaire might for example be answered by a wide range of English native speakers that differ largely in their economic and social status and educational levels. At the same time the questionnaire might cover an only very specific group of Arab native speakers, namely those that are well educated and affluent.

Furthermore, a user's cultural background affects the way opinions are expressed. This needs to be taken into account through either a culturally adapted conception of data gathering tools or through appropriate data processing that considers these differences.

Non-reactive methods are challenged by the difficulties of correctly assessing a user's cultural background and linguistic abilities. Analyses of logfiles and IP addresses can be considered only proxies with a limited certainty of the data collected. Analyses of IP addresses are for example used to obtain information about the location from which the Internet is accessed. Information about the location in turn helps to derive information about a user's linguistic and cultural background, but involve a certain error (Kralisch & Berendt, 2005; Kralisch, 2006).

Data Processing. Ubiquity of people leads to heterogeneous data sets due to different contexts. The context may attribute different values to each collected data item. User-

centred knowledge discovery hence requires data processing that takes background knowledge about the users and their context into account.

For example, if a user accesses a website despite major linguistic challenges, this might signify a higher relevance of the website's content or service. In cases where a relationship between use/access and relevance (or other attributes) is established, data processing becomes significantly more accurate if weighted measures that consider these challenges are used. In a similar manner, if the amount of generated content/ services is analysed as an indicator of need or interest, context information about difficulties of content/service generation render the analysis more accurate.

Given the increasing amount of multilingual data sets, knowledge discovery should also take into consideration research results regarding multilingual information retrieval tools or information retrieval tools that take cultural aspects into account. For example, Kralisch and Mandl (2005) provide a first overview how the users' cultural backgrounds affect the use of information retrieval tools.

Data Presentation. Information about the impact of language and culture on data and knowledge accessibility provides important insight into suitable forms of data presentation. Further insight can be obtained through appropriate data processing. Culture and language are two factors that affect people's abilities and preferences for certain forms of data presentation. For example, the outcomes of Kralisch and Berendt (2004) indicate that users from high power distant countries prefer a hierarchical form of knowledge presentation more than members of low power distant countries. Kralisch, Berendt, and Eisend (2005) propose design guidelines based on this and related findings. Further divergent preferences were found with regard to other cultural dimensions. In a similar manner, Yeo and Loo (2004) present cultural differences in preferences for classification schemes.

Research on data presentation forms also involves the development of technologies that are able to bridge the gap between different cultures and languages, such as multilingual information retrieval tools. However, where bridging the gap is not conceivable or feasible, adaptations to the users' cultural and/or linguistic needs are necessary. User-centred knowledge discovery should therefore also aim to discover the threshold of where adaptations to the user's linguistic and cultural needs are necessary and where other solutions are more efficient and/or appropriate.

In future work, we also plan to investigate the technical implications of these findings. In particular, we intend to explore how the ubiquity of people can be reflected in user and context modelling, and put to use in the processes by which these models enter KD and its deployment for user adaptation (see Section 2 and Heckmann, 2005; Jameson, 2001).

Finally, it should be noted that the studies presented in Section 4 were mostly intracultural or bicultural comparisons. In order to obtain a wide range of background knowledge, *multi*cultural comparative studies are necessary (see also Danet & Herring, 2003).

6 References[8]

Bauer, C., Chin, K., & Chang, V. (2000). *Web-Based Learning: Aspects of Cultural Differences.* Proc- of the 8th European Conf. on Information Systems (ECIS), July 3-5, Vienna, Austria.

Burk, D. (2004). *Privacy and Property in the Global Datasphere: International Dominance of Off-the-shelf Models for Information Control.* In: Sudweeks & C. Ess (2004), pp. 363-373.

Danet, B., & Herring, S. (2003). Introduction: The Multilingual Internet. *Journal of Computer-Mediated Communication, 9*((1)), http://jcmc.indiana.edu/vol9/issue1/intro.html.

Danet, B., & Herring, S. (in press). Multilingualism on the Internet. In: M. Hollinger & A. Pauwels (Eds.), *Language and Communication: Diversity and Change* (Vol. IX), Berlin: Mouton de Gruyter.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*, 319-340.

Day, D., & Evers, V. (1999). *Questionnaire development for multicultural data collection.* In: E. del Galdo & G. Prahbu (Eds.), Proc. of the Int. Workshop on International Products and Systems, May, 20-22, Rochester, UK.

de la Cruz, T., Mandl, T., & Womser-Hacker, C. (2005). *Cultural Dependency of Quality Perception and Web Page Evaluation Guidelines: Results from a Survey.* In: D. Day, V. Evers & E. del Galdo (Eds.), Proceedings of the 7th IWIPS, 15-27, July 7-9, Amsterdam, The Netherlands.

De Mooij, M. (2000). The Future is Predictable for International Marketers: Converging Incomes Lead to Diverging Consumer Behaviour. *International Marketing Review, 17*(2), 103-113.

Debnath, N., & Bhal, K. (2004). *Religious Belief and Pragmatic Ethical Framework as Predictors of Ethical Behavior: An Empirical Study in the Indian Context.* In: Sudweeks & C. Ess (2004), pp. 409-420.

DeLong, D., & Fahey, L. (2000). Diagnosing Cultural Barriers to Knowledge Management. *Academy of Management Executive, 14*(4), 113.

Dyson, L. (2004). *Cultural Issues in the Adoption of Information and Communication Technologies by Indigenous Australians.* In: Sudweeks & C. Ess (2004), pp. 58-71.

Ferguson, C. (1959). Diglossia. *Word, 15*(2), 325-340.

Global Reach. (2006). *Global Internet Statistics (by Language)*. global-reach.biz/globstats/index.php3

Halavais, A. (2000). National Borders on the World Wide Web. *New Media and Society, 2*(1), 7-28.

Heckmann, D. (2005). *Ubiquitous User Modeling.* PhD dissertation, Universität Saarbrücken, Saarbrücken, Germany, [online], Available at: w5.cs.uni-sb.de/publication/file/178/Heckmann05Diss.pdf

Heier, H., & Borgman, H. (2002). *Knowledge Management Systems Spanning Cutures: the Case of Deutsche Bank's HRbase.* Proceedings of the 10th ECIS, June 6-8, Gdansk, Poland.

Herring, S., & Estrada, Z. (2004). *Representations of Indigenous Language Groups of North and South America on the World Wide Web: In Whose Voice?* In: Sudweeks & C. Ess (2004), pp. 377-380.

Herring, S., Paolillo, J., Clark, B., Kouper, I., Ramos-Vielba, I., Scheidt, L. A., Stoerger, S., & Wright, E. (2006). *Linguistic Diversity and Language Networks on LiveJournal.* Proceedings of the INSNA Sunbelt Conference, Vancouver, Canada.

Hofstede, G. (1991). *Cultures and Organizations: Software of the Mind*. London: McGraw-Hill.

Hongladarom, S. (2004). *Global Culture, Local Cultures, and the Internet: The Thai Example.* In: Sudweeks & C. Ess (2004), pp. 187-201.

Jameson, A. (2001). Modeling both the context and the user. *Personal Technologies, 5 (1),* 29-33.

Jarvenpaa, S. L., & Tractinsky, N. (1999). Consumer Trust in an Internet Store: A Cross-Cultural Validation. *Journal of Computer-Mediated Communication, 5*(2), [online], Available at: http://jcmc.indiana.edu/vol5/issue2/jarvenpaa.html

Kobsa, A., & Teltzrow, M. (2005). Impacts of Contextualized Communication of Privacy Practices and Personalization Benefits on Purchase Behavior and Perceived Quality of Recommendation. In: M. Van Setten, S. McNean & J. Konstan (Eds.), *Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research (IUI 2005)*, 48-53, San Diego, CA, USA.

Kralisch, A. (2006). *The Impact of Culture and Language on the Use of the Internet: Empirical Analyses of Behaviour and Attitudes.* PhD dissertation, Humboldt University, Berlin. Published at http://edoc.hu-berlin.de/docviews/abstract.php?id=27410

Kralisch, A., & Berendt, B. (2004). *Cultural Determinants of Search Behaviour on Websites.* In: V. Evers, E. del Galdo, D. Cyr & C. Bonanni (Eds.), Proc. of the 6th IWIPS, 61-74, Product & Systems Internationalisation, Inc., July 8-10, Vancouver, Canada.

