Evaluation of User Model Effectiveness by Simulation

Shlomo Berkovsky, Ariel Gorfinkel, Tsvi Kuflik and Larry Manevitz

The University of Haifa
slavax@clsx.haifa.ac.il, agorfink@cs.haifa.ac.il, tsvikak@is.haifa.ac.il,
manevitz@cs.haifa.ac.il

Abstract. Accurate initialization of a User Model (UM) is important for every system that provides personalized services. However, there are systems where this initialization is critical yet no user data is available from prior interactions. For example, one in which both the repeat usage by a user is rare, and the total interaction on a single use is relatively limited.

Evaluating the effectiveness of a User Model and a particular instance of such a model is never an easy task, particularly through user studies. Moreover, evaluation typically focuses on the usability of an entire system rather than the performance of a specific UM instantiation. In this paper we propose evaluating the quality of UMs via simulation and comparison to a "gold standard". This standard is an approximation of the user's ideal model. We will demonstrate this through a case study of a museum's visitor guide system implemented in the Hecht Museum at the University of Haifa, Israel

Introduction

Evaluation of User Modeling Systems and techniques has always been a challenging task. Not only is the user modeling task uncertain and error prone, but the users themselves are not always able to assess the quality of their UM. Users provide their impression on using a system and they may provide a subjective evaluation of system performance. However, evaluation is typically done on an entire system. A wide range of techniques for this have been developed which focus on the system's users and their experience [6]. Likewise, trying to evaluate the effectiveness of a UM based on a system usability test is not straight forward. In the field of Information Retrieval (IR), sets of benchmark document collections (like TREC [11] and others) have been constructed to allow for the objective evaluation of IR systems. Such collections do not exist for user modeling in general, although there are a few data sets in the field of Recommender Systems that can be used for that purpose (Movielens [7] for example). In the remaining fields a carefully designed evaluation approach needs to be defined, such as the ones reported by [3]. Currently, it seems very difficult to create a standard set for user modeling evaluations given the diversity of application domains and tasks.

Museums are a popular location for research and development of applications of novel technologies. Since the appearance of mobile computers, there have been numerous projects focused on the development of mobile, adaptable museum visitor guide systems. An example of the wealth of research in the area is the survey of [1],

in which nine such systems were analyzed. However, even though museums are popular research sites, they present difficult challenges for the personalization of museum visitor guide systems. Typically, the system has no initial information on the visitors who will probably not return to that museum in the future. Despite the lack of initial information the system must be able to provide personalized service from the outset of the visit. To cope with this challenge, [2] suggested the idea of mediation (e.g. using UM data about the visitor, available from external systems), as a tool to bootstrap a UM in the museum. In this paper we address the problem of how to assess the UM's quality, assuming that external UM data is available.

In order to avoid the uncertainty and bias inherent in the evaluation of UMs by human users, we put forward evaluation by simulation. Within the simulation, different avatars, each with a stereotypical behavior, i.e. preset responses, are defined. The UM is being adapted continuously by the system according to the avatar's feedback based on its predefined stereotypical behavior. With an "infinite" number of these iterations the "gold standard" UM is generated. This serves as a reference to which the quality of an initial UM, and specifically the mediated UM can be compared. In other words, comparisons between the "gold standard" and a UM enable the assessment of that UM's quality. This in turn indicates the contribution of the bootstrapping techniques that generated it (such as the UM mediator mechanism).

User Modeling Mediation in the PIL Project

In the framework of the PIL project [5], a mobile, multimedia museum visitor guide system was developed. For every exhibit item, the system provides museum visitors with a list of available presentations sorted by inferred visitors' preferences. In order to personalize this ordering, the system uses a "content based" UM approach in which user's preferences are represented by a weighted vector of terms extracted from the presentations' text. The representation of both presentations and UMs is based on the classical IR vector space, where an *n*-dimensional term vector represents the text and terms are weighted using the TF*IDF weighting approach [9]. *n* is fixed and chosen by the number of terms in an overall dictionary. In other words, every UM has its own weighted vector representing these terms. The personalization is carried out through the ordering of available presentations. It is based on the similarity between the UM's vector and the presentations' vector.

