A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining

Honghua (Kathy) Dai, Bamshad Mobasher
{hdai, mobasher}@cs.depaul.edu
School of Computer Science, Telecommunication, and Information Systems
DePaul University, Chicago, Illinois, USA

Abstract

Personalization based on Web usage mining can enhance the effectiveness and scalability of collaborative filtering. However, without semantic knowledge about the underlying domain, such systems cannot recommend different types of complex objects based in their underlying properties and attributes. This paper provides an overview of approaches for incorporating semantic knowledge into Web usage mining and personalization processes. We present two general approaches to integrate semantic knowledge extracted from the content features of pages into the usage-based personalization process. Next, we present a general framework of integrating domain ontologies with Web Usage Mining and Personalization. In each case, we discuss how semantic knowledge is leveraged and represented in the preprocessing and pattern discovery phases, as well as how it is used to enhance usage-based personalization.

1 Introduction

One of the most widely used technologies for building personalization and recommendation systems is collaborative filtering (CF) [9]. Given a target user's record of activity or preferences, CF-based techniques, such as the k-Nearest-Neighbor (kNN) approach, compare that record with the historical records of other users in order to find the top k users who have similar tastes or interests. The mapping of a visitor record to its *neighborhood* could be based on similarity in ratings of items, access to similar content or pages, or purchase of similar items. The identified neighborhood is then used to recommend items not already accessed or purchased by the active user. The advantage of this approach over purely content-based approaches which rely on content similarity in item-to-item comparisons is that it can capture "pragmatic" relationships among items based on how their intended use or based on similar tastes of the users.

The CF-based techniques, however, suffer from some well-known limitations [16]. For the most part these limitations are related to the scalability and efficiency of the kNN approach which requires real-time computation in both the neighborhood formation and the recommendation phases. Web usage mining [17] techniques, that rely on offline pattern discovery from users' Web transactions, have been used effectively to improve the scalability of personalization systems based on traditional collaborative filtering [10, 13, 14].

However, the pure usage-based approach to personalization has an important drawback: the recommendation process relies on the existing user transaction data, thus items or pages added to a site recently cannot be recommenced. This is generally referred to as the "new item problem". A common approach to revolving this problem in collaborative filtering has been to integrate content characteristics of pages with the user ratings or judgments [5, 15]. Generally, in these approaches, keywords are extracted from the content on the Web site and are used to either index pages by content or classify pages into various content categories. In the context of personalization, this approach would allow the system to recommend pages to a user, not only based on a similar users, but also (or alternatively) based on the content similarity of these pages to the pages user has already visited.

Keyword-based approaches, however, are incapable of capturing more complex relationships among objects at a deeper semantic level based on the inherent properties associated with these objects. To be able

recommend different types of complex objects using their underlying properties and attributes, the system must be able to rely on the characterization of user segments and objects, not just based on keywords, but at a deeper semantic level using the domain ontologies for the objects.

This paper gives an overview approaches for incorporating semantic knowledge into Web usage mining and personalization processes. First, we discuss two general approaches to integrate semantic knowledge extracted from the content features of pages into the usage-based personalization process. Next, we present a general framework of integrating domain ontologies with Web Usage Mining and Web Personalization. In each case, we discuss how semantic knowledge is leveraged and represented in the preprocessing and pattern discovery phases, as well as how it is used to enhance usage-based personalization.

2 Elements of Usage-Based Personalization

Generally speaking, usage-based Web personalization systems involve 3 phases: data preparation and transformation, pattern discovery, and recommendation. Of these, the latter is a real-time component, while the other two phases are performed offline.

For a detailed discussion of preprocessing issues related to Web usage mining see [3]. Usage preprocessing results in a set of n pageviews, $P = \{p_1, p_2, \cdots, p_n\}$, and a set of m user transactions, $T = \{t_1, t_2, \cdots, t_m\}$, where each $t_i \in T$ is a subset of P. *Pageviews* are semantically meaningful entities to which mining tasks are applied (such as pages or products). Conceptually, we view each transaction t as an l-length sequence of ordered pairs:

$$t = \left\langle (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), \cdots, (p_l^t, w(p_l^t)) \right\rangle,$$

where each $p_i^t = p_j$ for some $j \in \{1, \cdots, n\}$, and $w(p_i^t)$ is the weight associated with pageview p_i^t in the transaction t representing its significance (usually, but not exclusively, based on time duration).