[8] All online sources were last accessed on June 20th, 2006.

Kralisch, A., & Berendt, B. (2005). Language-sensitive Search Behaviour and the Role of Domain Knowledge. *New Review of Multimedia and Hypermedia: Special Issue on Minority Language, Multimedia and the Web, 11*(2), 221-246.
Kralisch, A., Eisend, M. Berendt, B. (2005). Impact of Culture on Website Navigation Behaviour. In *Proc. of 11th Int. Conf. on Human-Computer Interaction, Las Vegas, NE, 22-27 July 2005.*
Kralisch, A., & Köppen, V. (2005). *The Impact of Language on Website Use and User Satisfaction: Project Description.* In: D. Bartmann et al. (Eds.), Proc. of the 13th ECIS, May 26-28, Regensburg, Germany.
Kralisch, A., & Mandl, T. (2005). *Intercultural Aspects of Design and Interaction with Retrieval Systems.* In *Proc. of 11th Int. Conf. on Human-Computer Interaction, Las Vegas, NE, 22-27 July 2005.*
Kralisch, A., & Mandl, T. (2006). *Barriers to Information Access across Languages on the Internet: Network and Language Effects.* Proceedings of the 39th Hawaii International Conference on System Sciences (HICSS-39), IEEE Computer Society, January 4-7, Poipu, HI, USA.
Mafu, S. (2004). From the Oral Tradition to the Information Era: The Case of Tanzania. *International Journal of Multicultural Societies, 6*(1), 53-78.
Marcus, A., & West Gould, E. (2000). *Cultural Dimensions and Global Web User-Interface Design: What? So what? Now what?* Proc. of the 6th Conf. on Human Factors and the Web (HFWeb), June 19, Austin, TX, USA.
Markova, M. (2004). *Barriers for Informationo Technology Adoption in Post-Soviet Central Asia.* In: Sudweeks & C. Ess (2004), pp. 277-281.
Milberg, S. J., Smith, H. J., & Burke, S. J. (2000). Information Privacy: Corporate Management and National Regulation. *Organizational Science, 11*, 35-37.
Palfreyman, D., & Al-Khalil, M. (2003). A Funky Language for Teenzz to Use: Representing Gulf Arabic in Instant Messaging. *Journal of Computer-Mediated Communication, 9*(1), http://jcmc.indiana.edu/vol9/issue1/palfreyman.html.
Panyametheekul, S., & Herring, S. (2003). Gender and Turn Allocation in a Thai Chat Room. *Journal of Computer-Mediated Communication, 9*(1), jcmc.indiana.edu/vol9/issue1/panya_herring.html.
Pargman, D., & Palme, J. (2004). *Linguistic Standardization on the Internet.* In: Sudweeks & C. Ess (2004), pp..385-388.
Phillipson, R. (1992). *Linguistic Imperialism*. Oxford: Oxford University Press.
Phillipson, R., & Skutnabb-Kangas, T. (2001). Linguistic Imperialism. In: R. Mesthrie (Ed.), *Concise Encyclopedia of Sociolinguistics*, 570-574, Oxford: Elsevier Science.
Reeder, K., Macfadyen, L. P., Chase, M., & Roche, J. (2004). *Falling through the (Cultural) Gaps? Intercultural Communication Challenges in Cyberspace.* In: Sudweeks & C. Ess (2004), pp. 123-134.
Siala, H., O'Keefe, R. M., & Hone, K. (2004). The Impact of Religious Affiliation on Trust in the Context of Electronic Commerce. *Interacting with Computers, 16*(1), 7-27.
F. Sudweeks & C. Ess (Eds.) (2004). Proceedings of the Fourth International Conference on Cultural Attitudes towards Technology and Communication (CATaC), School of Information Technology Murdoch University, June 27 - July 1, Karlstad, Sweden.
Tseliga, T. (in press). "It's all Greeklish to me!": Linguistic and sociocultural perspectives on roman-alphabeted Greek in asynchronous computer-mediated communication. In: B. Danet & S. Herring (Eds.), *The Multilingual Internet: Language, Culture and Communication Online*, New York: Oxford University Press.
Warschauer, M. (2003). *Technology and Social Inclusion: Rethinking the Digital Divide*. Cambridge, MA: MIT Press.
Wei, C., & Kolko, B. (2005). Resistance to Globalization: Language and Internet Diffusion Patterns in Uzbekistan. *New Review of Multimedia and Hypermedia, 11*(2), 205-220.
Wheeler, D. (2001). Old Culture, New Technology: A Look at Women, Gender and the Internet in Kuwait. In: C. Ess & F. Sudweeks (Eds.), *Culture, Technology, Communication: Towards an Intercultural Global Village*, Albany: Suny.
Wodak, R., & Wright, S. (2004). The European Union in Cyberspace: Democrative Participation via Online Multilingual Discussion Boards? In: B. Danet & S. Herring (Eds.), *The Multilingual Internet: Language, Culture and Communication Online*, New York: Oxford University Press.
Yeo, A., & Loo, W. (2004). *Identification and Evaluation of Classification Schemes: A User-Centred Approach.* In: V. Evers, E. del Galdo, D. Cyr & C. Bonanni (Eds.), Proceedings of the Sixth International Workshop on Internationalisation of Products and Systems (IWIPS), 75-87, Product & Systems Internationalisation, Inc., July 8-10, Vancouver, Canada.

Discovering user communities on the Web and beyond
(Invited Talk – Abstract)

Georgios Paliouras

National Centre for Scientific Research, Athens, Greece
http://iit.demokritos.gr/~paliourg/

In this talk I will provide an overview of our work on discovering user communities from Web usage data and will highlight interesting research directions within the realm of the Web and beyond. Particular emphasis will be given to the potential of user communities in addressing new challenges introduced by ubiquitous communication environments.

The motivation for our work has been two-fold: the pressing need for personalized information systems on the Web and the availability of recorded usage data. Our approach to this problem has been to provide methods to acquire knowledge from data, in order to construct models of user behaviour, particularly user communities. These user models can then be used for various types of personalization, such as adaptive Web sites and Web page recommendation. The talk will mention briefly our early work on the discovery of user communities and will emphasize mostly our recent work on the community Web directories and the use of grammars as a model of Web navigation. In contrast to most of the work on Web personalization and our own earlier work that has focused on particular Web sites, we are currently working with user navigation data for the entire Web, that exhibit a high degree of thematic diversity. Addressing this issue, we propose methods for discovering hierarchical models with probabilistic latent semantic analysis, as well as grammar inference methods for navigation modelling. I will present experimental results with real usage data, providing both positive and negative evidence for the proposed methods and raising interesting issues for further research.

Furthermore, I will present our view of how user communities could prove useful in new personalization tasks in the realm of the Web and beyond. Particularly, I will emphasize the challenges and opportunities introduced by new communication structures, leading us to ubiquitous communication environments. Our view is that the distributed nature of these new structures lends itself naturally to community modeling, which can also help in handling the additional information overload that can be generated in ubiquitous environments.

Sharing Sensor Data in Intelligent Environments

Tim Schwartz, Dominik Heckmann, Jörg Baus

Saarland University and DFKI
66041, Saarbrücken, Germany
schwartz@cs.uni-sb.de, heckmann@dfki.de, baus@dfki.de

Abstract. Instrumented environments are enriched with sensors,senders and computing devices in order to support intelligent applications for the human-computer interaction. Because the sensors and senders in the environment can deliver a large amount of data, these so-called "intelligent environments" form ideal playgrounds to test the novel idea of *ubiquitous knowledge discovery*. In this paper we will describe the sensor architecture and the management software that is installed in the SUPIE (Saarland University Pervasive Intelligent Environment) for sharing basic sensor data. As an example application for ubiquitous knowledge discovery we describe our positioning system that collects data from different types of senders and that derives the user's position by fusing all data that can be helpful for this task.

1 Introduction and Architecture

Before we start representing our ideas about knowledge discovery in instrumented environments, we will briefly introduce the Saarland University Pervasive Instrumented Environment (SUPIE), in which computational resources are embedded as well as distributed. The environment's hardware and software architecture has been designed for the seamless integration of various services and applications supporting different tasks such as our shopping assistant [1] and the pedestrian navigation system [2]. The software architecture consists of four hierarchical layers, see figure 1 and [3], where assistance applications are considered as the top level.

