A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content

Xin Jin, Yanzan Zhou, Bamshad Mobasher
{xjin,yzhou,mobasher}@cs.depaul.edu
Center for Web Intelligence
School of Computer Science, Telecommunication, and Information Systems
DePaul University, Chicago, Illinois, USA

Abstract

Web usage mining techniques, such as clustering of user sessions, are often used to identify Web user access patterns. However, to understand the factors that lead to common navigational patterns, it is necessary to develop techniques that can automatically characterize users' navigational tasks and intentions. Such a characterization must be based both on the common usage patterns, as well as on common semantic information associated with the visited Web resources. The integration of semantic content and usage patterns allows the system to make inferences based on the underlying reasons for which a user may or may not be interested in particular items. In this paper, we propose a unified framework based on Probabilistic Latent Semantic Analysis to create models of Web users, taking into account both the navigational usage data and the Web site content information. Our joint probabilistic model is based on a set of discovered latent factors that "explain" the underlying relationships among pageviews in terms of their common usage and their semantic relationships. Based on the discovered user models, we propose algorithms for characterizing Web user segments and to provide dynamic and personalized recommendations based on these segments. Our experiments, performed on real usage data, show that this approach can more accurately capture users' access patterns and generate more effective recommendations, when compared to more traditional methods based on clustering.

1 Introduction

Web users exhibit different types of behavior depending on their information need and their intended tasks. These behavior "types" are captured implicitly by a collection of actions taken by users during their visits to a site. The actions can range from viewing pages or buying products to interacting with online applications or Web services. For example, in an e-commerce site, there may be many user groups with different (but overlapping) behavior types. These groups may include visitors who engage in "window shopping" by browsing through a variety of product pages in different categories, visitors who are goal-oriented showing interest in a specific product category, or visitors who tend to place items in their shopping cart, but not purchase those items. Identifying these behavior types may, for example, allow a site to distinguish between those who show a high propensity to buy versus those who don't. This, in turn, can lead to automatic tools to tailor the content of pages for those users accordingly.

Web usage mining techniques (Srivastava *et al.* 2000), which capture usage patterns from users' navigational data, have achieved great success in various application areas such as Web personalization and recommender systems (Mobasher, Cooley, & Srivastava 2000; Mobasher, Dai, & T. Luo 2002; Nasraoui *et al.* 2002; Pierrakos *et al.* 2003), link prediction and analysis (Sarukkai 2000), Web site evaluation or reorganization (Spiliopoulou 2000; Srikant & Yang 2001), and e-commerce data analysis (Ghani & Fano 2002; Kohavi *et al.* 2004). An important problem in Web usage mining is to identify the underlying user goals and functional needs that lead to common navigational activity.

Most current Web usage mining systems use different data mining techniques, such as clustering, association rule mining, and sequential pattern mining to extract usage patterns from users' navigational data. Generally, these usage patterns are standalone patterns at the pageview level. They, however, do not capture the intrinsic characteristics of Web users' activities, nor can they quantify the underlying and unobservable factors that lead to specific navigational patterns. Thus, to better understand the factors that lead to common navigational patterns, it is necessary to develop techniques that can automatically characterize the users' underlying navigational objectives and to discover the hidden semantic relationships among users as well as between users and Web objects. This, in part, requires new approaches that can seamlessly integrate different sources of knowledge from both usage, as well as from the semantic content of Web sites.

The integration of content information about Web

objects with usage patterns involving those objects provides two primary advantages. First, the semantic information provides additional clues about the underlying reasons for which a user may or may not be interested in particular items. This, in turn, allows the system to make inferences based on this additional source of knowledge, possibly improving the quality of discovered patterns or the accuracy of recommendations. Secondly, in cases where little or no rating or usage information is available (such as in the case of newly added items, or in very sparse data sets), the system can still use the semantic information to draw reasonable conclusions about user interests.

Recent work (Mobasher *et al.* 2000; Anderson, Domingos, & Weld 2002; Dai & Mobasher 2002; Ghani & Fano 2002) has shown the benefits of integrating semantic knowledge about the domain (e.g., from page content features, relational structure, or domain ontologies) into the Web usage mining and personalization processes. There has also been a growing body of work in enhancing collaborative filtering systems by integrating data from other sources such as content and user demographics (Claypool *et al.* 1999; Pazzani 1999; Melville, Mooney, & Nagarajan 2001). Content-oriented approaches, in particular, can be used to address the "new item problem" discussed above. Generally, in these approaches, keywords are extracted from the content of Web pages and are used to recommend other pages or items to a user, not only based on user ratings or visit patterns, but also (or alternatively) based on the content similarity of these pages to other pages already visited by the user.