For many data mining tasks, such as clustering and association rule discovery, as well as for collaborative filtering based on the kNN technique, we can represent each user transaction as a vector over the n-dimensional space of pageviews. Given the transaction t above, the transaction vector $\overrightarrow{t}$ is given by:

$$\overrightarrow{t} = \left\langle w_{p_1}^t, w_{p_2}^t, \cdots, w_{p_n}^t \right\rangle,$$

where each $w_{p_j}^t = w(p_i^t)$, for some $i \in \{1, \cdots, n\}$, in case p_j appears in the transaction t, and $w_{p_j}^t = 0$, otherwise. Thus, conceptually, the set of all user transactions can be viewed as an $m \times n$ transaction-pageview matrix, denoted by TP.

Given a set of transactions as described above, a variety of unsupervised knowledge discovery techniques can be applied to obtain patterns. These techniques such as clustering of transactions (or sessions) can lead to the discovery of important user or visitor segments. Other techniques such as item (e.g., pageview) clustering and association or sequential pattern discovery can be used to find important relationships among items based on the navigational patterns of users in the site. In each case, the discovered patterns can be used in conjunction with the active user session to provide personalized content. This task is performed by a recommendation engine.

3 Integrating Web Usage Mining and Personalization with Content Features

3.1 Extracting Semantic Features from Site Content

One direct source of semantic knowledge that can be integrated into the mining and personalization processes are content features associated with items on a Web site. These features include keywords, phrases, category names, or other textual content embedded as meta-information. Content preprocessing involves the extraction of relevant features from text and meta-data. For features extracted from meta-data, feature weights are usually provided as part of the domain knowledge specified by the analyst. For features extracted from text

weights can be derived automatically, for example as a function of the term frequency and inverse document frequency (tf.idf) which is commonly used in information retrieval.

Further preprocessing on content features can be performed by applying text mining techniques. This would provide the ability to filter the input to, or the output from, usage mining algorithms. For example, classification content features based on a concept hierarchy can be used to limit the discovered usage patterns to those containing pageviews about a certain subject or class of products. Similarly, performing clustering or association rule mining on the feature space can lead to composite features representing concept categories.

Each pageview p can then be represented as a k-dimensional feature vector, where k is the total number of extracted features (or composite features) from the site in a global dictionary. This vector is given by: $p = \{\langle fw(p, f_1), fw(p, f_2), \ldots, fw(p, f_k)\rangle\}$ where $fw(p, f_j)$, is the weight of the jth feature in pageview $p \in P$, for $1 \leq j \leq k$. For the whole collection of pageviews in the site, we have the $n \times k$ pageview-feature matrix $PF = \{p_1, p_2, \ldots, p_n\}$.

3.2 Content-Enhanced Personalization

There are at least two basic choices as to when content features can be integrated into the usage-based personalization process: pre-mining integration or post-mining integration.

The pre-mining integration involves the transformation of user transactions, as described earlier, into "content-enhanced" transactions containing the semantic features of the underlying pageviews. While, in practice, there are several ways to accomplish this transformation, the most direct approach involves mapping each pageview in a transaction to one or more content features. The range of this mapping can be the full feature space, or feature sets (composite features) which in turn may represent concepts or concept categories. Conceptually, the transformation can be viewed as the multiplication of the transaction-pageview matrix TP with the pageview-feature matrix PF. The result is a new matrix $TF = \{t'_1, t'_2, \ldots, t'_m\}$, where each t'_i is a k-dimensional vector over the feature space. Thus, a user transaction can be represented as a content feature vector, reflecting that user's interests in particular concepts or topics.