The actual assistance applications of our intelligent environment use the knowledge representation and services of the lower layers to provide an intelligent user interface. The shopping assistant application provides product information and personalized advertisements to the user, this also includes the animated agent [4]. As the user interacts with real products on the shelf, their actions are recognized by a RFID reader and sent as events to the application. In response, the assistant proactively serves product information to the user, either on a tablet or a wall mounted plasma display. The user can also use their PDA for multi-modal interaction with the shopping assistance application, which entails the fusion of speech, handwriting, intra and extra gestures. Our navigation application also runs on a PDA and is based on the information provided by the location model and the positioning service. On the handheld, a graphical map and speech synthesis are provided. Besides the mobile device, the system utilizes nearby public displays to present arrows that indicate the direction to go.

All these aforementioned applications have access to a knowledge representation layer. This layer models some parts of the real world like an office, a shop, a museum

or an airport, see [5]. It represents persons, things and locations as well as times, events and their properties and features. A hierarchical symbolic location model represents places at different levels of granularity, like cities, buildings and rooms, and serves as a spatial index to the situational context. In order to generate localized presentations and navigational aid, the positions of the user, the buildings and the displays have to be known. Therefore the symbolic world model is supplemented by a geometric location model, containing the necessary data.

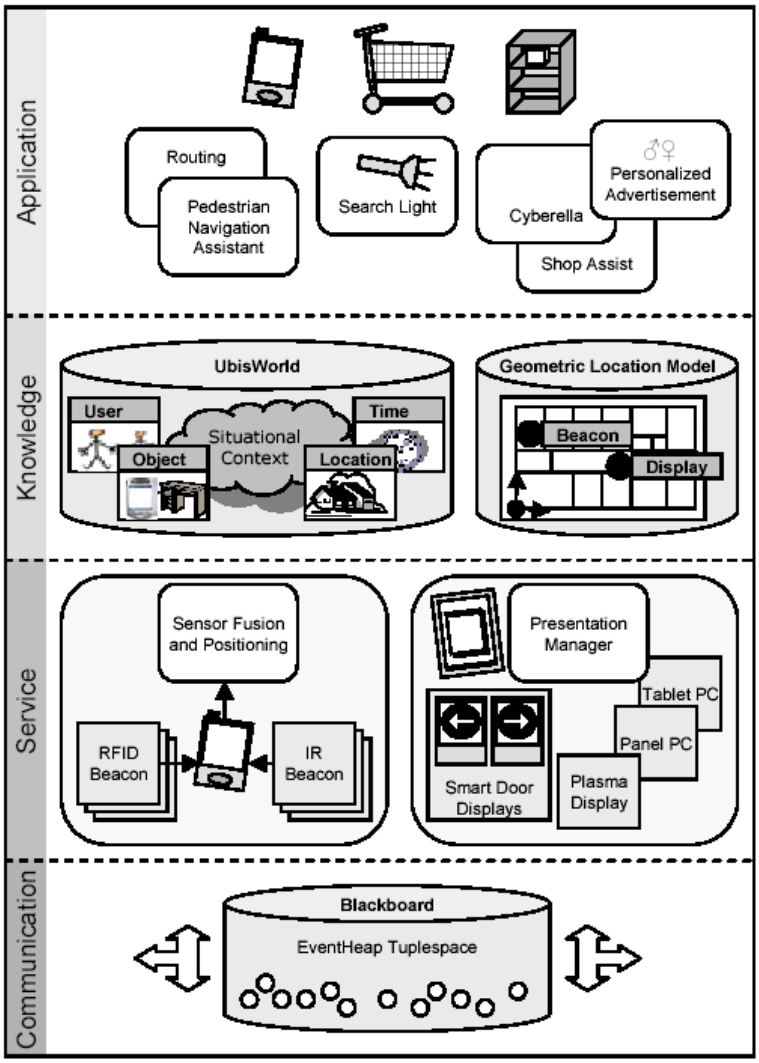


Fig. 1. The four-layered architecture of instrumented environments with communication, service, knowledge and application

Our software architecture's service layer provides multiple applications at the same time with information about a user's position in the environment and offers access to the public presentation service. It hides the technical complexity of these services behind a simple interface, which is based on blackboard events. For the positioning service we adopt a heterogeneous sensor approach, where a mobile terminal receives coordinates from infrared beacons as well as active RFID tags and estimates its own position using a dynamic Bayesian networks approach. The positioning service is presented in more detail later in the article as example for gathering sensor data. The presentation service provides a simple interface that allows applications to present Web content such as HTML and Flash on any display, which is connected to the presentation manager.

The communication and coordination within the intelligent environment is based on a commonly accessible tuplespace. Processes post their information to the space as tuples (collections of ordered type-value fields) or read them from the space in either a destructive or non-destructive manner. As the backbone of our communication layer we have chosen the EventHeap server and API, developed at Stanford University as a framework for their iRoom project (see [6]). Similar implementations are available from Sun [7] and IBM [8].

1.1 Our Notion of Ubiquity

The presented approach can be classified as *ubiquitous computing*. Mark Weiser's classification of a ubiquitous computing system is based on two fundamental attributes: namely *ubiquity* and *transparency*, see [9]. Ubiquity denotes that the interaction with the system is available wherever the user needs it. Transparency denotes that the system is non-intrusive and is integrated into the everyday environment. Further inspiring statements by Weiser are:

- Ubiquitous computing is fundamentally characterized by the connection of things in the world with computation.
- The main idea of ubiquitous computing: integrate computing into objects of daily life but hide its existence if possible.
- Things in the world can be actively supported by integrating computing devices or adding additional identification badges or labels. Things can be connected into Intelligent Environments via e.g. wireless lan.
- The real power of the concept comes not from any one of these devices; it emerges from the interaction of all of them. The hundreds of processors and displays are not a "user interface" like a mouse and windows, just a pleasant and effective "place" to get things done.

2 Sensors

There are different kinds of sensors integrated in our intelligent environment ranging from smart sensors boards which are able to sense, e.g. lighting conditions, temperature or physical interactions with objects, bio-sensors (see [10]), to passive/active RFID-Tags and Bluetooth-Dongles, which are used for our positioning services described in the next section.

Instead of describing all these sensors in a shallow fashion, we put the focus on the sensors for location and describe them in detail in the following.

The knowledge about the user's position is valuable information in a variety of applications. Because the Global Positioning System (GPS) that is normally used for such purposes is not available in buildings a different technology has to be used. One idea is to equip the user with a sender and to instrument the building with respective sensors. These sensors detect the signal the user's sender is dispatching and send this information to a centralized server that can then calculate the position of the user. This kind of localization is often called tracking or *exocentric localization* ([11]) because the user is sharing her position with the environment (the sender shouts "I'm here, I'm here!"). The opposite approach is to place the senders in the building and to let the user wear a device that is equipped with the respective sensor. The senders in the building send a signal ("You're near me, you're near me!"), the user's sensor detects this signal and the personal device of the user calculates the position. This is called positioning or *egocentric localization* because no information is send to the environment and thus the user's privacy is better protected.

Sensors/senders that are often used for the purpose of localization are: Ultrasound, infrared, and various radio based devices like WiFi, RFID and Bluetooth. These technologies differ in cost and reliability, where cheaper senders often provide less accuracy than higher priced ones. One of the problems of a localization system is that they normally use just one kind of sensor/sender technology and such a system works only in those buildings that provide the respective infrastructure. Our idea is to use different sensors/senders and to use a sensor fusion approach to calculate the users position. The advantage of this system is that it can work if there is just one of the sensor/sender technology available or it can derive a better position if more sensors/senders are available. Because such a system tries to always reach the highest possible accuracy (using all of the sensors/senders at the current position) we call it an Always Best Positioned (ABP) system (in analogy to Always Best Connected). We think that such an ABP system is a good example for ubiquitous knowledge discovery.

Our ABP system currently uses active RFID tags and infrared beacons as senders and an RFID reader card and the built-in infrared port of a PDA as sensors (see Figure 2). To fuse the sensory data we use an approach that is based on geo referenced Dynamic Bayesian Networks (see [11, 12] for a detailed description). The system is an egocentric system (as explained above) so the user can decide if she wants to reveal her position to the instrumented environment. She can do so by clicking a corresponding menu entry which causes the positioning engine to send the positioning information on the iROS Event Heap (see [13] for more information about the Event Heap).

