Case-Based to Content-Based User Model Mediation

Shlomo Berkovsky[1], Ariel Gorfinkel[1], Tsvi Kuflik[1] and Larry Manevitz[1]

Abstract. Systems providing personalized services to users have a need to build and maintain a User Model (UM). However, at the onset of providing services, such a system has no prior knowledge about a user and it may benefit from information imported from external sources. Due to lack of standards in the representation of UMs, commercial competition, and privacy issues, distinct personalized service-providing systems build their own specific models and store their information in incompatible manners. Thus, although much data on a specific user might exist in other systems; it is typically unavailable for use in the initial phase of the given system. This work puts forward the design of a user model mediation idea. This is demonstrated in an initial implementation in a specific system (Museum Visitors' guide system) under the PIL project, where the user is modelled by a "bag of words" vector and the initial information is imported from a case-based modelled user (in an external trip planning system).

1 INTRODUCTION

In order to provide users with personalized services, systems build and maintain a UM for each user. In general, a UM may be comprised of various details of personal information such as the user's age, education, income, life style, interests, preferences, past interactions with the system, etc.

Different systems use various methods, and techniques from diverse research areas, such as information retrieval, artificial intelligence and behavioural sciences for the construction of a UM [4]. Furthermore, every system stores UMs according to its own representation and chooses only the specific parts of user data that are relevant for providing its personalized services. Thus, large portions of user-related data that are heterogeneous both in representation and in content are distributed over various systems.

The notion of general (i.e. application-independent) user modeling was initially proposed in [3]. Their system - 'General User Modeling System' (GUMS) allowed the developers of user-adaptive applications to define simple user stereotype hierarchies. GUMS determined the basic functionality of general user modeling systems: providing selected runtime personalized services that can be configured during development time.

Most of the general systems that were developed, classified the collected UM to one of the predefined stereotypes using different inference methods. For example, [7] allowed stereotypical assumptions about the user and users groups to be represented in a first-order predicate logic, so that inferences across different assumptions could be defined in a first-order modal logic.

In [5], the authors discussed the development of UserML, an XML-like knowledge representation language developed for the purposes of describing UMs from various application domains. When used as a uniform user modeling language across multiple systems, UserML has a potential to facilitate transfer of UMs in distributed environments and further composition of UMs, accumulated in different systems.

Currently, even though all the required data may be potentially available, systems usually can not assemble comprehensive UMs due to commercial competition, privacy issues, and representation heterogeneity. By using UM mediators that bridge over different approaches and representations, the heterogeneity problem can be solved. Systems will be able to continue using their own methods of UM representation, and yet be able to exchange relevant parts of UMs with other systems, and enrich their UMs.

2 UBIQUITOUS USER MODELING

Personalization systems reside on the Web, in personal devices, and virtually everywhere. Thus, whenever a new user is introduced to a system, it has the potential to gather data about that user from systems all around- this "all-around located" data is ubiquitous in that sense, and the creation of UMs from ubiquitous data is therefore named "ubiquitous user modeling".

A rather simplistic approach of providing personalization in a ubiquitous environment was suggested in [12]. This approach suggested building an application adaptation framework using a personal smart card. The smart card stored and processed a UM, thus partially solving the privacy and availability issues which are essential in decentralized ubiquitous environments. Compared to a solution, where the profiles are stored in a central remote server, the use of smart cards made the profiles available in any context, enhanced privacy and security by allowing the users to fully control their own profiles, and avoided communication delays. However, the smart card remained a single "point of failure", storing sensitive personal information that could be disclosed by a malicious attacker.

Generation of centralized UMs in a ubiquitous environment by composing partial UMs that are stored by different systems was presented by [8]. The paper represented a ubiquitous general UM stored on a central server as a composition of partial UMs, stored by various personalization applications. Every system only maintains the inference mechanism needed for extracting the needed UM data and updating the general model. Although the general model was composed in a distributed manner, it was stored in a central server, which is a single "point of failure" in this case as well.

In [10], the authors highlighted the significance of cross-system personalization that will allow UM data sharing across different systems in a user-centric way. This approach allows information transfer between different systems, and gives the users the ability to control their UMs. Cross-system personalization might be implemented through a central Unified User Context Model (UUCM) [11]. The paper detailed three main stages of a cross-system communication protocol:

- Negotiation – achieving an understanding on the type of information that is needed, i.e., agreeing on common ontology and vocabulary.
- Personalization – extracting data which is relevant to the activity and transferring it to the target system.
- Synchronization – replicating and updating of the stored user model upon completion of personalization tasks.