In order to bootstrap the UM for the visitor, a UM mediator system is used. This system converts user information taken from a simulated trip planning system to the context of the museum visitor guide system [2]. While planning the trip, the user reads about the available products and selects those that seem of interest. The UM representation in the trip planning system is "case-based". In this system a case is a trip planned to the northern part of Israel, including a set of attractions that a user selected. In order to generate a content-based UM the mediator receives cases of the specific user from the case-based UM. Terms are extracted by the mediator from descriptions of cases items which are obtained from the knowledge base. Features extracted from acquired case descriptions are converted to the features representing the exhibitions' presentations. Figure 1 depicts the mediation process:

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** retrieves the user's model from the **Trip Planning** system (case items).
3. The **Museum Mediator** accesses an external **knowledge-base** (**KB**), and asks for relevant descriptions for the retrieved case items.
4. The **Museum 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 **Museum Mediator** can treat the assembled UM of **Exhibition A**, as an additional data source for UM data.

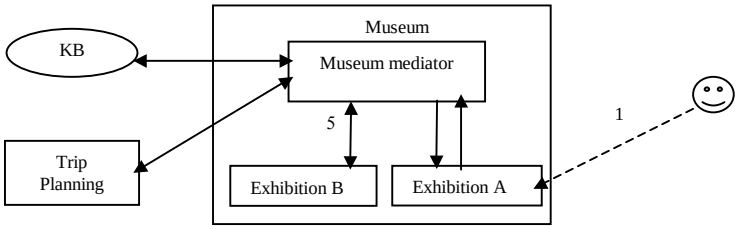

Fig. 1. User model mediation

UMs are adapted throughout the visit. The adaptation is done by asking the visitor to rate explicitly every delivered presentation on a scale of 1 to 5. This rating, together with the presentation's vector is used to update the vector of terms representing the user preferences, by applying the Rocchio algorithm, [8].

User Modeling Evaluation by Simulation

A major question regarding the above mediation mechanism is how we can evaluate the performance of the user modeling mediation component. In user modeling tasks there are only limited data sets allowing evaluation of the accuracy of personalized services. Such sets exist for Collaborative Recommender Systems like the Movielens database [7] and a few others. There are tasks such as information searching for which evaluating the contribution of a UM can be done by comparison of task performance, as described in [9]. However, the case of the museum visitor guide the is different. At the beginning of the visit the system has very little information about the visitors. Consequently, we use UM mediation for bootstrapping a UM of the museum's visitor guide system. The museum visitors have individual preferences and it is difficult to assess whether a visitor received the "best" ordering of presentations as well as assessing the accuracy of the mediation.

In order to address this challenge, simulation is used in the following fashion. Avatars which behave according to stereotypical user behaviors were defined. The stereotypical behavior is characterized by a pre-defined response to every presentation in the museum. For each avatar a "gold standard" UM, the avatar's ideal model, is generated. The "gold standard" is created by simulating a visit to the museum where the avatar sees each presentation many times. The avatar first chooses an exhibit item

randomly. There, it randomly chooses one presentation from those available. The avatar then gives feedback to the system based on its stereotypical behavior. This process is repeated 25,000 times.

Figure 2 presents the cosine similarity between a UM of an avatar and the avatar's previously obtained "gold standard" during a simulated visit. The graphs presented on Figure 2 illustrate the adaptation of UMs over time.