Since every room in our lab has as least one computer in it and we have some public displays installed in exposed places, we use an inexpensive but also coarse exocentric (tracking) system that uses standard USB Bluetooth dongles to detect the presence of Bluetooth enabled phones. A small Java program is installed on the public displays and the office computers that scans for bluetooth devices. The result of each scan is also posted on the iROS Event heap as a list of all detected Bluetooth-addresses, the "friendly names" and of course the position of Bluetooth dongle itself. This can be

Fig. 2. iPAQ with attached RFID sensor and built-in infrared sensor (left). Active RFID tag and infrared beacon (right).

considered as raw sensor data because it does not contain the position of the detected devices. An application that wants to use this data has to infer the position out of these lists.

3 Sharing Sensor Data and Context Information

The presented event heap technology proves to be sufficient for sharing the sensor data within the instrumented environment itself. However, for sharing the sensor data and further inferred user model and context information with external systems and applications like the SPECTER system, see [14], new tasks of sharing and privacy handling have to be solved.

In our approach, we link the SUPIE event heap with the UbisWorld[1] situational statement service, see [14]. The concept of sharing with external systems is split up

[1] UbisWorld is based on the new concept of ubiquitous user modeling which means that networked systems constantly track the users behavior at many times and in many ways. See `http://www.ubisworld.org`

within UbisWorld into exchanging and integrating statements about sensor data and context information. The former is realized by a user model and special context server (www.u2m.org) that provides a service-based architecture for distributed storage and retrieval of statements about users and situations.

We developed the RDF-based user model and situation exchange language UserML to enable decentralized systems to communicate over user models as well as situational and contextual factors. The idea is to spread the information among all adaptive systems, either with a mobile device or via ubiquitous networks. UserML statements can be arranged and stored in distributed repositories in XML, RDF or SQL. Each mobile and stationary device has an own repository of situational statements, either local or global, dependent on the network accessability. A mobile device can perfectly be integrated via wireless lan or bluetooth into the intelligent environment, while a stationary device could be isolated without network access. The different applications or agents produce or use UserML statements to represent the user model information. UserML forms the syntactic description in the knowledge exchange process. Each concept like the user model auxiliary *hasProperty* and the user model dimension *timePressure* points to a semantical definition of this concept which is either defined in the general user model ontology GUMO, the UbisWorld ontology, which is specialized for ubiquitous computing, or the general SUMO/MILO ontology. More about these ontologies and the used protocols can be found in [15].

Figure 3 shows the input and output information flows *add*, *request* and *report* of the SITUATIONSERVICE. They are denoted as (yellow) arrows. The numbers in the (or-

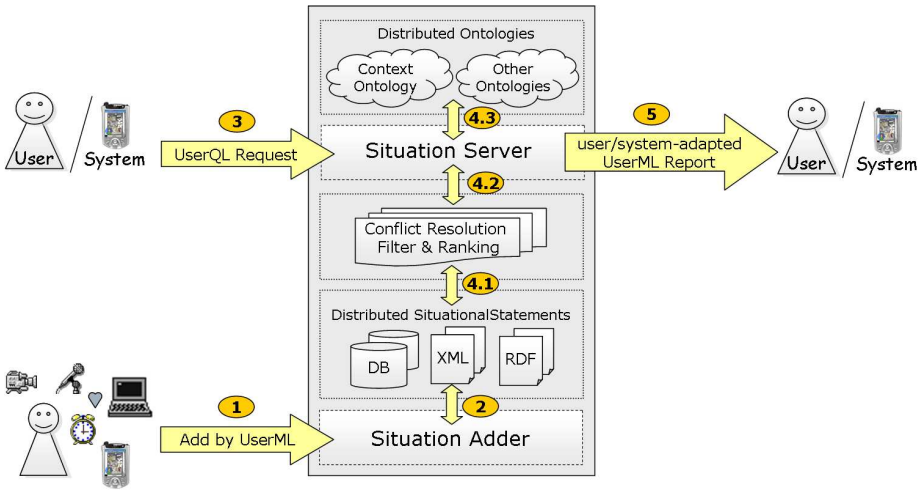


Fig. 3. General procedural view to the SITUATIONSERVICE

ange) ovals present the procedural order. Number (1) visualizes the sensors, users and systems that add statements via `UserML`. The statements are sent to the so called *Sit-*

uation Adder, a parser that preprocess the incoming data and distributes them to the different repositories, as indicated by number (2). If now a request is sent to the *Situation Server* via `UserQL` from a user or a system, see number (3), the repositories are selected from which the statements are retrieved as shown at number (4.1). Then conflict resolution strategies are applied, see number (4.2), and the semantic interpretation as indicated by number (4.3). Finally, see number (5), the adapted output is formatted and sent via HTTP in form of an `UserML` report back to the requesting user or system.

The integration of statements is achieved with an accretion model together with a multilevel conflict resolution method [5], which also solves the problem of contradictory information. What statements can be retrieved and how they are integrated depends on several layers of metadata attached to the statements by means of reification. From the outermost to the innermost layer, these are: administration, privacy, explanation, and situation. They establish a sequence of access constraints which have to be met in order to obtain the reified statement. The privacy layer in this sequence is of special interest. It implements the following privacy attributes: key, owner, access, purpose, and retention. The UbisWorld service checks these attributes in order to deliver as much information as possible without violating the users preferences. Combined with the other layers, complex situational access constraints can be established.

4 Summary

We have described the foundation for sharing sensor data within the Saarland University Pervasive Instrumented Environment by event heap technology, while we have directed the focus on location sensors and the diversity within the overall architecture. As an example for ubiquitous knowledge discovery we described how to fuse the data from different sender/sensor-technology to derive the position of a user. This example also shows a method for integrating privacy issues into the aspect of sharing sensor data of instrumented environments with external systems.

Acknowledgments

This research is being supported by the German Science Foundation DFG (Deutsche Forschungsgemeinschaft) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Project EM 4, BAIR, and its Transfer Unit on Cognitive Technologies for Real-Life Applications, TFB 53, Project TB 2, RENA.

References

1. Wasinger, R., Schneider, M., Baus, J., Krüger, A.: Multimodal Interaction with an Instrumented Shelf. In: Proceeding of Artificial Intelligence in Mobile Systems 2004 (AIMS 2004). (2004) pages 36–43
2. Krüger, A., Butz, A., Müller, C., Wasinger, R., Steinberg, K., Dirschl, A.: The Connected User Interface: Realizing a Personal Situated Navigation Service. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI 2004), ACM Press (2004) 161–168

3. Stahl, C., Baus, J., Brandherm, B., Schmitz, M., Schwartz, T.: Navigational- and shopping assistance on the basis of user interactions in intelligent environments. In: Proceedings of the IEE International Workshop on Intelligent Environments (IE), University of Essex, Colchester, UK (2005)
4. Kruppa, M., Spassova, L., Schmitz: The Virtual Room Inhabitant. In: Proceedings of the 2nd Workshop on Multi-User and Ubiquitous User Interfaces (MU3I), San Diego (CA). (2005)
5. Heckmann, D. LNAI3946. In: Situation Modeling and Smart Context Retrieval with Semantic Web Technology and Conflict Resolution. Springer-Verlag, Berlin Heidelberg (2006) 34–47
6. Fox, A., Johanson, B., Hanrahan, P., Winograd, T.: Integrating information appliances into an interactive workspace. IEEE Computer Graphics and Applications **20** (2000) 54–65
7. Freeman, E., Hupfer, S., Arnold, K.: JavaSpaces Principles, Patterns and Practice. Addison Wesley (1999)
8. Wyckoff, P.: Tspaces. IBM Systems Journal (1998)
9. Weiser, M.: Some computer science issues in ubiquitous computing. Communications on the ACM **36** (1993) 75–84
10. Brandherm, B., Schultheis, H., von Wilamowitz-Moellendorff, M., Schwartz, T., Schmitz, M.: Using physiological signals in a user-adaptive personal assistant. In: Proceedings of the 11th International Conference on Human-Computer Interaction (HCII-2005), Las Vegas, Nevada, USA (2005)
11. Schwartz, T., Brandherm, B., Heckman, D.: Calculation of the User-Direction in an Always Best Positioned Mobile Localization System . In: Artificial Intelligence in Mobile Systems (AIMS) 2005, in adjunction with MobileHCI 2005, Salzburg, Vienna. (2005)
12. Brandherm, B., Schwartz, T.: Geo Referenced Dynamic Bayesian Networks for User Positioning on Mobile Systems. In Strang, T., Linnhoff-Popien, C., eds.: Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, Munich, Germany, Springer-Verlag Berlin Heidelberg (2005) 223–234
13. Borchers, J., Ringel, M., Tyler, J., Fox, A.: Stanford interactive workspaces: A framework for physical and graphical user interface prototyping. In: IEEE Personal Communications Special Issue on Smart Homes, June 2002. (2002)
14. Kröner, A., Heckmann, D., Wahlster, W.: SPECTER: Building, exploiting and sharing augmented memories. In: Workshop on Knowledge Sharing for Everyday Life (KSEL06), Kyoto, Japan (2006) 1–8
15. Heckmann, D.: Ubiquitous User Modeling. PhD thesis, Department of Computer Science, Saarland University, Germany (2005)