The involved systems communicated through the mechanism of "context passport" using a mediating architectural layer. UUCM is

[1] The University of Haifa, Haifa, Israel.

based on a shared ontology- each system extracts the required information from the user's passport, performs the required personalization activities, and finally updates the user's passport. To succeed in its mission, the UUCM should have the following two features: generality (to be usable in a variety of domains), and expressiveness (to be able to express a wide variety of facts about the users). The fact that the shared ontology should be developed a-priori is the main drawback of this approach. Furthermore, every system used the ontology, which made it inflexible and unsusceptible to frequent changes.

In [6], the authors introduced GUMO, publicly available General User Model Ontology, which facilitates uniform interpretation of distributed UMs. GUMO is represented through a modern semantic language (OWL [9]) and can be freely used by any personalization system. A common ontology simplifies the exchange of UM data between different systems, which makes it possible to overcome the inherent problem of syntactical and structural heterogeneity between systems. The main problem of GUMO is that it is based on a single central ontology, which prevents dynamicity and frequent changes. Moreover, an initial stage of engineering and construction of a comprehensive all-including ontology requires a vast effort.

To summarize, the above approaches might be insufficient in the dynamic environment of today's information world, since both the available information sources and the needs and interests of the users change frequently, whereas personalization services should keep proper functioning and represent high levels of accuracy. This raises an intriguing research question of developing a mechanism that can easily adapt to the dynamicity of the environment, and at the same time allows the systems to provide an accurate customization of personalized services.

3 USER MODEL MEDIATION

A UM mediator generates UMs on demand, using available users' data according to the specifications of a target system. This is done by querying and receiving partial UMs from various source systems, translating them to the context of the target system, and building an integral UM according to the target system's method of representation from them [1].

Such kind of UM mediator is dynamic in the sense that it is not bound by a specific representation. Any system that requires a UM for bootstrapping may receive it, regardless of the specific personalization technique it uses, and the frequency the UM representation changes, unlike centralized apporaches. Since there are not many techniques used for user modeling, it seems feasible to have a set of specific mediators to be activated in any given scenario.

Another point to be made is that the mediator does not save any data regarding users, thus unlike a personal smart card, user data is not easily breached.

Figure 1 illustrates the stages of the mediation process as described below:

1. A user is requesting a service from a system.
2. In order to provide a personalized service to that user the system requests a UM from the mediator.
3. The mediator identifies the system's application domain and UM representation technique.
4. The mediator extracts from the knowledge base (KB) a set of systems that may provide partial UMs related to the target system's domain.
5. The mediator queries these systems for their UMs of the specific user.
6. Systems that actually store relevant UMs, respond and send the appropriate UM to the mediator.
7. The mediator converts, integrates and assembles the partial UMs (using the KB) into a UM needed by the target application.
8. The generated domain-specific UM is sent to the target system, which is now capable of providing more accurate personalization.

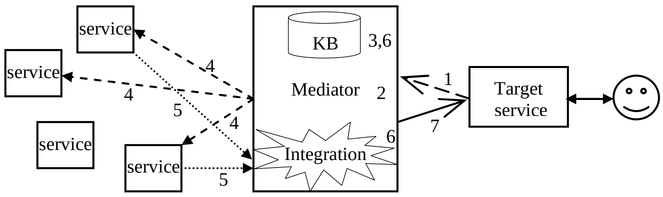

Figure 1. UM Mediator Architecture

4 CURRENT IMPLEMENTATION

A UM mediator system is being developed in the domains of tourism and cultural heritage. This system converts user information from an external trip planning system (Trip@dvice [13]) to a personalized museum visitor's guide system (in use under the PIL project at the Hecht museum at the University of Haifa [16]). In addition, it also serves as an intra-museum UM mediator, since the museum has different exhibitions, whose UM's representation is only partially related.

The UM representation in Trip@dvice is "cased-based". A case is a set of products that a user selected while planning a trip (such as attractions to visit, accommodations etc). In Trip@dvice, the personalization process determines a relevance score for each case item according to the user's current preferences and previous travel plans, and travel plans of users that have similar preferences. In order for a user to select preferred products when planning a trip, descriptions of various offered products, are presented to her/him. The selected products are recorded as a case representing the user's preferences.