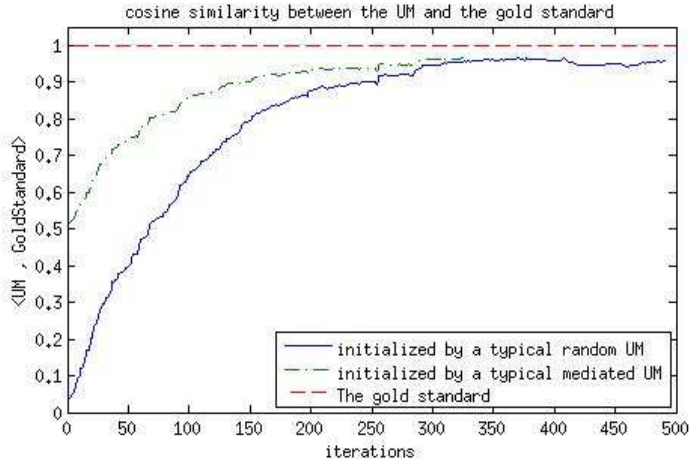

Fig.2. – Convergence graphs to a "gold-standard"

We see that different UMs of the same avatar converge to the same "gold standard," confirming the assumption that a "gold standard" exists and it is independent of the initialization. A comparison between the two initialization methods clearly shows that the UM bootstrapped by the mediator is initially closer to the "gold standard" than a randomly bootstrapped UM. The more similar a UM is to the "gold standard," the more accurate is the personalization received. Thus, the mediated UM enables better personalization at the early stages of a visit than a randomly initialized UM. Moreover, the mediated UM converges faster to the "gold standard". This shows that the mediated model also reduces the time until a UM suits the avatar relatively well. Museum visits are typically brief, thus in real life only a small number of presentations will be seen by the visitor. Therefore, a model that quickly becomes a suitable match to the visitor is desired. Since the avatars are an approximation to real visitors, these results indicate that having a mediated model can indeed help the museum guide achieve better personalization.

Conclusions and Future Work

An approach for evaluating user modeling mediation quality by simulation has been presented. Initial results show that, as expected, the mediated UM converges to the "gold standard" faster than a randomly bootstrapped UM. However, a systematic class of avatars that "covers" all user spaces should be developed.

In the near future we plan to experiment with other initialization methods, and evaluate the impact of these modifications on the accuracy of the mediated model. One of the potential modifications is the use of WordNet [4], which will allow semantic enhancements in the conversion of terms from one domain to another.

Acknowledgments

PIL was developed as part of the collaboration between ITC/irst and the University of Haifa and the experimentation is conducted at the Hecht museum at the University of Haifa. We thank the "Caesarea Edmond Benjamin de Rothschild Foundation Institute for Interdisciplinary Applications of Computer Science" (C.R.I.), the Neuro-computation Laboratory, and the HIACS Research center.

References

1. J. Baus, C. Kray, (2003), A survey of mobile guides, Workshop on Mobile Guides at: Mobile Human Computer Interaction '03
2. S. Berkovsky, A. Gorfinkel, T. Kuflik and L. Manevitz, (2006), “Case-Based to Content-Based User Model Mediation”, in proceedings of the Workshop on Ubiquitous User Modeling, in conjunction with the European Conference on Artificial Intelligence (ECAI), Riva del Garda, Italy, August 2006.
3. D. N. Chin and M. E. Crosby, (2002) Introduction to the Special Issue on Empirical Evaluation of User Models and User Modeling Systems, User Modeling and User-Adapted Interaction 12: 105-109.
4. C. Fellbaum (1998), “WordNet - An Electronic Lexical Database”, The MIT Press Publishers.
5. http://www.cri.haifa.ac.il/index.html?http://www.cri.haifa.ac.il/Connections/PIL/peach.php
6. Gena, C., & Weibelzahl, S. (2007). Usability engineering for the adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), “The adaptive web: Methods and strategies of web personalization”. Berlin: Springer (to appear).
7. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J., (1999), “An Algorithmic Framework for Performing Collaborative Filtering”, in proc. of SIGIR Conference.
8. Rocchio J.J., (1971), Performance Indices for Document Retrieval. In: G. Salton (ed):The SMART Retrieval System – Experiments in Automatic Documents Processing, Englewood, NJ, pp. 57-67.
9. Santos E., Nguyen H., Zhao Q., Pukinskis E. (2003). “Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application”. In proc. of the 9th International Conference of User Modeling, pp. 292-296.
10. G. Salton, M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
11. TREC (2007) http://trec.nist.gov/