Estimation of global term weights for distributed and ubiquitous IR

Hans Friedrich Witschel

witschel@informatik.uni-leipzig.de

Abstract. This paper reports on information retrieval experiments aimed at application in ubiquitous or P2P environments. The main question to be investigated is whether global term statistics such as IDF – which normally require a global view on the document collection – can be replaced with estimates obtained from a representative (i.e. well-balanced) reference corpus without losing too much effectiveness. Experiments with the British National Corpus (as a reference corpus) and two different IR test collections show that this is indeed possible. More interestingly still, lists of estimates can be compressed to a great extent without degrading performance, indicating that robust information retrieval is possible with very little knowledge of term characteristics, namely just an extended list of stop words. This makes it possible to distribute compressed term lists onto mobile devices without taking up too much bandwidth or storage capacity.

1 Introduction

As the importance of ubiquitous computing continues to grow, people will start to carry more and more personal information, but also information of general interest – or of interest within a certain community – with them on mobile devices. Besides ubiquity of computing, there is another important trend towards decentralised search, using e.g. P2P information retrieval (P2PIR) techniques. Integrating P2PIR or other distributed search techniques into ubiquitous environments means enabling people to share their data and search other people's data without even noticing. The limited resources available in most ubiquitous environments create new challenges for information retrieval algorithms. This paper discusses an approach to reduce storage and bandwith consumption of distributed IR by using compressed global information about index terms estimated from a reference corpus.

Most information retrieval weighting schemes are composed of a local and a global component, the local one measuring the extent to which a term is representative of a document's content (e.g. term frequency = TF) and the global one measuring a term's informative content in general (e.g. inverse document frequency = IDF). Estimating global term weights almost always requires access to the complete document collection which is often not feasible.

There are various reasons why global information may be difficult to obtain: often contents of a database change dynamically and quite quickly so that costly computation of global collection statistics cannot be performed at the same rate at which new items arrive. Sometimes – e.g. in the case of adaptive filtering or on-line new event detection – not even a part of the collection is known in advance.

Another area where full access to data collections is impossible is the rapidly growing field of distributed and peer-to-peer applications. In distributed environments, a global view on the data is often not guaranteed (e.g. because databases are not freely accessible), and more often not desirable: one of the advantages of P2P technologies which is cited very often is the fact that no central data repository needs to be maintained. This helps to reduce costs and to avoid the risks of failure. In these circumstances, it would be desirable to be able to estimate global term characteristics in advance, i.e. independently of the collection used for retrieval and then use these estimates on all sorts of collections. The collection used to obtain – once and for all – these term weight estimates will be called *reference corpus* throughout this paper.

It should be noted that – apart from global term weights such as IDF – there are other global statistics such as average document length (*avdl*) which is used e.g. in the well-known BM25 weighting formula for document length normalisation. In this paper, we will not be concerned with global statistics which are not directly related to terms, partly because not all weighting schemes use them and partly because estimating them from reference corpora seems unfeasible. If one wishes to employ weighting schemes that incorporate such statistics, one could estimate them (e.g. *avdl*) from the local document collection of a peer or database node and hope that their distribution does not vary too much among the different nodes.

Another point of interest will be the question to what extent term lists with global estimates – once we have obtained them – can be compressed: in many applications in ubiquitous environments, it is desirable to keep the term list as small as possible since it has to be distributed among all peers or participants and most mobile devices have limited storage and bandwidth capacities.

All in all, this paper is concerned with providing lists of global term weights and their compression, i.e. the focus is on reducing resource consumption for single nodes, but also on reducing message overhead by avoiding all communication related to sampling of term information from network nodes. However, apart from these strategies, no other network-related optimisations (e.g. adaptation to network properties, replication, optimisation of distributed queries etc.) are used.

The remainder of this paper is organised as follows: in section 2, we review some related work, section 3 details the theoretical background needed for the estimation of term weights, section 4 presents experimental results and section 5 concludes.

2 Related Work

As indicated above, there are various fields of application for global term weight estimates, obtained independently of a specific collection. Early work on dynamic and quickly growing collections [20] (and later [5]) found that IDF does not need to be re-computed constantly when adding new documents. Chowdhury [5] found that a training set of 40-50% of the documents in a collection is normally sufficient for robust IDF estimation.

In the field of on-line new event detection and adaptive filtering, we are faced with the problem that the collection is not known in advance, but constantly growing. There are two basic techniques that have been used in this field: either use a reference corpus from which IDF estimates are derived [7,16] or increment IDF estimates when new documents arrive, or combinations of both strategies [23]. Zhai et al. [7] note that the size of the reference corpus seems to play a certain role, but no systematic evaluation is performed.

In distributed IR, one often replaces IDF with ICF (inverse collection frequency, [2]) or IPF (inverse peer frequency, [8]), i.e. instead of counting the number of *documents* in which a certain term appears, one uses the number of *databases or peers* that contain the term. This is used for resource (i.e. database) selection but also for merging results from different databases. Various result merging algorithms have also found their way into hierarchical P2P networks (e.g. [14]). Alternatively, one may acquire and merge document frequency information from each database, but often one finds that a sample is sufficient: [19] show that only a part of collection-wide information needs to be disseminated in distributed IR and that the extent to which this is necessary depends on the allocation of documents to databases: random assignment requires fewer information than allocation based on content.

Sampling document frequency (DF) statistics from databases or peers is also often used in P2P information retrieval. Some approaches use exhaustive statistics [10], but since this scales poorly, one often operates with just samples of peers from which statistics are obtained and merged (e.g. [18]). A reference corpus for IDF estimation has been used in [12].

All in all, the existing approaches have either conducted evaluation of estimation from a subset of the retrieval collection or – if they chose to use a reference corpus – have not evaluated the effects of this choice at all. This paper will try to add value to the current state of the art by closely examining the possibility of term weight estimation from a fixed reference corpus. If estimates from reference corpora yield good effectiveness, they are highly useful: it will not be necessary to do tedious and costly sampling of term lists of participants' shared content before we can start to do retrieval.

As mentioned above, we will also be interested in how and to what extent term lists with frequency estimates can be pruned in order to fit onto mobile devices with small storage capacities. There are various ways in which information retrieval data can be compressed: besides lossless compression techniques (cf. [22]) which use efficient data structures to compress data, there are lossy ones (like the one presented here) that prune entries from term or inverted lists.

Although it is common to remove stop words before indexing, most of the existing lossy compression techniques focus on inverted lists: for example, Carmel et al. [3] prune entries from the inverted list whenever a term's occurrence in a document suggests that the term is not a good indicator of the document's content (e.g. the term is too rare in the document).

The approach taken in this paper is different: instead of eliminating terms or term-document pairs that seem "unimportant", rare (i.e. informative) terms will be expunged from the term list. The idea is to make a coarser distinction w.r.t. to the informative content of terms, i.e. to treat all rare terms as equally informative. Thus, although inverted lists remain uncompressed, term lists can be considerably compressed. Compressing these lists is important in distributed environments because they have to be disseminated throughout the network.