Within PIL, each and every exhibit item has several different presentations; each presents the item from a different perspective and the goal of the personalization process is to recommend the user a presentation that best matches his/her interests and information needs. The entire set of terms from the presentations is called "bag of words", thus each presentation is represented as a weighted set of terms. The weights are obtained via the TF*IDF method known from Information retrieval, in which the weight of a term is proportional to the term's frequency in the presentation and to the scarcity of the term in other presentations [14]. A perspective's weighted vector is calculated in a similar way, by considering each perspective as the set of presentations which provide information about various items from that same point of view. The UMs in PIL are "content-based", where a user's preferences are represented by a weighted vector of terms that signify the visitor's preferences, and subsequently by a vector of cosine similarity of the user's UM to the perspectives' representation.

For the purposes of personalization, the presentations that provide information about an item, are being sorted according to their similarity to the visitor's interests (as determined by the UM) before being offered to him/her.

In order to generate a content-based UM from the case-based UM, terms are extracted by the mediator from descriptions of the cases' items in the Trip@dvice user's model. First, the mediator retrieves the cases from the case-based UM. Then, it obtains the free-text case items' descriptions. These descriptions are obtained in two ways: from the KB of tourism attractions, and also seeks out additional information about a case item from the Web. The "bag of words" representation of the UM as extracted from the case is compared to the "bag of words" representation of the exhibition using the well-known cosine similarity metric [14]. This allows deducing features' weights by considering the case item's relevance to the user (as stored in the Trip@dvice system).

Furthermore, every museum's exhibition has its own "bag of words", thus matching terms from one exhibition to another allows intra-museum UM mediation

For the purpose of user modeling, features extracted from acquired case description are converted to the features representing the exhibitions' presentations. First order conversion is aimed at matching the exact same feature (a one to one matching) - the weight of a feature in the target representation is given its weight in the other representation. There are cases in which first order conversion is not good enough, e.g., "ship" and "boat" should probably be considered as the same term even though they are not exactly the same term. Currently, only the first order conversion is implemented in our system.

Second order conversion is aimed at matching between semantically related features. One possibility of doing so is using WordNet [2] to match a feature to a synonym feature. In WordNet, English terms are organized into synonyms sets that describe the same semantic concept. This matching can be enhanced using various machine learning techniques, which weights the relationship between the feature and its synonym features. We are currently working on implementing the second order conversion.

The functional flow of our work is depicted below:

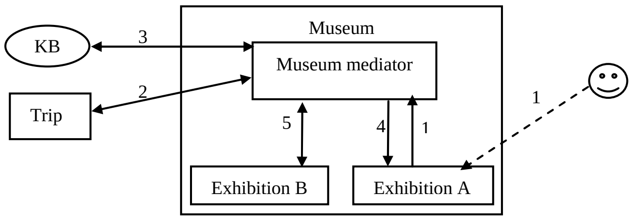

Figure 2. User model mediation in the museum

1. A visitor comes to the museum and enters exhibition *A*. The museum guide system requests an initial UM from the museum mediator.
2. The museum mediator verifies with the trip planning system that the user interacted with the system and retrieves the UM represented by the case items.
3. The mediator addresses an external knowledge-base (KB), and asks for relevant descriptions for the retrieved case items.
4. The mediator assembles a UM by converting features extracted from the case descriptions into UM features in the context of exhibition *A*.
5. When the user arrives to exhibition *B*, the mediator repeats the process and provides with a new conversion of features. The mediator treats the assembled UM of exhibition *A*, as an additional data source.

For example, let us assume that a user's trip planning case consists of the following locations: "Acco's port", "Acco's walls", and the "National maritime museum". These locations have archaeological and historic importance, and the maritime theme is common to them all. Visiting the national maritime museum which has many ship models, might suggest that the visitor is somewhat interested in ship-building methods, thus the technological perspective of the museum presentations, which explains the way that things were made, is probably a good perspective for the user. Let us also assume that the user has started his/her museum visit in the "Phoenicians' exhibition". The descriptions of these user's case locations are collected from the Trip@dvice system. Additional information describing these locations is collected from the Web, or from an external knowledge-base.

Using the Lucene search engine [15] the collected descriptions are indexed and the weight (TF*IDF) is calculated for their terms. The relevance score of a case item to the user is a factor that is considered during the process of determining terms' weights. Some of the collected terms also appear in the presentations of the exhibition, thus in the first order conversion, these terms' weight is conveyed to the generated UM. By using WordNet, every term's synonym set of terms can be found. The exhibition's terms are scanned for the synonym set of terms, which upon detection are added a weight, while taking into account several factors like the size of the term's synonyms set.

During the visit in the "Phoenicians' exhibition" the initial UM is modified to provide better personalization using implicit and explicit feedback from the user. Upon reaching the "Ancient ship exhibition", which has disparate items and consequently dissimilar presentations and only partially related "bag of words" perspectives' representation, the mediator will recalculate a UM relevant to the new exhibition. The mediator will use the UM resulted from the "Phoenicians' exhibition" visit, and generate a new UM for the "ancient ship" exhibition.