In most cases, however, these techniques involve independently learning user and content models, while integrating these after the fact in the recommendation process. In this paper, we are interested in developing a unified model of usage and content which can seamlessly integrate these sources of knowledge during the mining process. We believe that such an approach would be better able to capture the hidden semantic associations among Web objects and users, and thus result in patterns that can more closely represent the true interests of users and the context of their navigational behavior.

Latent semantic analysis (LSA) based on singular value decomposition (SVD) can capture the latent or hidden semantic associations among co-occurring objects (Deerwester *et al.* 1990). It is mostly used in automatic indexing and information retrieval (Berry, Dumais, & OBrien 1995), where LSA usually takes the (high dimensional) vector space representation of documents based on term frequency as a starting point and apply dimension reducing linear projection, such as Singular Value Decomposition (SVD) to generate a reduced latent space representation. LSA has been applied with remarkable success in different domains. Probabilistic Latent Semantic Analysis (PLSA), is a probabilistic variant of LSA which provides a more solid statistical foundation than standard LSA and has many applications in information retrieval and filtering, text learning and related fields (Hofmann 1999; 2001; Brants, Chen, & Tsochantaridis 2002; Brants & Stolle 2002). Approaches based on PLSA have also been used in the context of co-citation analysis (Cohn & Chang 2000; Cohn & Hofmann 2001).

In this paper we propose a Web usage mining approach based on PLSA. We begin with Web navigational data and Web site content information, and use these two sources of knowledge to create a joint probabilistic model of users' navigational activities. We then use the probabilistic model to discover and characterize Web user segments that capture both common navigation activity of users, as well as content characteristics which lead to such behavior. Based on the discovered patterns, we propose a recommendation algorithm to provide dynamic content to an active user. The flexibility of this model allows for varying degrees to which content and usage information is taken into account. It can, therefore, be utilized for personalization even when there is inadequate semantic knowledge about Web objects or sparse historical usage information.

We have conducted experiments on usage and content data collected from two different Web sites. The results show that our approach can successfully distinguish between different types of Web user segments according to the types of tasks performed by these users or the interest they showed in semantic attributes of the visited objects. Our results also suggest that the proposed approach results in more effective personalized recommendations when compared to other model-based approaches such as those based on clustering.

The paper is organized as follows. In Section 2 we provide an overview of our unified Probabilistic Latent Semantic Model as applied to both Web usage data and Web content information. Our algorithms for discovering Web user segments, based on the joint probabilistic model, and to generate recommendations are described in Section 3. Finally, in Section 4 we provide some examples of the discovered patterns and present our experimental evaluation.

2 Probabilistic Latent Semantic Models of Web User Navigations

The overall process of Web usage mining consists of three phrases: data preparation and transformation, pattern discovery, and pattern analysis. The data preparation phase transforms raw Web log data into transaction data that can be processed by various data mining tasks. In the pattern discovery phase, a variety of data mining techniques, such as clustering, association rule mining, and sequential pattern discovery can be applied to the transaction data. The discovered patterns should then be analyzed and interpreted for use in various applications, such as personalization, or for further analysis.

The usage data preprocessing phase (Cooley, Mobasher, & Srivastava 1999) results in a set of n pageviews, $P = \{p_1, p_2, \ldots, p_n\}$ and a set of m user

Attributes	Product A Pageview	Product B Pageview
Price (500-999)	0	1
Price (1000-1499)	1	0
Brand (HP)	1	0
Brand (Kodak)	0	1
Category (electronics)	1	1
Sub-category (computers)	1	0
Sub-category (camera)	0	1
...	...	...

Figure 1: Example of a hypothetical pageview-attribute matrix

sessions, $U = \{u_1, u_2, \ldots, u_m\}$. A *pageview* is an aggregate representation of a collection of Web objects (e.g. pages) contributing to the display on a user's browser resulting from a single user action (such as a click through, product purchase, or database query). The Web session data can be conceptually viewed as a session-pageview matrix (also called usage observation data), $UP_{m\times n}$, where the entry UP_{ij} corresponds to a weight associated with the pageview p_j in the user session u_i. The weights can be binary (representing the existence or not existence of a pageview within the session), based on the amount time spent on a page, or based on user ratings (such as in a collaborative filtering application).

Another important source of information in the discovery of navigational patterns is the content of the Web site. Each pageview contains certain semantic knowledge represented by the content information associated with that pageview. By applying text mining and information retrieval techniques, we can represent each pageview as an attribute vector. Attributes may be the keywords extracted from the pageviews, or structured semantic attributes of the Web objects contained in the pageviews. For instance, in an e-commerce site there many be many pageviews associated with specific products or product categories. Each product page can be represented by the product attributes (product name, price, category, etc). For example, suppose that a pageview A represents information about an HP laptop computer. This pageview may be represented as a vector (price=1200, brand=HP, sub-category=computer, ...). Similarly a pageview B about a Kodak camera can be represented as (price=600, brand=Kodak, sub-category=camera, ...). Applying content preprocessing techniques (Mobasher *et al.* 2000) to the Web site content, results in a set of s distinctive attribute values, $A = \{a_1, a_2, \ldots, a_s\}$ which comprise the *content observation data*. We can view these content observations as an attribute-pageview matrix $AP_{s\times n}$, where the entry AP_{tj} means pageview p_j contains the distinctive attribute value a_t. A portion of the attribute-pageview matrix for the above hypothetical example is depicted in Figure 1.