3 Estimators

Let us treat the problem of estimating global term weights as one of statistical inference: given a sample of, say, general English language, which was generated according to some unknown probability distribution, we wish to make guesses about this distribution. This problem is well-studied in the field of language and speech processing (cf. e.g. [15]), more precisely in language modelling where one tries to predict the next word, given a sequence of predecessors. However, in our task, we will not care about the neighbouring words, but rather try to predict a word's probability regardless of its predecessors, i.e. estimate a *unigram language model*.

3.1 Term weighting

Once we have estimated a term's probability from our sample, how do we turn this estimate into a term *weight*? This depends on the weighting scheme which we are going to use, but generally terms with low probabilities will be treated as informative ones. Before we start the estimation, we should therefore have a closer look at IR weighting schemes in order to know what exactly we will need to estimate: some weighting schemes (like TF.IDF) employ *document frequency* of terms, whereas others make use of *collection frequencies*.

In the case of IDF, the *document frequency* DF_t of term t is turned into a global term weight via:

$$IDF_t = log\frac{N}{DF_t} = -log\frac{DF_t}{N} \tag{1}$$

where N is the number of documents in the collection. When looking at the right hand side of the equation, we can interpret $p_{doc}(t) = \frac{DF_t}{N}$ as a term's probability of being contained in an arbitrary document ($-\log p$ being a measure of an event's informative content or its *pointwise information*, often used in information theory, cf. [17]). This means that if we have a reasonable estimate $\hat{p}_{doc}(t)$

of $p_{doc}(t)$, we can calculate an IDF approximation as $-\log \hat{p}_{doc}(t)$.

An information retrieval model that makes use of collection instead of document frequencies is the language modelling approach: here, we estimate the probability that a query q was generated from a document d's language model:

$$p(q|d) = \prod_{t \in q} p(t|d)^{tf_{t,q}} \tag{2}$$

where $tf_{t,q}$ is the term's frequency in the user query.

Using a maximum likelihood estimate for $p(t|d)$, i.e. the term's relative frequency in the document, will assign zero probability to documents that happen not to contain one of the query terms. Hence, this probability is smoothed using the collection as a back-off, e.g. using Dirichlet priors ([24]):

$$p(t|d) = \frac{tf_{t,d} + \mu p_{coll}(t)}{|d| + \mu} \tag{3}$$

where $tf_{t,d}$ is the term's frequency in the document, μ is a free parameter and $p_{coll}(t)$ is the term's back-off probability, normally computed from the whole collection. Smoothing has been shown ([24]) to have an effect similar to IDF weighting. This is because documents that *do not* contain an informative term t will be punished because of a very small factor $p(t|d)$ in the above product.

In the experiments below, an alternative formulation of language model weighting was used, namely a Kullback-Leibler divergence-based approach [13] in combination with Dirichlet smoothing. Since no parameter tuning was to be performed, μ was set to the average document length (which is a frequent choice although it has been shown that larger values often perform better [24]).

Now, if we want to apply the formulae above, we need to have an estimate $\hat{p}_{coll}(t)$ which can either be calculated as a maximum likelihood estimate from the collection or estimated from our reference corpus. The experiments to be described below will examine the differences between these two possibilities.

3.2 Estimating unigram language models

We now turn to estimating the probabilities mentioned above from a sample of English language. For doing this, we will need a large collection of documents, representative of general English language usage. The idea is that terms which are common in general English usage will be common in many other fields. Terms not contained in a representative sample of English are likely to be highly specific and thus informative.

One can argue a long time about the right way to compile such a sample, but for now, we will just assume the British National Corpus (BNC) to be such a collection, i.e. let us trust for a moment in the BNC compilers' expertise in having supplied a representative sample. The BNC consists of 4054 documents and approx. 100 million running words.

The simplest way to estimate $\hat{p}_{doc}(t)$ or $\hat{p}_{coll}(t)$ from the sample would be to use maximum likelihood estimates, i.e. $\frac{DF_t}{N}$ and $\frac{CF_t}{|C|}$, respectively where DF_t and CF_t are the document and collection frequencies of term t, N is the number of documents and $|C|$ is the number of word tokens in the sample. However, even if our sample is very large, we will invariably run into situations where some term has not been observed in the sample. Hence, we need to reserve some probability mass for unseen events.

The Good-Turing estimator provides one solution to this problem. It reserves $P_0 = \frac{E(n_1)}{N}$ for unseen events and $P_r = \frac{r*}{N}$ to events seen r times with

$$r* = (r+1)\frac{E(n_{r+1})}{E(n_r)} \tag{4}$$

where N is the sample size (i.e. the number of tokens in the sample) and n_r is the number of terms seen exactly r times in the sample. Instead of using n_r directly in the formula (which is often 0 for high frequencies), Gale and Sampson [9] propose – using Zipf's law – to fit a line through the $r - n_r$-curve in the log domain, a technique which they called "Simple Good-Turing" estimation and which yields smoothed values $E(n_r)$. [1].

For estimating $\hat{p}_{doc}(t)$, we will take n_r to be the number of terms that occur in exactly r documents. For $\hat{p}_{coll}(t)$, n_r will be the number of terms with a collection frequency of r.

Good-Turing estimates are known to be very accurate. However, like most other estimators, they require to estimate the total number of unseen events. The probability mass P_0 is then divided equally among all these in order to obtain an estimate of the probability of an individual unseen term.

There is a vast amount of literature about the estimation of unseen classes, especially in biology where one is interested to know the total number of species (e.g. butterflies) after having caught a number of animals. Some estimators have been proposed for language modelling as well ([6,1]) but they often yield widely differing and often too small estimates. There are even researchers who claim that there is an infinite number of (English) word types [11]. For the purposes of the experiments to be described below, let us consider a lower bound for the number of unseen classes which can be derived using P_0 itself [4]: Let n be the true number of all classes and D be the number of classes observed in our sample. We then define the sample coverage CV to be the sum of the probabilities of the observed classes. That means that $CV = 1 - P_0$. If we assume for a moment that all classes are equally probable, i.e. each class occurs with probability $1/n$, then the total probability of all observed classes is $CV = D \cdot 1/n$. This yields the estimator $\hat{n} = \frac{D}{1-P_0}$. Of course, the assumption of equiprobable word types is wildly wrong (cf. Zipf's law!), but in the case of lower entropy (i.e. non-equiprobable classes) we expect to see fewer classes in the sample and thus $\hat{n}$ is a lower bound for n.

[1] In the experiments, the simple GT software available from G. Sampson's website was used to obtain the estimates

For the BNC, we find that there are $n_1 = 153,857$ terms with a frequency of 1. Therefore, $P_0 = \frac{E(n_1)}{N}$ is approx. 0.0016 (remember that $N = 100 \cdot 10^6$). We observe $D = 348,285$ different term types in the BNC and hence expect that there are at least another $\hat{n} = 348,843$ unseen term types in English language. The probability of an individual unseen term is therefore $4.59 \cdot 10^{-9}$ and its pseudo-frequency (obtained by multiplying with N) is 0.46. The pseudo-frequency of words seen once in the BNC is 0.60, that of terms seen twice 1.51 and so on. This illustrates how some probability mass has been taken away from the observed terms and given to unseen ones.

We will later see that exact estimates of very low-frequent items are of little interest, i.e. it suffices to have some coarse estimate for them as long as we are able to separate informative from non-informative words. In order to see this, the experiments were conducted both with (hopefully accurate) Good-Turing estimates and the well-known Laplace (or *add-one*) method which consists in adding one token of each unseen class to the sample, yielding estimates linear in the maximum likelihood estimates but non-zero for unseen events. Laplace estimates are known to be inaccurate for low-frequent terms (they reserve too much probability mass for them, cf. [15] chapter 6), but we will see that this does not affect retrieval performance.

4 Experiments

In this section, we will use Good-Turing and Laplace estimation to derive $\hat{p}_{doc}$ and $\hat{p}_{coll}$ estimates from the BNC and then use these to perform retrieval with TF.IDF and language model term weighting on two very different collections: on the one hand, we use the TREC collection (disks 4 and 5 minus the congressional records, using queries from TREC-7 and TREC-8). On the other hand, the small and very specific MEDLINE collection was used which consists of only 1,033 medical abstracts, judged for relevance w.r.t. 30 queries. Since many more *unjudged* abstracts can be obtained from the same source, we can enlarge this collection in order to obtain more robust DF or CF estimates, but do retrieval evaluation only on the small collection.

4.1 First results

Table 1 shows results for TREC-7 and TREC-8 using medium-sized TREC queries (i.e. title and description) and for MEDLINE.