5 OPEN QUESTIONS AND FUTURE WORK

Future work will deal with experimentation and evaluation of the mediation in the cultural heritage setting described above. In addition to the obvious benefits of evaluation, the evaluation of the quality of the generated UM will allow the mediator to select between several possible UMs that were deduced from conflicting partial models.

An additional issue to be explored is the decisions the mediator should take regarding the quality of the information available from superficial sources in constructing the requested UM. The confidence level of the mediator in the information sources and in their relevance can be used as a factor while integrating partial UMs (sources that are more trusted will have a larger effect on the generated model). The confidence level can also be provided to the target system allowing it to determine to which extent the provided UM can be trusted and to decide its course of action. In our specific case, the relevance score of a case item might be a starting point to calculate the confidence level of the mediator in the information gathered from that item. The confidence level in the entire generated model might be an accumulation of the confidence levels in the gathered data.

We also intend on addressing third-order conversion, which is aimed at implicitly deducing relationships between the features. For example, if a person is interested in under-water archaeology, he/she might be somewhat interested in diving techniques. We consider using domain ontologies to describe these relationships.

Future work will also extend the current implementation to integrate UMs from several sources, and diverse domains. As stated above, several factors like the confidence level can play a role in the integration of UMs from several sources. For example, if a model of a user already exists in the target system, the mediator might consider it as a very reliable source in the process of generating a more comprehensive model. Other heuristics to solve the question of conflicting partial UMs, e.g., regarding the last time the UM was updated, will be explored.

One might also consider extending the bag of words UM representation in the museum to ontology, as a different form of content-base representation. This could prove to be another interesting case-study.

REFERENCES

[1] S. Berkovsky, T. Kuflik, F. Ricci, "Entertainment Personalization Mechanism through Cross-Domain User Modeling", in proceedings of the INTETAIN Conference, November-December 2005.

[2] C. Fellbaum, “WordNet - An Electronic Lexical Database”, The MIT Press Publishers, 1998.

[3] T. W. Finin, D. Drager, “A general user modeling system”, in Proceedings of the 6th Canadian Conference on Artificial Intelligence, Montreal, Canada, pp. 224-230, 1986.

[4] U. Hanani, B. Shapira, P. Shoval, “Information Filtering: Overview of Issues, Research and Systems”, in User Modeling and User Adapted Interactions, vol. 11 (3), pp. 203-259, 2001.

[5] D. Heckmann, A. Krueger, “A User Modeling Markup Language (UserML) for Ubiquitous Computing”, in Proceedings of the 9th International Conference on User Modeling, Pittsburgh, PA, pp. 393-397, 2003.

[6] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, M. von Wilamowitz-Moellendorff, “GUMO – the General User Model Ontology”, in Proc. of the International Conference on User Modeling, Edinburgh, UK, 2005.

[7] A. Kobsa, W. Pohl, “The BGP-MS user modeling system”, in User Modeling and User-Adapted Interaction, vol. 4(2), pp. 59-106, 1995.

[8] J. Kay, B. Kummerfeld, P. Lauder, “Managing private user models and shared personas”, in Proc. of Workshop on User Modeling for Ubiquitous Computing, Pittsburgh, PA, 2003.

[9] D. L. McGuinness, F. van Harmelen, “OWL Web Ontology Language Overview”, World Wide Web Consortium, 2004. http://www.w3.org/TR/owl-features/

[10] B. Mehta, C. Niederee, A. Stewart, “Towards Cross-system Personalization”, in Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, Las Vegas, NV, 2005.

[11] C. Niederee, A. Stewart, B. Mehta, M. Hemmje, “A Multi-Dimensional, Unified User Model for Cross-System Personalization”, in Proc. of Workshop on Environments for Personalized Information Access, Gallipoli, Italy, pp 34-54, 2004.

[12] O. Potonniee, “Ubiquitous personalization: A smart card based approach”, in proceedings of the 4th Gemplus Developer Conference, Singapore, 2002.

[13] F. Ricci, B. Arslan, N. Mirzadeh, A. Venturini, “ITR: a Case-Based Travel Advisory System”, in proceedings of the ECCBR Conference, Aberdeen, Scotland, 2002.

[14] G. Salton, M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983

[15] [Apache Lucene project]http://lucene.apache.org/

[16] [PIL project] http://www.cri.haifa.ac.il/peach.htm.