The PLSA model can be used to identify the hidden associations among variables in co-occurrence observation data. For the usage observations, a hidden (unobserved) factor variable $z_k \in Z = \{z_1, z_2, \cdots, z_l\}$ is associated (with certain probability) with each observation (u_i, p_j) corresponding to an access by user u_i to a Web resource p_j (i.e., an entry of matrix UP). Similarly, the hidden factor z_k is also probabilistically associated with each observation (a_t, p_j) (an entry of the attribute-pageview matrix AP). Our goal is to discover this set of latent factors $Z = \{z_1, z_2, \cdots, z_l\}$ from the usage and content observation data. The assumption behind this joint PLSA model is that the discovered latent factors "explain" the underlying relationships among pageviews both in terms of their common usage patterns, as well as in terms of their semantic relationships. The degree to which such relationships are explained by each factor is captured by the derived conditional probabilities that associate the pageviews with each of the latent factors.

The probabilistic latent factor model can be described as the following generative model:

1. select a user session u_i from U with probability $Pr(u_i)$;
2. select a latent factor z_k associated with u_i with probability $Pr(z_k|u_i)$;
3. given the factor z_k, generate a pageview p_j from P with probability $Pr(p_j|z_k)$.

As a result we obtain an observed pair (u_i, p_j), while the latent factor variable z_k is discarded. Translating this process into a joint probability model results in the following:

$$Pr(u_i, p_j) = Pr(u_i) \bullet Pr(p_j|u_i),$$

where,

$$Pr(p_j|u_i) = \sum_{z_k \in Z} Pr(p_j|z_k) \bullet Pr(z_k|u_i).$$

Furthermore, using Bayes' rule, we can transform this probability into:

$$Pr(u_i, p_j) = \sum_{z_k \in Z} Pr(z_k) \bullet Pr(u_i|z_k) \bullet Pr(p_j|z_k).$$

Similarly, each attribute-page observation (a_t, p_j) can be modeled as:

$$Pr(a_t, p_j) = \sum_{z_k \in Z} Pr(z_k) \bullet Pr(a_t|z_k) \bullet Pr(p_j|z_k).$$

Combining them together, the total likelihood $L(U, A, P)$ of two observation data matrices is then described as:

$$L(U, A, P) = \alpha \sum_{i,j} UP_{ij} \bullet \log \Pr(u_i, p_j) + (1 - \alpha) \sum_{t,j} AP_{tj} \bullet \log \Pr(a_t, p_j),$$

where α is the combination parameter, which is used to adjust the relative weights of usage observations and attribute observations.

Thus, the process of generating a model that "explains" observations (U, P) and (A, P) amounts to estimating parameters $Pr(z_k)$, $Pr(u_i|z_k)$, $Pr(a_t|z_k)$ and $Pr(p_j|z_k)$ which can maximize the overall likelihood, $L(U, A, P)$.

Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin 1977) is a well-known approach to perform maximum likelihood parameter estimation in latent variable models. It alternates two steps: (1) an expectation (E) step where posterior probabilities are computed for latent variables $z_k \in Z$, based on the current estimates of the parameters; and (2) a maximization (M) step, where parameters are updated for given posterior probabilities computed in the previous E-step. The EM algorithm is guaranteed to reach a local optimum.

To apply the EM algorithm in this context, we begin with some initial values of $Pr(z_k)$, $Pr(u_i|z_k)$, $Pr(a_t|z_k)$ and $Pr(p_j|z_k)$. In the expectation step we compute:

$$Pr(z_k|u_i, p_j) = \frac{Pr(z_k) \bullet Pr(u_i|z_k) \bullet Pr(p_j|z_k)}{\sum_{z \in Z} Pr(z) \bullet Pr(u_i|z) \bullet Pr(p_j|z)}$$

and, similarly

$$Pr(z_k|a_t, p_j) = \frac{Pr(z_k) \bullet Pr(a_t|z_k) \bullet Pr(p_j|z_k)}{\sum_{z \in Z} Pr(z) \bullet Pr(a_t|z) \bullet Pr(p_j|z)}.$$