Various data mining tasks can now be performed on the content-enhanced transaction data. For instance, if we apply association rule mining on such data, then we can get a group of association rules on content features. As an example, consider a site containing information about movies. This site may contain pages related to the movies themselves, actors appearing in the movies, directors, and genres. Association rule mining process could generate a frequent itemset: {"British", "Romance", "Comedy" $\Rightarrow$ "Hugh Grant"}, suggesting that users who are interested in British romantic comedies may also like the actor Hugh Grant. During online recommendation, the user's active session (which is also transformed into a feature representation) is compared with the discovered patterns. Before recommendations are made, the matching patterns must be mapped back into Web pages or Web objects. In the above example, if the active session matches the left hand side of the association rule, the recommendation engine could recommend other Web pages that contains the feature "Hugh Grant".

The post-mining integration of semantic features into personalization involves combining the results of mining (performed independently on usage and content data) during the online recommendation phase. An example of this approach was presented in [12], where clustering algorithms were applied to both the transaction matrix TP and the transpose of the feature matrix PF. Since both matrices have pageviews as dimensions, the centroids of the resulting clusters in both cases can be represented as a set of pageview-weight pairs where the weights signify the frequency of the pageview in the corresponding cluster. We call the patterns generated from content data "content profiles", while the patterns derived from usage data are called "usage profiles". Though they share the same represenation, they have different semantics: usage profiles represent a set of transactions with similar user behavior patterns, while content profiles contain a group of Web pages with similar content.

Each such profile, in turn, can be represented as a vector in the original n-dimensional space of pageviews. This aggregate representation can be used directly in the recommendation phase: given a new user, u who has accessed a set of pages, P_u, so far, we can measure the similarity of P_u to the discovered profiles, and recommend to the user those pages in matching profiles which have not yet been accessed by the user. Note that this approach does not distinguish between recommendations emanating from the matching content and usage profiles. Also note that there are many other ways of combining usage profiles and content profiles

during the online recommendation phase. For example, we can use content profiles as the last resort in the situation where usage profiles can not provide sufficient recommendations.

The integration of content features with usage-based personalization is desirable when we are dealing with sites where text descriptions are dominant and other structural relationships in the data are not easy to obtain, e.g., news sites or online help systems, etc. This approach, however, is incapable of capturing more complex relations among objects at a deeper semantic level based on the inherent properties associated with these objects. To be able recommend different types of complex objects using their underlying properties and attributes, the system must be able to rely on the characterization of user segments and objects, not just based on keywords, but at a deeper semantic level using the domain ontologies for the objects. In the next section, we present a framework for integrating domain ontologies with the personalization process.

4 Integrating Web Usage Mining and Personalization with Domain Ontologies

At a conceptual level, there may be many different kinds of objects within a given site that are accessible to users. At the physical level, these objects may be represented by one or more Web pages. For example, the movie site mentioned earlier may contain pages related to the movies, actors, directors, studios, etc. Conceptually, each of these entities represents a different type of semantic object. During a visit to this site, a user may access several of these objects together during a session. In contrast to content features, ontological representation of domain knowledge contained in the site makes it possible to have a uniform architecture to model such objects, their properties, and their relationships.

In this section we will present a general framework for fully utilizing domain ontologies in Web usage mining and personalization. Figure 1 lays out a general process for such an integrated approach. As before, it is composed of 3 main phases: preprocessing, pattern discovery and online recommendation. Each of these phases must now take into account the object properties and their relationships.

