A User Modeling Server for Determination of Semantically-Enriched User Interests in Ubiquitous Environments

Fedor Bakalov[1], Birgitta König-Ries[1], Andreas Nauerz[2], and Martin Welsch[2]

[1]Friedrich-Schiller University of Jena, 07743 Jena, Germany,
{fedor.bakalov | birgitta.koenig-ries}@uni-jena.de
[2]IBM Research and Development, 71032 Böblingen, Germany,
{andreas.nauerz | martin.welsch}@de.ibm.com

Abstract. This paper describes a user modeling server capable of harvesting user interests in a nonintrusive manner based on the content accessed by users through mobile devices. The server enables mobile agents to log information about the content accessed by the user either in form of URL, text markup, or URI of an ontology concept referring to an encountered concept. Based on the information about occurrences of terms and semantic relations among them, the server clusters the terms into three groups, namely, interested, partially interested, and not interested.

1 Introduction

As mobile technology gets more advanced and mature, more and more interactions happen between human beings and mobile devices [2]. Users check the latest news through mobile phones, use these devices to purchase anything from soft-drinks to cinema tickets to subway tokens, access library collections via PDAs, museums use handhelds to guide their visitors through the exhibitions, and so on. All these interactions provide a huge potential for identifying user interests. Knowing about a user's interests is particularly important in mobile environments as here, it is essential that information is custom-tailored to the user and her current interests. However, due to the limited computational and storage capabilities of mobile devices, performing user modeling at the device side is very difficult and sometimes not feasible at all. To overcome this limitation, we propose a user modeling server for "outsourcing" the user interest modeling functionalities. This approach enables the seamless integration of user information obtained via different devices. Based on the information about the accessed resources or on terms describing these resources as well as semantic relations among the terms, the server determines to which degree the user is interested in a certain term. In the remainder of the paper, we will first provide details on the user model and will then discuss the different steps involved in building this model Section 2. We will then, in Section 3 briefly present the prototypical implementation of the user modeling server and provide initial test results. Finally, Section 4 concludes the paper and outlines the directions for our future work.

2 User Model

In the scope of this paper, we focus on modeling user interest, hence we describe the user **interest** model. We define user interest as a fact indicating that a given user is interested to a certain degree in a certain term. Here, the term is a reference to a concept denoting either a real world object, like company, geographic location, or person, or an abstract notion, like area of science or technology. The concepts themselves are stored in the underlying *domain model*, which is represented as an OWL ontology providing machine-processable semantics of the contained entities[1]. The degree of interest denotes the extent to which the user is interested in a given term. We distinguish three levels of interest identified by the following linguistic variables: *interested*, *partially interested*, and *not interested*. Following [4], we model the user interests as time dependent features. We assume that a user might be interested in a certain term only for a certain period of time. Thus, the interest user model is represented as a collection of tuples *(U, T, I, V)*, where U is the user ID, T is the URI of an instance from the domain model, I is the linguistic variable indicating the degree of interest, V is the time period of the interest validity.

Our approach to determining user interests involves the following activities. The server harvests the terms that users encounter through various devices, like their PDAs, mobile phones, portable computers but also their desktop machines. Afterwards, the collected terms are semantically enriched by referring to the corresponding instances in the underlying domain model. Finally, for every collected term, the server determines the degree of interest based on term frequency and the semantic relations among the terms in the domain model.

Harvesting Terms. The mobile devices need to be able to log occurrences of terms encountered by their user and submit them to the server. Depending on the capabilities of the mobile device but also on the user's privacy preferences, there are several possibilities to achieve this: The device might know (part of) the domain model and might be able to determine appropriate terms from this model itself. In this case, it can send the URIs of encountered terms to the server. Alternatively, the device leaves the determination of suitable terms to the server. In this case, it will submit the URL or markup of visited resources. As an example for the first case, consider a museum electronic guide. It might log evidence of a user stopping by a certain exhibit, which is annotated by a number of domain concepts, which can then be transferred to the server. In the second case, the server can process the whole resource and extract text fragments of certain types. For this purpose, we leverage at third-party Web service, Calais[2], a named-entity recognition service, which can receive an HTML or plain text document as an input and return an annotated document in RDF format. The user modeling server uses Calais for extraction of named entities, such as company, location, person, which are then used to update the domain

[1] For more information on the domain model see Section 4 in [1]
[2] http://www.opencalais.com/

and user models. For every extracted entity, the server checks if the domain model contains a matching instance and inserts one, if this is not yet the case. The concept this entity should be an instance of is determined via a mapping from the Calais types to the concepts of our domain model that we developed. Afterwards, the server makes an entry in the user log where it specifies the URI of the domain instance and the number of occurrences in the document. Quite obviously, there is a tradeoff between the two approaches: The first requires more storage space and processing power on the mobile device, but submits less information about the user to the server. With this approach, the server knows only about the terms the user is interested in, not the precise resources visited.