In the maximization step, we compute:

$$\Pr(z_k) \propto \alpha \sum_{i,j} UP_{ij} \bullet \Pr(z_k|u_i, p_j) + (1 - \alpha) \sum_{t,j} AP_{tj} \bullet \Pr(z_k|a_t, p_j),$$

$$Pr(u_i|z_k) = \frac{\sum_{p_j \in P} UP_{ij} \bullet Pr(z_k|u_i, p_j)}{\sum_{u \in U, p_j \in P} UP_{ij} \bullet Pr(z_k|u, p_j)},$$

$$Pr(a_t|z_k) = \frac{\sum_{p_j \in P} AP_{tj} \bullet Pr(z_k|a_t, p_j)}{\sum_{a \in A, p_j \in P} AP_{tj} \bullet Pr(z_k|a, p_j)},$$

$$Pr(p_j|z_k) \propto \alpha \sum_{i} UP_{ij} \bullet Pr(z_k|u_i, p_j) + (1 - \alpha) \sum_{t} AP_{tj} \bullet Pr(z_k|a_t, p_j).$$

Iterating the above computation of expectation and maximization steps monotonically increases the total likelihood of the observed data $L(U, A, P)$ until a local optimal solution is reached. Furthermore, note that varying the combination parameter α allows the model to take into account the usage-based and content-based relationships among pageviews in varying degrees, as may be appropriate for a particular Web site. For, instance, in a content-rich Web site, α may be set to 50%, equally taking into account relationships from both sources. On the other hand, if adequate content information is not available, α can be set to 1, in which case the model will be based solely on the usage observations.

The computational complexity of this algorithm is $O(mnl + snl)$, where m,n,s, and l represent the number of user sessions, pageviews, attribute values, and factors, respectively. Since both the usage observation matrix and the attribute matrix are very sparse, the memory requirment can be dramaticlly reduced using efficient sparse matrix implementation.

3 A Recommendation Framework Based on the Joint PLSA Model

Applying the EM algorithm, as described in Section 2, will result in estimates of $Pr(z_k)$, $Pr(u_i|z_k)$, $Pr(a_t|z_k)$, and $Pr(p_j|z_k)$, for each $z_k \in Z$, $u_i \in U$, $a_t \in A$, and $p_j \in P$. In this context, the hidden factors $z_k \in Z$ correspond to users' different task-oriented navigational patterns. In this section, we discuss how the generated probability estimates can be used for common Web usage mining tasks and applications. We, first, present an algorithm for creating aggregate representations that characterize typical Web user segments based on their common navigational behaviors and interests. These aggregate representations constitute the discovered user models. We then present a recommendation algorithm that combines the discovered models and the ongoing activity of a current user to provide dynamic recommendations.

3.1 Characterizing Web User Segments

We can use the probability estimates generated by the model to characterize user segments according to users' navigational behavior. Each segment will be represented as a collection of pageviews which are visited by users who are performing a similar task. We take each latent factor z_k, generated by the model, as corresponding to one such user segment. To this end we will use $Pr(u_i|z_k)$, which represents the probability of observing a user session u_i, given that a particular user segment is chosen.

A particular advantage of the probabilistic factor model, in contrast to probabilistic mixture models, is that a particular user session can be seen as belonging to not just one, but a combination of segments represented by the latent factors. For instance, a user session may correspond (with different probabilities) to two different factors z_1 and z_2. This is important in the context of user navigational patterns, since a user may, indeed, perform different information seeking or functional tasks during the same session. We describe our approach for characterizing user segments based on the latent factors below.

For each hidden factor z_k, the top user sessions with the highest $Pr(u|z_k)$ values can be considered to be the "prototypical" user sessions corresponding to that factor. In other words, these user sessions represent

the typical navigational activity of users who are performing a similar task. Thus, we can characterize task-oriented user segments based on the top sessions corresponding to each factor.

For each user segment z_k, we choose all the user sessions with probability $Pr(u_i|z_k)$ exceeding a certain threshold μ. Since each user session $\vec{u}$ can also be viewed as a pageview vector in the original n-dimensional space of pageviews, we can create an aggregate representation of the collection of user sessions related to z_k also as a pageview vector. The algorithm to generate this aggregate representation of user segments is as follow.

1. Input: $Pr(u_i|z_k)$, user session-page matrix UP, threshold μ.
2. For each z_k, choose all the sessions with $Pr(u_i|z_k) \geq \mu$ to get a user session set R.
3. For each z_k, compute the weighed average of all the chosen sessions in R to get a page vector $\vec{v}$ as:

$$\vec{v} = \frac{\sum_R \vec{u}_i \bullet Pr(u_i|z_k)}{|R|}$$

   where $|R|$ denotes the total number of chosen sessions for the factor z_k.
4. For each factor z_k, output page vector $\vec{v}$. This page vector consists of a set of weights, for each pageview in P, which represents the relative visit frequency of each pageview for this user segment.