	TREC-7				TREC-8			
Weighting	MAP	P5	P10	R	MAP	P5	P10	R
TFIDF	0.2075	0.504	0.460	0.532	0.2482	0.524	0.448	0.602
+ BNCGT	0.1952	0.512	0.452	0.510	0.2350	0.524	0.454	**0.525**
+ BNCLAP	0.1945	0.512	0.450	0.508	0.2348	0.524	0.458	**0.523**
LM	0.2163	0.532	0.482	0.530	0.2578	0.512	0.438	0.582
+ BNCGT	**0.2085**	0.520	0.472	0.530	**0.2399**	**0.468**	0.426	**0.556**
+ BNCLAP	**0.2085**	0.520	0.472	0.531	**0.2398**	**0.468**	0.424	**0.555**

	MEDLINE			
Weighting	MAP	P5	P10	R
TFIDF	0.5054	0.680	0.6033	0.884
+ BNCGT	**0.5378**	**0.767**	**0.673**	0.884
+ BNCLAP	**0.5349**	**0.773**	**0.667**	0.884
LM	0.4606	0.700	0.6033	0.884
+ BNCGT	**0.5224**	0.730	**0.657**	0.884
+ BNCLAP	**0.5214**	0.727	**0.657**	0.884

Table 1. Mean average precision (MAP), precision at 5 and 10 documents (P5 and P10) and recall after 1000 documents retrieved (R) of TF.IDF weighting and language models (LM) and their correspondents using BNC Good-Turing (GT) and Laplace (LAP) estimates on TREC-7, TREC-8 and MEDLINE. Statistically significant differences (using a two-tailed t-test at a 95% confidence level) between a scheme and its corresponding BNC estimation scheme are marked with **bold** font.

From these figures, we can see that:

- On TREC, BNC estimates generally lead to small losses in mean average precision (MAP). Differences are never significant for TF.IDF, but sometimes for language models, i.e. language models seem less robust against poor CF estimates. Looking at precision at 5 or 10 documents, BNC estimates often compare more favourably to the TREC estimates, especially with TF.IDF. This means that we can expect good early precision with BNC estimates which is very often sufficient for the user of the retrieval system. As far as recall is concerned, BNC estimates retrieve about the same amount of relevant documents for TREC-7 and MEDLINE, but do significantly worse for TREC-8.
- On MEDLINE, BNC estimates always yield better performance, both in terms of mean average precision and in terms of early precision (P5 and P10). Almost all differences are statistically significant. Figure 1 shows what happens if we enlarge the MEDLINE collection by an ever growing amount of abstracts from the same source and use this enlarged collection for estimation. At some point, the quality of MEDLINE estimates outperform the BNC estimates (3,000 abstracts for TF.IDF, 30,000 for LM). The differences in MAP between BNC and the large MEDLINE collection become significant from 16,000 abstracts upwards with TF.IDF, but never for language

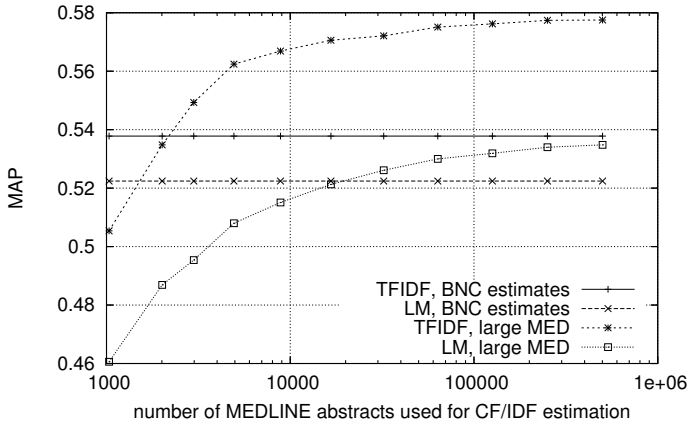


Fig. 1. Mean average precision as a function of the number of MEDLINE abstracts used for estimation.

models. Note that language models perform worse than TF.IDF probably because we have not tuned the μ parameter: MEDLINE queries are quite long and long queries normally need higher μ-values [24].
This exercise shows that for very small collections, it is profitable to use "outside" estimates, preferably from the same source, but if this is not available, also from a general-purpose collection like the BNC.

– There seems to be no difference between Good-Turing and Laplace estimates in terms of retrieval performance. A possible explanation could be (see below) that we do not need very exact estimates, rather only a rough distinction of which words are common and which are rare/informative.

4.2 Compression

For using BNC estimates, we need to supply a term list which contains all the terms we have seen in the BNC, together with their estimated collection or document frequency (or maybe directly their IDF) and of course an estimate assigned to words not contained in that list (unseen events). The smaller this list is, the better it is suited for ubiquitous environments. In this section, we will therefore examine two ways of compressing the term list:

Reducing the sample size If we use only a fraction of the BNC for estimation, we can expect the number of word types in the sample to be smaller than in the whole corpus and hence our list to be smaller, too. For example, we find that when using a random sample 30% the size of the whole BNC, we can compress the term list by approx. 50%. However, when we increase the sample, the size of the vocabulary does not increase linearly but slower. Figure 2 shows

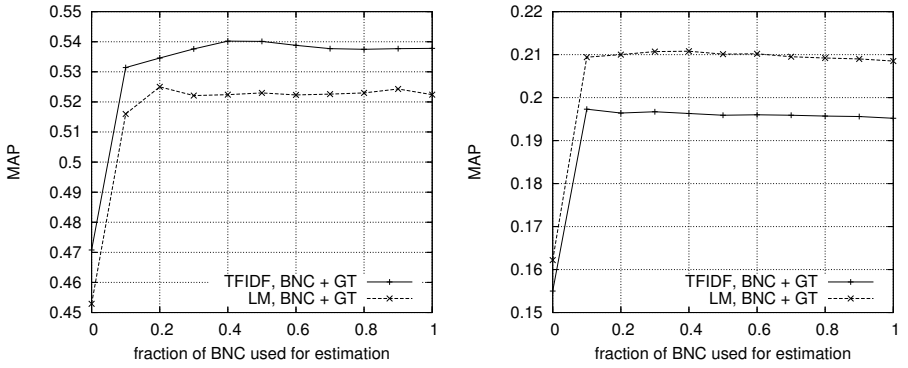


Fig. 2. Mean average precision as a function of BNC sample size for (a) MEDLINE and (b) TREC-7 (Good-Turing estimation).

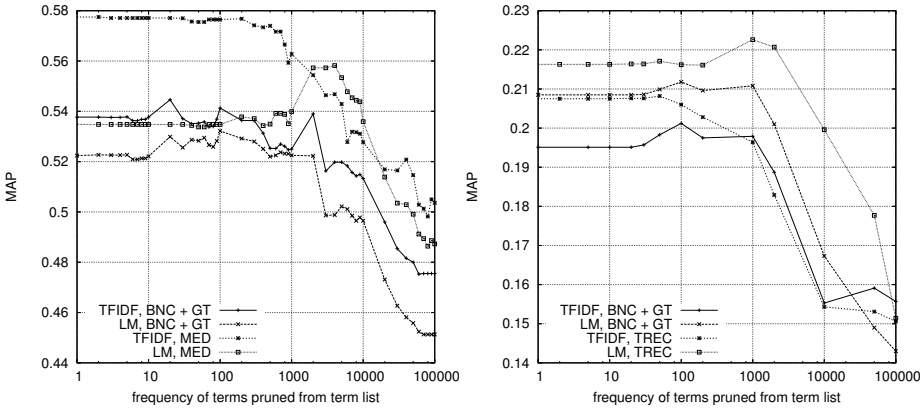
mean average precision as a function of sample size for MEDLINE and TREC respectively. Note that a sample size of zero indicates that uniform values were used for IDF and CF (namely 1.0) and a fraction of 1 indicates that the whole BNC was used (as before). As we can see, using 30% of the BNC is already close to the final value in all cases, so we can safely compress the list this way.

Pruning low-frequency words This method expunges all words which were seen only n times (for n=1,2,3...) from the term list (i.e. treats them as if they were unseens). The reduction in term list size we can achieve with this technique is notably higher than by reducing sample size due to the fact that e.g. – according to Zipf's law – approx. 50% of all terms have a frequency of 1. Figure 3 shows that performance only starts to become unstable or to drop dramatically when we prune terms with frequency ≥ 100 from the term list.

