User Profile Refinement using explicit User Interest Modeling

Gerald Stermsek, Mark Strembeck, Gustaf Neumann
Institute of Information Systems and New Media
Vienna University of Economics and BA Austria
{firstname.lastname}@wu-wien.ac.at

Abstract: In this paper, we present an approach to refine user profiles that were derived from Web server logs in an automated procedure. In most application scenarios, such automatically derived profiles can only deliver a preliminary result and require human interaction for further refinement. We describe the individual steps to enhance and refine derived user profiles which can be used for personalization purposes (e.g. information filtering). In particular, the user can choose to refine the profile manually or use supporting techniques, such as ontologies, that assist him in the refinement process. In addition to information included in automatically derived profiles, the user thus explicitly provides information to refine his profile.

1 Introduction

The constantly growing information supply in Internet-based information systems poses high demands on concepts and technologies to support users in filtering relevant information. Nevertheless, not every user may be willing to define his user profile from scratch as this can be a complex and time consuming task. Therefore, the first phase of our approach derives a preliminary user profile. This first phase is introduced in [SSN07]. In particular, the first phase is based on log-file analysis using descriptive statistics and network analysis methods. Note that we keep users informed about the data we gather via P3P policies (cf. [CLM02]), and we only derive profiles for users who agree with these policies (for details see [SSN07]). An automatically derived profile can then be seen as preliminary profile that covers a user's interests but needs to be further refined and elaborated. The user, thus, has to review the preliminary user profile to make sure that it represents his interests.

In this paper, we now focus on the refinement of a derived profile. In particular, we discuss an approach to adapt the preliminary user profile in order to define a more sophisticated user profile which better fits the user's information needs. The remainder of this paper is structured as follows. Section 2 gives on overview of our approach for user profile definition. In Section 2.1 the extension of profiles is discussed and Section 2.2 explains the refinement process. We briefly discuss related work in Section 3. Section 4 concludes the paper.

2 Approach Overview

In general, the user profiles that we derive from Web server logs (see [SSN07]) provide the following information:

- *Categories*: Categories represent user interests and are derived from meta-data provided along with the Web pages the user visited.

- *Structural Information*: If structural information is available we use to this information to derive relationships between categories.

A simple example of a derived user profile is shown in Figure 1. In this example, the automatically derived user profile indicates that the respective user is interested in soccer. A hierarchy structure of interest categories is not mandatory, though. If no structural information is available the user profile results in a simple list of categories. However, depending on the context of the Information Filtering system a hierarchy structure of interest categories may be used for weighting purposes in the information filtering process (see e.g. [SWM02]).

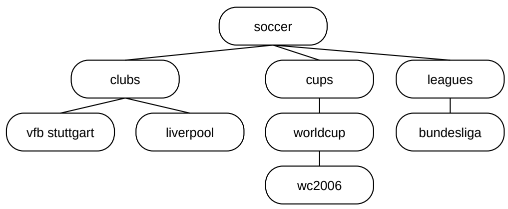

Figure 1: Example of a derived (preliminary) user profile

A high-level view of our approach is shown in Figure 2. The first two steps have already been elaborated in [SSN07] and, thus, are printed with dashed borders in Figure 2. In the following Sections we now describe the subsequent steps of our approach.

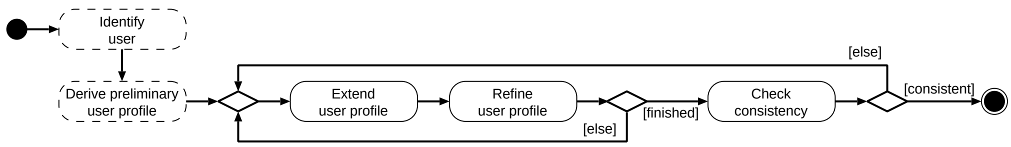

Figure 2: High level view of the user profiling approach

2.1 Extend user profile

The process to extend the user profile is depicted in Figure 3. At first the preliminary derived user profile has to be fetched and presented to the user. Afterwards, the user has four possibilities to alter the user profile:

- *Predefined categories*: With this option the user is offered a list of predefined categories which he can add to his user profile. This list is typically domain-dependent.
- *Manually*: Another option is to allow the user to add arbitrary user-defined categories to his user profile. This may not be suitable for all users and all domains but allows for a freely customizable user profile.
- *External source*: Additional user interests can also be imported from an external source. A user can, for example, import filtering keywords of an already configured news aggregator and add them as categories to his user profile.