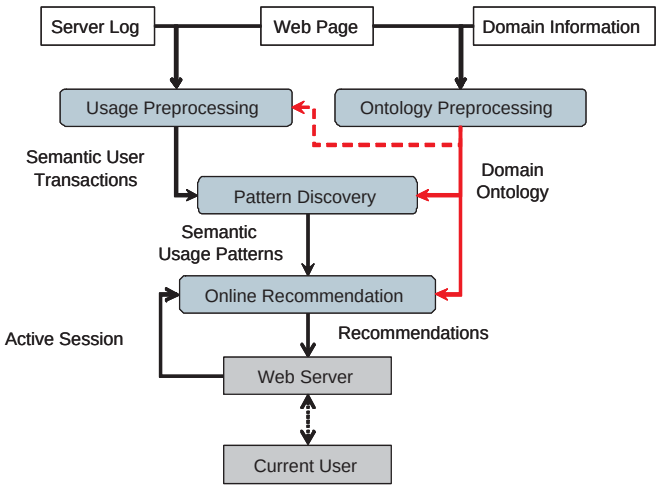

Figure 1: A Framework for Personalization Based on Domain Ontologies

4.1 Ontology Preprocessing

The ontology preprocessing phase takes as input domain information (such as database schema and metadata, if any) as well as Web pages, and generates the site ontology. For simple Web sites, ontologies can be easily designed manually or derived semi-automatically from the site content. However, it is more desirable to have automatic ontology acquisition methods for large Web site, especially in web sites with dynamically

generated Web pages. E-commerce web sites, for instance, usually have well-structured Web content, including predefined metadata or database schema. Therefore it is easier to build automatic ontology extraction mechanisms that are site-specific.

There have been a number of efforts dealing with the ontology learning problem [4, 8, 11, 2]. A wide range of information, such as thesauri, content features, and database schema can help to identify ontologies. Many of these approaches have focused on extracting ontological information from the Web, in general. In [1] the notion of "Semantic Web Mining" was introduced, including a framework for the extraction of a concept hierarchy and the application of data mining techniques to find frequently occurring combinations of concepts.

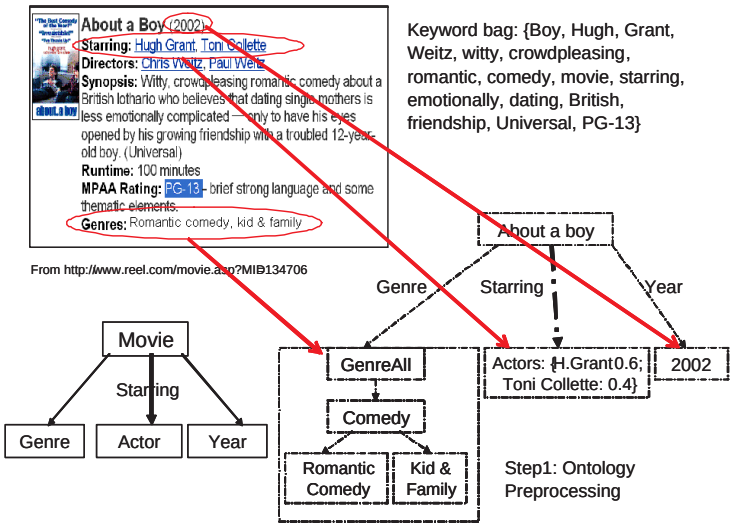

Figure 2: Example of Ontology Preprocessing

Figure 2 shows a "Movie" instance "About a Boy" and its related attributes and relations extracted from a Web page. The schema of the concept "Movie" is shown at the bottom left portion of the figure. Here we treat the concepts "Genre" and "Year" as attributes of the concept "Movie". The instances of the ontology are shown at the bottom right of the figure. The "Genre" attribute contains a partial order among labels representing a concept hierarchy of movie genres. We use a restriction of this partial order to represent the Genre to which the Movie instance belongs. The diagram also shows a keyword bag containing the important keywords in that page.

4.2 Pattern Discovery

As depicted in Figure 1 domain ontologies can be incorporated into usage preprocessing to generate semantic user transactions (pre-mining), or they can be integrated into pattern discovery phase to generate semantic usage patterns. Here we will focus on the latter approach.

Given a discovered usage profile (a set of pageview-weight pairs), as described earlier, we can transform it into a domain-level aggregate representation of the underlying objects ([6]. To distinguish between the representations we call the original discovered pattern an "item-level" usage profile, and we call the new profile based on the domain ontology a "domain-level" aggregate profile. The item-level profile is first represented as a weighted set of objects: $pr = \{\langle o_1, w_{o_1}\rangle, \langle o_2, w_{o_2}\rangle, \ldots, \langle o_n, w_{o_n}\rangle\}$ in which each o_i is an object in the underlying domain ontology and w_i represents o_i's significance in the profile pr. The profile represents a set of objects accessed together frequently by a group of users (as determined through Web usage mining). Objects, in the usage profile, that belong to the same class are combined to form an aggregated pseudo object belonging to that class. An important benefit of aggregation is that the pattern volume is significantly reduced, thus relieving the computation burden for the recommendation engine. Our goal is to create an aggregate representation of this weighted set of objects to characterize the common interests of