We can sort the pageviews in $\vec{v}$ by weight so that the top elements correspond to the most frequently visited pages within the user segment. In this way, each user segment is characterized by an "aggregate" representation of all individual users' navigational activities from that user group. In the following, by "user segments" we mean their aggregate representations as described above.

The characterization of Web user segments, by itself, can help analysts to understand the behavior of individual or groups of users based on their navigational activity as well as their interest in specific content information. However, the probabilistic latent factor model also provides the flexibility to perform a variety of other supervised or unsupervised analysis tasks.

For example, we can categorize pages in a Web site, according to common usage patterns corresponding to different user segments. Specifically, given a Web page p, for each factor z, we can compute

$$Pr(z|p) = \frac{Pr(p|z) \bullet Pr(z)}{\sum_{z'} Pr(p|z') \bullet Pr(z')}.$$

Then, we can select the dominant z with the highest $Pr(z|p)$ as the class label for this page.

We can also use a similar approach to classify users, or to predict the liklihood that a user may visit a previously unvisited page. Specifically, given a user u_i, we can compute the probability of a (a previously unvisited) page P_j being visited by u_i as:

$$Pr(p_j|u_i) = \sum_z Pr(p_j|z) \bullet Pr(z|u_i).$$

In the following section we present an approach for Web personalization, based on the joint probabilistic user models generated using the latent factor model.

3.2 Using the Joint Probability Model for Personalization

Web personalization usually refers to dynamically tailoring the presentation of a Web site according to the interests or preferences of individual (or groups of users). This can be accomplished by recommending Web resources to a user by considering the current user's behavior together with learned models of past users (e.g., collaborative filtering). Here we will use the probabilistic user models generated via our joint PLSA framework to generate recommendations.

Given a set of user segments and an active user session, the method of generating a top-N recommendations is as follows.

1. Represent each user segment C as an n-dimensional pageview vector using the approach described above, where n is the total number of pages. Thus, $C = \langle \omega_1^C, \omega_2^C, \cdots, \omega_n^C \rangle$, where ω_i^C is the weight associated with pageview p_i in C. Similarly, the active user session S is represented as $S = \langle S_1, \cdots, S_n \rangle$, where S_i is set to 1, if pageview p_i is visited, and to 0, otherwise.
2. Choose the segment that best matches the active user session. Here we use the standard cosine coefficient to compute the similarity between the active user session and the discovered user segments.

$$match(S, C) = \frac{\sum_n (S_n \times \omega_n^C)}{\sqrt{\sum_n (S_n)^2 \times \sum_n (\omega_n^C)^2}}$$

3. Given the top matching segment C_{top} and the active user session S, a recommendation score, $Rec(S, p)$ is computed for each page $p \in C_{top}$ as:

$$Rec(S, p) = \sqrt{weight(p, C_{top}) \bullet match(S, C)}.$$

   Thus, each page will receive a normalized value between 0 and 1. If the page p is already in the current active session S, its recommendation value is set to zero.
4. Choose the top N pages with the highest recommendation values to get a top-N recommendation set.

The above approach for generating recommendations is not unique to the PLSA model. The traditional approach for discovering user segments is based on clustering of user records. In (Mobasher, Dai, & T. Luo 2002), k-means clustering was used to partition user sessions into clusters. The centroid of each cluster was

then obtained as an n-dimensional pageview vector, in a similar manner as described above. The cluster centroids, thus, provide an aggregate representation of common user patterns corresponding to each segment. This process was called *Profile Aggregation Based on Clustering Transactions* (PACT). The discovered user segments were then used to generate recommendations using the algorithm described above. In our experiments, discussed below, we use PACT as a point of comparison for the effectiveness of generated recommendations.

4 Experimental Evaluation

To evaluate the effectiveness of the PLSA-based model, we perform two typed of evaluation using two different data sets. First, we evaluate individual user segments to determine the degree to which they actually represent activities of similar users. Secondly, we evaluate our recommendation algorithm, based on the user segments, in the context of top-N recommendation framework. In each case, we compare our approach with the clustering approach for the discovery of Web user segments (PACT) (Mobasher, Dai, & T. Luo 2002), as described above.

4.1 Description of the Data Sets

In our experiments, we have used Web server log data from two Web sites. The first data set is based on the server log data from the host Computer Science department. After data preprocessing, we have identified 21,299 user sessions (U) and 692 Web pageviews (P), where each user session consists of 9.8 pageviews in average. We refer to this data set as the "CTI data." In this data set we used the time spent on each pageview as the weight associated with that pageview in the given session. Since most Web pages are dynamically generated, we do not adopt any content information from this site. Hence, this data set is used to evaluate our approach when only usage information is available for analysis.