- *Remove*: The user also has the possibility to remove interests from his user profile if they do not (longer) represent his user interests.

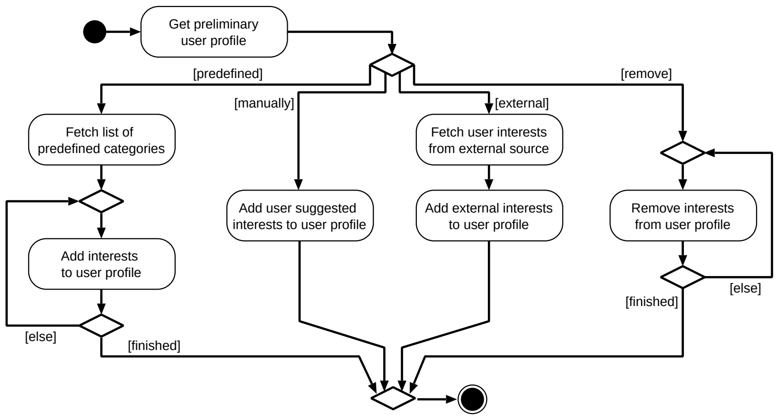

Figure 3: Sub-process to extend user profiles

The result of this process is then used as input for the next step of user profiling.

2.2 Refine user profile

In this step of the proposed user profiling refinement approach, the user can finalize his user profile. This can, again, be done manually or automatically. The corresponding process is depicted in Figure 4.

- *Manual refinement*: In this case the user has the possibility to refine the current user profile to fit his needs. To do this he can add or remove explicit relationships between categories. When adding an explicit relationship the user has to indicate the related terms and define them as related.
- *Automatical refinement*: If the user chooses not to define structural relationships manually he may use an ontology-assisted approach, for example. In this case, the user then has to select an appropriate domain ontology or, if not available, use a general purpose ontology (e.g. WordNet [Fe98]). This ontology then serves as a basis to derive term relationships. In [BH06] different measures of semantic relatedness are discussed which can be used to structure interest categories as graphs. If the user is not satisfied with the automatically derived term relationships he may further refine them manually, of course.

The two steps of adding interest categories and modifying the hierarchy structure can be iterated until the user is satisfied with the user profile. Finally a brief error check of the current user profile is conducted (cf. Figure 2). This includes spell checking and the indication of duplicates. As individual user profiles may be very specific we suggest to just indicate spelling errors and duplicates rather than correcting them automatically. The user then can decide on how to proceed on these issues.

Figure 5 depicts our example from Figure 1 after the refinement process. As can be seen the user removed the category `liverpool` which was derived from his log file entries. In our

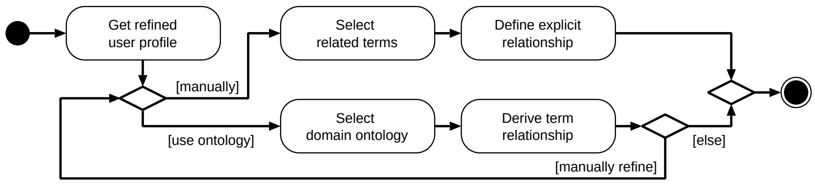

Figure 4: Sub-process to refine user profiles

example, this was just an accidental hit and the user has no long-term interest in Liverpool. Instead, he added a new category `mario gomez` and defined an explicit relationship between `mario gomez` and `vfb stuttgart`. The user also added another interest category named `wc2010` and defined a relationship with `worldcup`, expressing his interest in the forthcoming world cup.