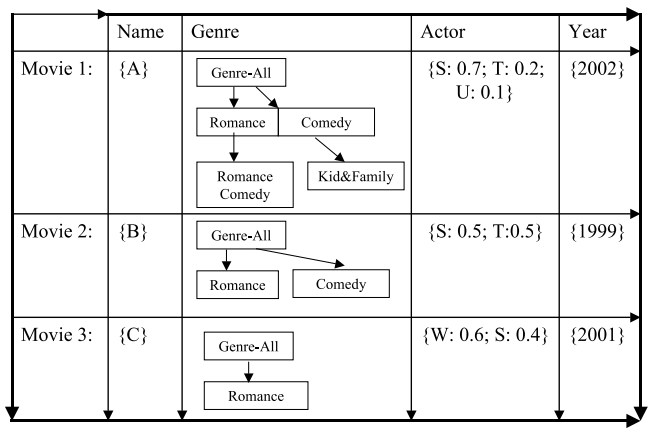

Figure 3: A Weighted Set of Objects in a Usage Profile from a Movie Web Site

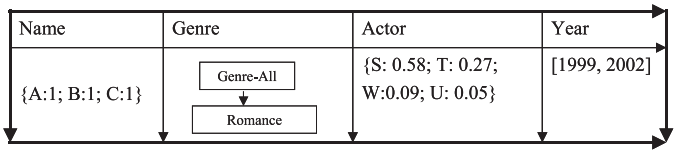

Figure 4: An Example of Domain-Level Aggregate Profile from a Movie Web site

the user segment captured by the usage profile at the domain level.

The aggregation process requires that a "Combination Function " ψ_a be defined for each attribute a of an object in the domain ontology. Figures 3 and 4 show an example of such process. Each movie object has attribute "Name", "Actor", "Genre" and "Year". For the attribute *Name*, we are interested in all the movie names appearing in the instances. Thus we can define ψ_{Name} to be the union operation performed on all the singleton *Name* attributes of all movie objects. On the other hand, the attribute *Actor* contains a weighted set of objects belonging to class **Actor**. In fact, it represents the relation "Starring" between the actor objects and the movie object. In such cases we can use a vector-based weighted mean operation as the combination function. In this case, the weight for an actor object o is determined by $w'_o = \frac{\sum_i w_i \cdot w_o}{\sum_i w_i}$.

Applying ψ_{Actor} to our example will result in the aggregate actor object $\{\langle$S, 0.58$\rangle$, $\langle$T, 0.27$\rangle$, $\langle$W, 0.09$\rangle$, $\langle$U, 0.05$\rangle\}$. As for the attribute *Year*, the combination function may create a range of all the *Year* values appearing in the objects. Another possible solution is to discretize the full *Year* range into decades and find the most common decades that are in the domains of the attribute. Applying ψ_{Year} to our example may result in an aggregate instance $Year'$ of attribute *Year*, namely $[1999, 2002]$.

The attribute *Genre* of concept **Movie** contains a partial order representing a concept hierarchy among different *Genre* values. The combination function, in this case, can perform tree (or graph) matching to extract the common parts of the conceptual hierarchies among all instances. Applying ψ_{Genre} to the example will result in an aggregate instance $Genre'$ of attribute *Genre* with the value $\{$"Romance"$\}$.

Figure 3 shows the item-level usage profile and its representation as a weighted set of objects. Figure 4 depicts the resulting domain-level aggregate profile. Note that the original item-level profile gives us little information about the reasons why these objects were commonly accessed together. However, after we characterize this profile at the domain-level, we find some interesting patterns: they all belong to *Genre* "Romance", and the actor S has a high score compared with other actors. This might tell us that this group of users are interested particularly in the movies belonging to "Romance" and are particularly fond of the actor S.