The figures show results for pruning BNC term lists, but also term lists taken from the collections themselves. In the latter case, pruning low-frequent terms means to treat them not as unseens but rather as if they had frequency 1.

MEDLINE results are quite unstable: there is even some increase in effectiveness for language models when pruning terms up to a frequency of 5,000. It will be a matter of future work to examine why this happens. One possible explanation is that many medium-frequency terms are highly informative and hence treating them as very rare gives good results.

Generally, however, this method is extremely robust: in all cases, we can get rid of all terms with frequency ≤ 100 without any notable change in effectiveness. Since both collections consist of about 100 mio. running words, this corresponds to a relative frequency threshold of 10^{-6}. It also corresponds to a reduction in term list size of more than 90% and leaves over approx. 40,000 terms.



(a) (b)

Fig. 3. Mean average precision as a function of the frequency of terms pruned from the term list (Good-Turing estimation from BNC or estimation from the collection itself) for (a) MEDLINE and (b) TREC-7. Note that a frequency of 0 indicates that nothing was pruned, i.e. the full term list was used.

This behaviour seems to indicate that exact term weights are not very important as long as we can separate non-content from content words. This coarse distinction is obviously possible by just knowing the 40,000 most frequent words and classifying them as uninformative (a set of words which will be called "stop words" in the rest of this paper).

4.3 Qualitative analysis

In order to find out what makes BNC estimates poor or good, let us look at queries which perform considerably better or worse when compared to estimates computed from the retrieval collection. We may then examine how query terms were weighted in each case.

Above, we have seen that BNC estimation tends to perform slightly worse on TREC. Although differences were sometimes quite large, they were rarely significant. This means that there is a large variation (standard deviation) across queries. In fact, when looking at MAP differences between queries and their BNC counterparts, we find that roughly 50% have a loss in effectiveness whereas the other half gains. As far as MEDLINE is concerned, most queries (approx. 70%) gain from BNC estimates when only the small collection is used for estimation and most queries (again around 70%) are degraded when compared to using the large MEDLINE. Generally, results on MEDLINE are more consistent, in terms of variation of results both among queries and among different weighting schemes.

However, in all cases, by examining poorly performing queries we can identify three main reasons of why BNC estimates are sometimes poorer than estimates from the retrieval collection:

1. Failure to identify collection-specific stop words: A good TREC example for this is query 428 "Do any countries other than the U.S. and China have a declining birth rate?". Here, "China" and "U.S." are very common in TREC, but weighted much too highly by the BNC estimates and hence we retrieve many documents about these two countries, but not about declining birth rates. Examples of MEDLINE-specific stop words which the BNC fails to identify include "disease" and "enzyme".
2. The reverse effect: sometimes words quite common in the BNC should have been weighted more highly because they are rare or informative in TREC or MEDLINE. A TREC example is the word "quilt", MEDLINE examples include "breast" or "head".
3. Missing discriminatory power: with IDF weighting, it often occurs that IDF estimates from the BNC are too flat, i.e. differences among them are too small in order to single out which term is – by far – the most important.

The reason why BNC estimates are quite superior to the small MEDLINE collection's estimates is simply the fact that words like "interest" or "include" are not recognised as being uninformative, obviously because they occur rarely in that small collection. This effect is, however, quickly remedied by adding some 1,000 abstracts to the collection (cf. fig. 1).

5 Conclusions and Future Work

In this paper, a series of experiments was performed which investigated the possibility of replacing global term weights from a retrieval collection with estimates from a reference corpus. Some interesting conclusions can be drawn from the results that were obtained:

- There seems to be a minimum amount of text needed in order to obtain good estimates of term weights. Although it is hard to quantify this exactly, the results indicate that around 1 to 10 million running words should generally be sufficient. Hence, very small collections are likely to benefit from estimates obtained from a larger corpus.
- The compression experiments indicate that what is actually needed as a term weight is obviously not very much: a list of approximately 40,000 high-frequent words ("stop words") produces results of the same quality as with full frequency information for all words. That means that all we need to know about words is a binary classification into informative and uninformative.
- Such a list of "stop words" can be estimated using a reference corpus without losing too much effectiveness. However, some collection-specific peculiarities (e.g. medical stopwords like "disease" in MEDLINE) which the reference corpus is unable to predict lead to a small loss in performance which we will

have to accept if we use reference corpora. These differences are sometimes, but not always significant.

All in all, using reference corpora in distributed and ubiquitous environments seems a promising idea: it yields better estimates than using small samples of a collection and it is much cheaper in terms of communication cost than obtaining large samples from a large number of participants or databases.

Additionally, compressed term lists which only contain about 40,000 terms do not take up much disk space (especially if they are further compressed, e.g. using Bloom filters) and still guarantee good retrieval performance on mobile devices where disk space is scarce. Besides distributed or P2P information retrieval, these term lists can also be applied in tasks like on-line new event detection or adaptive filtering.

For the future, it will be interesting to do more experiments with other test collections of different sizes, both highly specialised ones and general-language ones, in order to see more clearly which effects are attributable to a collection's size and which have to do with its degree of specialisation. Another focus of future work will be experimentation in a distributed setting, comparable to that of [21], where a reference corpus was already used in a P2P simulation.

References

1. S. Boneh, A. Boneh, and R.J. Caron. Estimating the Prediction Function and the Number of Unseen Species in Sampling with Replacement. *Journal of the American Statistical Association*, 93(444):372–379, 1998.
2. J. Callan. Distributed Information Retrieval. In W.B. Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.
3. D. Carmel, D. Cohen, R. Fagin, E. Farchi, M. Herscovici, Y. S. Maarek, and A. Soffer. Static index pruning for information retrieval systems. In *Proc. of SIGIR '01*, pages 43–50, 2001.
4. A. Chao and S.M. Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87:210–217, 1992.
5. A. R. Chowdhury. *On the design of reliable efficient information systems.* PhD thesis, Illinois Institute of Technology, 2001.
6. B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.
7. C. Zhai et al. Threshold Calibration in CLARIT Adaptive Filtering. In *Proc. of TREC-7*, pages 96–103, 1998.
8. F. M. Cuenca-Acuna et al. PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. In *12th International Symposium on High Performance Distributed Computing (HPDC)*, 2003.
9. W. Gale and G. Sampson. Good-Turing frequency estimation without tears. *Journal of Quantitave Linguistics*, 2(3):217–37, 1995.
10. F. Klemm and K. Aberer. Aggregation of a Term Vocabulary for Peer-to-Peer Information Retrieval: a DHT Stress Test. In *Proc. of DBISP2P'05*, 2005.
11. A. Kornai. How many words are there? *Glottometrics*, 4:61–86, 2002.

12. A. Z. Kronfol. FASD: A Fault-tolerant, Adaptive, Scalable, Distributed Search Engine, 2002.
13. J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proc. of SIGIR '01*, pages 111–119, 2001.
14. J. Lu and J. Callan. Merging retrieval results in hierarchical peer-to-peer networks. In *Proc. of SIGIR '04*, pages 472–473, 2004.
15. C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
16. R. Papka and J. Allan. On-Line New Event Detection using Single Pass Clustering. Technical report, University of Massachusetts, 1998.
17. S. Robertson. Understanding inverse document frequency: on theoretical arguments. *Journal of Documentation*, 60(5):503–520, 2004.
18. C. Tang, Z. Xu, and M. Mahalingam. pSearch: Information Retrieval in Structured Overlays. *ACM SIGCOMM Computer Communication Review*, pages 89–94, 2003.
19. C. L. Viles and J. C. French. Dissemination of collection wide information in a distributed information retrieval system. In *Proc. of SIGIR '95*, pages 12–20, 1995.
20. C. L. Viles and J. C. French. On the update of term weights in dynamic information retrieval systems. In *CIKM '95: Proceedings of the fourth international conference on Information and knowledge management*, pages 167–174, 1995.
21. H. F. Witschel and T. Böhme. Evaluating profiling and query expansion methods for P2P information retrieval. In *P2PIR'05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, pages 1–8, 2005.
22. I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Francisco, 1999.
23. Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proc. of SIGIR '98*, pages 28–36, 1998.
24. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

Author index