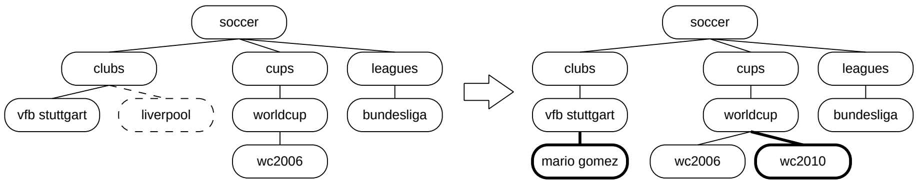

Figure 5: Example of a refined user profile

As mentioned above, defining a hierarchy structure is not mandadory but may revalue the user profile. A possibility to use the hierarchy structure of interest categories is to use categories from different levels to filter different information streams. An information system may, for example, use interest categories near the root category to select an appropriate RSS feed [RS06] for a respective user (e.g. sport news) and categories from the leaf nodes to filter information within this RSS feed (e.g. mario gomez, wc2006, wc2010).

3 Related work

Web usage mining (WUM) is the application of data mining techniques to large Web data repositories, some algorithms commonly used include association rule mining, sequential pattern generation, and clustering (see, e.g., [CMS99]). WUM produces aggregated results to better understand Web usage and improve the service provided to the customer (cf. [FSS00]). In contrast, our approach concentrates on data mining at the level of individual user data and produces non-aggregated results which can be used for the purpose of personalization, e.g. to form user profiles for information filtering.

Ontology-based user profiling [GCP03] uses ontologies to represent user interests via concept hierarchies. Compared to other concepts of user profile representation using ontologies means a semantic revaluation (cf. [GCP03]). However, general-purpose ontologies with a high number of concepts are often not appropriate for profiling a single user profile (cf. [GA05]). Ontologies often represent the shared knowledge of either a particular community or a group of users and therefore they may fail to capture an individual user's specific understanding of a domain [GA05].

In [HK04] Holland and Kießling present an approach for mining user preferences from user log data using strict partial order preferences. Eliciting complex user preferences from simple Web server logs may be diffucult. Holland and Kießling therefore suggest to use application server logs as they are a better source for user preferences. The refinement process presented in this paper can be applied to the approach of [HK04] as well.

4 Conclusion and Future Work

In this paper, we presented an approach to extend and refine automatically derived user profiles. Our approach benefits from the combination of automatic and manual user profiling. Automatically deriving a first version of a user profile relieves the user from the complex and time consuming task to define his user profile from scratch. This enables the user to better concentrate on the refinement process. We have already elaborated scripts to preprocess Web server log files and to automatically derive preliminary user profiles using the R software environment (see [SSN07]). The approach presented in this paper results in more elaborated user profiles which better fit the user's interests. The user can refine the profile manually or use supporting techniques, such as ontologies, that assist him in the refinement process. We are currently building a graphical tool that supports the presented refinement approach.

References

[BH06] Budanitsky, A. and Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. 32(1):13–47. 2006.

[CLM02] Cranor, L., Langheinrich, M., and Marchiori, M. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification, W3C Recommendation. April 2002.

[CMS99] Cooley, R., Mobasher, B., and Srivastava, J.: Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*. 1(1):5–32. 1999.

[Fe98] Fellbaum, C. (Hrsg.): *WordNet: An Electronic Lexical Database*. The MIT Press. Cambridge, MA, USA. 1998.

[FSS00] Fu, Y., Sandhu, K., and Shih, M.-Y.: A Generalization-Based Approach to Clustering of Web Usage Sessions. In: *WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*. S. 21–38. London, UK. 2000. Springer-Verlag.

[GA05] Godoy, D. and Amandi, A.: User profiling for web page filtering. *IEEE Internet Computing*. 9(4):56–64. 2005.

[GCP03] Gauch, S., Chaffee, J., and Pretschner, A.: Ontology-based personalized search and browsing. *Web Intelligence and Agent System*. 1(3-4):219–234. 2003.

[HK04] Holland, S. and Kießling, W.: User Preference Mining Techniques for Personalized Applications. *Wirtschaftsinformatik*. 46(6):439–445. 2004.

[RS06] RSS Advisory Board. RSS 2.0 Specification (2.0.8). August 2006. http://www.rssboard.org/rss-specification.

[SSN07] Stermsek, G., Strembeck, M., and Neumann, G.: A User Profile Derivation Approach based on Log-File Analysis. In: *Proc. of the International Conference on Information and Knowledge Engineering*. June 2007.

[SWM02] Shepherd, M., Watters, C., and Marath, A.: Adaptive user modeling for filtering electronic news. In: *Proc. of the 35th Annual Hawaii International Conference on System Sciences (HICSS)*. S. 1180–1188. 2002.