The second data set is from the server logs of a local affiliate of a national real estate company. The primary function of the Web site is to allow prospective buyers visit various pages and information related to some 300 residential properties. The portion of the Web usage data during the period of analysis contained approximately 24,000 user sessions from 3,800 unique users. The preprocessing phase for this data was focused on extracting a full record for each user of properties they visited. This required performing the necessary aggregation operations on pageviews in order to treat a property as the atomic unit of analysis. In addition, the visit frequency for each user-property pair was recorded, since the number of times a user comes back to a property listing is a good measure of that user's interest in the property. Finally, the data was filtered to limit the final data set to those users that had visited at least 3 properties. In average, each user has visited 5.6 properties. In our final data matrix, each row represented a user vector with properties as dimensions and visit frequencies as the corresponding dimension values. We refer to this data set as the "Realty data."

For the "Realty data," in addition to the usage observations, we also extracted the content information related to the properties. Each property has a set of attributes, including price, number of bedrooms, number of bathrooms, size, garage size (cars), and school district. After content preprocessing, we built an attribute-page matrix (similar to Figure 1) to represent the content observations. In this matrix, each column represents a property, and each row is a distinct attribute value associated with the properties.

Each data set (the usage observations) was randomly divided into multiple training and test sets to use with 10-fold cross-validation. The training sets were used to build the models while the test sets were used to evaluate the user segments and the recommendations generated by the models. In our experiments, the results presented in the following sections represent averages over the 10 folds.

By conducting some sensitivity analysis, we chose 30 factors in the case of CTI data, and 15 factors for the Realty data. Furthermore, we employ a "tempered" version of the EM algorithm as described in (Hofmann 2001) to avoid "overtraining."

4.2 Examples of Extracted Latent Factors

We begin by providing an example of the latent factors generated using the joint PLSA model for the Realty data. Figure 2 depicts six of the 15 factors extracted from this data set. The factors were generated using a combination of usage and content observations associated with the real estate properties. For each of the factors, the most significant attributes (based on the probability of their association with that factor) are given along with the description of the attributes.

Factors 1 and 2 in this figure clearly indicate the interest of users in properties of similar price and size. The model, however, distinguishes the two groups based on other attribute information. In particular, factor 1 represents interest in "town homes" located in the ANK school district, while factor 2 represents 2-story family units in the WDM school district. Factors 3 and 4 both represent larger properties in a higher price range, but again, they are distinguished based on user's interests in different school districts. Finally, factors 5 and 6 both represent much lower priced properties of the same type. However, the relationship between these two factors is more nuanced than the other cases above. Here, factor 6 represents a special case of factor 5, in which users have focused particularly in the DSM school district. Indeed, our experiments show that the joint PLSA model can capture overlapping interests of a similar group if users in different items at various levels of abstraction.

	Pr(attribute|factor)	attbribute	
factor 1	0.1126	#price	100,000-199,999
	0.0967	#garage	2
	0.0818	#size	1,000-1,999
	0.0689	#school	ANK
	0.0552	#bathroom	1.5
	0.0423	#bedroom	3
	0.0419	#year	1970-1979
	0.0417	#style	townhome
factor 2	0.1118	#price	100,000-199,999
	0.0928	#style	2_story
	0.0916	#school	WDM
	0.0711	#exterior	vinyl_siding
	0.0689	#garage	2
	0.0660	#year	1990-1999
	0.0632	#size	1,000-1,999
	0.0613	#bathroom	1.5
	0.0598	#bedroom	3
factor 3	0.0907	#year	1990-1999
	0.0788	#bedroom	4
	0.0758	#style	2_story
	0.0740	#bathroom	2.5
	0.0673	#garage	3
	0.0628	#price	200,000-299,999
	0.0588	#size	3,000-3,999
	0.0385	#school	ANK
factor 4	0.0883	#style	2_story
	0.0870	#bathroom	2.5
	0.0729	#price	200,000-299,999
	0.0729	#size	3,000-3,999
	0.0729	#year	1980-1989
	0.0672	#garage	3
	0.0540	#bedroom	5
	0.0521	#school	WDM
	0.0456	#year	1990-1999
factor 5	0.0895	#bathroom	1
	0.0681	#size	1,000-1,999
	0.0627	#price	<100,000
	0.0564	#style	ranch
	0.0549	#bedroom	2
	0.0505	#style	1_story
	0.0484	#garage	2
	0.0423	#bedroom	3
factor 6	0.1119	#school	DSM
	0.1012	#bedroom	3
	0.0797	#style	1_story
	0.0739	#year	1950-1959
	0.0716	#price	<100,000
	0.0660	#size	1,000-1,999
	0.0525	#bathroom	1.5
	0.0448	#garage	2