Partitioning Terms into Interest Groups. As described in Section 2, we distinguish three levels of interest degree and identify them by linguistic variables: *interested*, *partially interested*, and *not interested*. In order to determine the interest degree, we have identified two automatic update methods: log-based updates and inference-based updates. **Log-based updates** are performed by the server using the user log that stores occurrences of encountered terms. For every term in the user model, the server calculates the term frequency value as $TF_{i,j}=\frac{t_{i,j}}{\sum_k t_{k,j}}$ where t is the number of occurrences of $term_i$ for $user_j$, and the denominator is the total number of occurrences of all terms registered for $user_j$. **Inference-based updates** leverage the semantic relations among the instances in the domain model. This is when the server identifies new interests by propagating interest from the terms for which the user model already contains information about (e.g. it has been determined based on the term frequency). Term frequency and interest propagation values contribute to a term's cumulative weight, an aggregate value which is used to combine multiple updates. Based on the cumulative weight the server calculates the term's membership degree in the three fuzzy sets: *interested*, *partially interested*, and *not interested*. The fuzzy sets, in their turn, are determined by performing *Fuzzy C-Means Clustering* over the cumulative weights of all terms stored in the given user model. To determine whether $user_i$ is interested, partially interested, or not interested in $term_j$, the server calculates membership degrees of this term in the above mentioned fuzzy sets and assigns to the interest status of this term the linguistic variable of the fuzzy set to which the term has the highest membership.

3 Implementation and Test Results

The user modeling server has been prototypically implemented and consists of the following components (see Figure 1). The *user model* is implemented as a relational database. It stores information about user interests and logs containing the user browsing history and user model updates. The *domain model* is implemented as an RDF triple store deployed in the Sesame Framework[3]. Communication with the user model and domain model is enabled through the user

[3] http://www.openrdf.org/

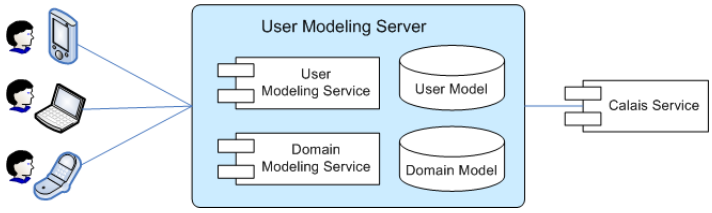

Fig. 1. System architecture

modeling and domain modeling services, respectively. The *user modeling service* provides such operations as add new log entry, perform updates, and query the model. Whereas the *domain modeling service* is used for manipulating and querying the content of the domain model. The prototype has been evaluated with three fictional users representing three countries, namely, Germany, Russia, and France. Every user used a specially designed interface to read news related to the country he or she is representing. The content of visited pages was processed with the *user modeling service*, which created three corresponding user models. The initial test results (Table 1) show that the generated models contain stable distribution of terms among the *interested* and *partially interested* clusters independent of the model's size.

User	News Topic	Visited Pages	Total Terms	Interests		
				Interested	Part. Interested	Not Interested
Klaus	Germany	418	669	4	55	610
Dmitry	Russia	72	173	5	14	154
Isabelle	France	20	56	4	23	29

Table 1. Test results

4 Conclusion and Future Work

In this paper, we have described a server that allows to build user interest models based on a user's access to resources with different devices. The server has been prototypically implemented and first experiments show that user interests are correctly modeled. Up to now, only the second approach of term harvesting has been implemented - a proof-of-concept implementation of the first approach is ongoing work. Also, we have used the user-interest model up to now only to personalize user's access to Portal pages [3]; we believe, however, that it is equally suitable for other uses in particular in ubiquitous environments.

References

1. Fedor Bakalov, Birgitta König-Ries, Andreas Nauerz, and Martin Welsch. Ontology-based multidimensional personalization modeling for the automatic generation of

mashups in next-generation portals. In *Proceedings of the 1st International Workshop on Ontologies in Interactive Systems held in conjunction with HCI 2008*, Liverpool, United Kingdom, September 2008.
2. Dominikus Heckmann. *Ubiquitous User Modeling: Volume 297 Dissertations in Artificial Intelligence - DISKI.* IOS Press, Inc., 2006.
3. Andreas Nauerz, Fedor Bakalov, Birgitta König-Ries, and Martin Welsch. Personalized recommendation of related content based on automatic metadata extraction. In *Proceedings of the 2008 conference of the Centre for Advanced Studies on Collaborative Research, Richmond Hill, Ontario, Canada*, page 5, 2008.
4. Andreas Schmidt. Ontology-based user context management: The challenges of dynamics and imperfection. In *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. Part I., ser. Lecture*, pages 995–1011. Springer, 2006.