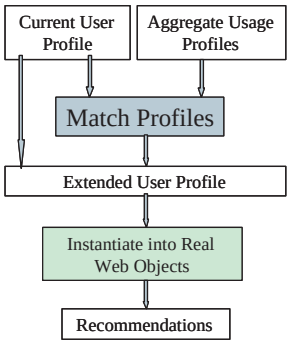

Figure 5: Online Recommendation Enhanced by Domain Ontologies

4.3 Online Recommendation Phase

In contrast to transaction-based usage profiles, semantic usage profiles capture the underlying common properties and relations among those objects. This fine-grained domain knowledge, captured in aggregate form enables more powerful approaches to personalization. As before, we consider the browsing history of the current user, i.e., active session, to be a weighted set of Web pages that the user has visited. The same transformation described in the last subsection can be used to create a semantic representation of the user's active session. We call this representation the *current user profile*.

Figure 5 presents the basic procedure for generating recommendations based on semantic usage profiles. The recommendation engine matches the current user profile against the discovered domain-level aggregate profiles. The usage profiles with matching score greater than some pre-specified threshold are considered to represent this user's potential interests. A successful match implies that the current user shares common interests with the group of users represented by the usage profile. The matching process results in an *extended user profile* which is obtained by applying the aggregation process described above to the matching domain-level profiles together with the original user profile.

The recommendation engine then instantiates the user's extended profile to real Web objects and will recommend them to the user. We can also exploit structural relationships among classes during the recommendation process. For example, if a concept hierarchy exists among objects, and the recommendation engine can not find a good match for a user profile at a certain concept level, then it can generalize to a more abstract level (e.g., from "romantic comedy" to "romance").

5 Conclusions and Future Work

This paper explores various approaches for integrating semantic knowledge into the personalization process based on Web usage mining. We have considered approaches based on the extraction of semantic features from the textual content contained in a site and their integration with Web usage mining tasks and personalization both in the pre-mining and the post-mining phases of the process. We have also presented a framework for Web personalization based on full integration of domain ontologies and usage patterns. The examples provided throughout the paper reveal how such a framework can provide insightful patterns and smarter personalization services. We leave some interesting research problems for open discussion and future work. Most important among these are techniques for computing similarity between domain objects and aggregate domain-level patterns, as well as learning techniques to automatically determine appropriate combination functions used in the aggregation process.

References

[1] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *International Semantic Web Conference (ISWC02)*, 2002.

[2] P. Clerkin, P. Cunningham, and C. Hayes. Ontology discovery for the semantic web using hierarchical clustering. In *Semantic Web Mining Workshop at ECML/PKDD-2001*, Freiburg, Germany, 2001.

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

[4] M. Craven, D. DiPasquo, D. Freitag, A.K. McCallum, T.M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 1/2(2-3):69–113, 2000.

[5] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining Content-based and Collaborative Filters in an Online Newspaper. In *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation.* University of California, Berkeley, Aug. 1999.

[6] H. Dai and B. Mobasher. Using ontologies to discover domain-level web usage profiles. In *2nd Semantic Web Mining Workshop at ECML/PKDD-2002*, 2002.

[7] B. Ganter and G. Stumme. Creation and merging of ontology top-levels. In *DBFusion 2002*, 2002.

[8] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *Proc. 18th International Conf. on Machine Learning*, pages 170–177, 2001.

[9] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *ACM SIGIR*, 1999.

[10] B. Mobasher, R. Cooley and J. Srivastava. Automatic personalization based on Web usage mining. In *Communications of the ACM*, (43) 8, August 2000.

[11] A. Maedche and S. Staab. Learning ontologies for the semantic web. In *Semantic Web Worshop 2001*, Hongkong, China, 2001.

[12] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000)*, Greenwich, UK., Sep. 2000.

[13] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Web Information and Data Management*, pages 9–15, 2001.

[14] B. Mobasher, H. Dai, T.Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.

[15] M. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review*, Dec. 1999, pp. 393-408.

[16] B. Sarwar, G. Karypis, J.A. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *ACM'00 Conference on Electronic Commerce*, pages 158–167, 2000.

[17] J. Strivastrava, R.t Cooley, M. Deshpande, and P-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), Jan. 2000.