Figure 2: Examples of extracted factors from the Realty data

4.3 Evaluation of Discovered User Segments

In order to evaluate the quality of individual user segments we use a metric called the *Weighted Average Visit Percentage* (WAVP) (Mobasher, Dai, & T. Luo 2002). WAVP allows us to evaluate each segment individually according to the likelihood that a user who visits any page in the (aggregate representation of the) segment will visit the rest of the pages in that segment during the same session. Specially, let T be the set of transactions in the evaluation set, and for a segment S, let T_S denote a subset of T whose elements contain at least one page from S. Now, the weighted average similarity to the segment S over all transactions is computed (taking both the transactions and the segment as vectors of pageviews) as:

$$WAVP = (\sum_{t \in T_S} \frac{\vec{t} \bullet \vec{S}}{|T_S|})/(\sum_{p \in S} weight(p, S))$$

Note that a higher WAVP value implies better quality of a segment, in the sense that the segment represents the actual behavior users based on their similar activities.

In these experiments we compare the WAVP values for the generated segments using the PLSA model and those generated by PACT model. Figures 3 and 4 depict these results for the CTI and Realty data sets, respectively. In each case, the segments are ranked in the decreasing order of WAVP. The results show clearly that the probabilistic segments based on the latent factors provide a significant advantage over the clustering approach (p-value < 0.05, 95%).

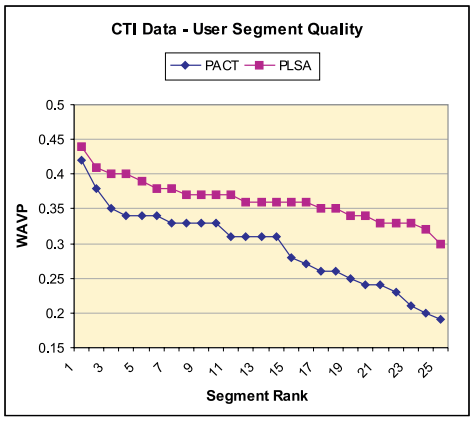

Figure 3: Comparison of user segments in the CTI site; PLSA model v. k-means clustering (PACT). (p-value over the average WAVP for all segments is < 0.05, 95%)

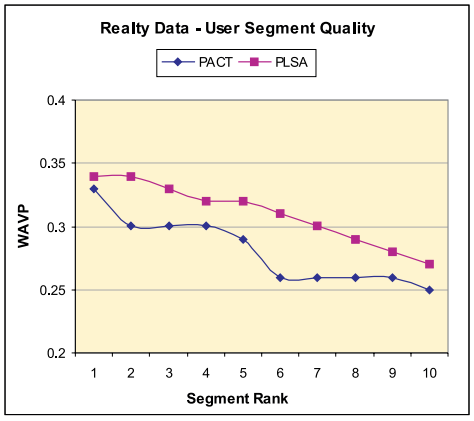

Figure 4: Comparison of user segments in the real estate site; PLSA model v. k-means clustering (PACT). (p-value over the average WAVP for all segments is < 0.05, 95%)

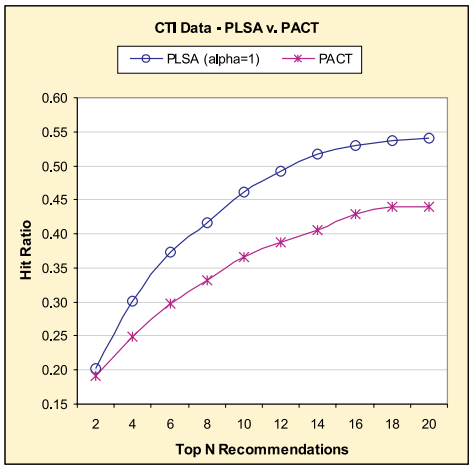

Figure 5: Comparison of generated page recommendations based on PLSA segments versus PACT segments in the CTI site. (for $N >= 4$, p-value < 0.05, 95%)

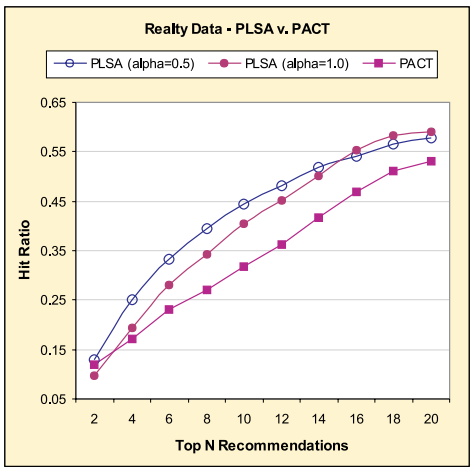

Figure 6: Comparison of generated property recommendations based on PLSA segments versus PACT segments in the real estate site. (for $N \in [4, 10]$, p-value < 0.05, 95%)

4.4 Evaluation of the Recommendation Algorithm

For evaluating the recommendation effectiveness we use a measure called *Hit Ratio* in the context of top-N recommendations. For each user session in the test set, we took the first K pages as a representation of an active user to generate a top-N recommendation set. We then compared the recommendations with the pageview $K+1$ in the test session. If there is a match, this is considered a hit. We define the Hit Ratio as the total number of hits divided by the total number of user sessions in the test set. Note that the Hit Ratio increases as the value of N (number of recommendations) increases. Thus, in our experiments, we pay special attention to smaller number recommendations that result in good hit ratios.

In these experiments we compared the recommendation accuracy of the PLSA-based algorithm with that of PACT. In each case, the recommendations were generated according to the algorithms presented in Section 3.2. The recommendation accuracy was measured based on hit ratio for different number of generated recommendations. These results are depicted in Figures 5 and 6 for the CTI and realty data sets, respectively. In the case of the CTI data, we used $\alpha = 1$, solely taking into account the usage observations. In the case of the Realty data, however, we compared the result of PACT recommendations with the PLSA-based recommendations both with $\alpha = 1$ (usage-only), as well as with $\alpha = 0.5$ (equally weighted content and usage).

The result show a general advantage for the PLSA model. In most realistic situations, we are interested in a small, but accurate, set of recommendations. Generally, a reasonable recommendation set might contain 5 to 10 recommendations (At these levels, the difference of the performance of the PLSA model and the clustering approach is statistically significant.). Indeed, this range of values seem to represent the largest improvements of the PLSA model over the clustering approach. In the case of the Realty data, the combined usage-content model provides a small gain in accuracy over the usage-only model (particularly at lower numbers of recommendations). The more significant advantage of the combined content-usage model, however, is in its ability to generate recommendations in the face of sparse or insufficient usage data, as well as in providing a better semantic characterization of user segments.

5 Conclusions

Users of a Web site exhibit different types of navigational behavior depending on their intended tasks or their information needs. However, to understand the factors that lead to common navigational patterns, it is necessary to develop techniques that can automatically characterize users' tasks and intentions, both based on their common navigational behavior, as well as based on semantic information associated with visited resources. In this paper, we have used a joint probabilistic latent semantic analysis framework to develop a unified model of Web user behavior based on the usage and content data associated with the site. The probabilistic model provides a great deal of flexibility as the derived probability distributions over the space of latent factors can be used for a variety of Web mining and analysis tasks. In particular, we have presented algorithms based on the joint PLSA models to discover and characterize Web user segments and to provide dynamic and personalized recommendations based on these segments.

Our experimental results show clearly that, in addition to greater flexibility, the PLSA approach to Web usage mining, generally results in a more accurate rep-

resentation of user behavior. This, in turn, results in higher quality patterns that can be used effectively in Web recommendation.

References

Anderson, C.; Domingos, P.; and Weld, D. 2002. Relational markov models and their application to adaptive web navigation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*.

Berry, M.; Dumais, S.; and OBrien, G. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* 37:573–595.

Brants, T., and Stolle, R. 2002. Find similar documents in document collections. In *Proceedings of the Third International Conference on Language Resources and Evaluation(LREC-2002)*.

Brants, T.; Chen, F.; and Tsochantaridis, I. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*.

Claypool, M.; Gokhale, A.; Miranda, T.; Murnikov, P.; Netes, D.; and Sartin, M. 1999. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*.

Cohn, D., and Chang, H. 2000. Probabilistically identifying authoritative documents. In *Proceedings of the Seventeenth International Conference on Machine Learning*.

Cohn, D., and Hofmann, T. 2001. The missing link: A probabilistic model of document content and hypertext connectivity. In Todd K. Leen, T. G. D., and Tresp, V., eds., *Advances in Neural Information Processing Systems 13*. MIT Press.

Cooley, R.; Mobasher, B.; and Srivastava, J. 1999. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems* 1(1).

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

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Hashman, R. 1990. Indexing by latent semantic indexing. *Journal of the American Society for Information Science* 41(6).

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society* B(39):1–38.

Ghani, R., and Fano, A. 2002. Building recommender systems using a knowledge base of product semantics. In *Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at the 2nd Int'l Conf. on Adaptive Hypermedia and Adaptive Web Based Systems*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal* 42(1):177–196.

Kohavi, R.; Mason, L.; Parekh, R.; and Zheng, Z. 2004. Lessons and challenges from mining retail e-commerce data. *To appear in Machine Learning*.

Melville, P.; Mooney, R.; and Nagarajan, R. 2001. Content-boosted collaborative filtering. In *Proceedings of the SIGIR2001 Workshop on Recommender Systems*.

Mobasher, B.; Dai, H.; Luo, T.; Sun, Y.; and Zhu, J. 2000. Integrating web usage and content mining for more effective personalization. In *E-Commerce and Web Technologies: Proceedings of the EC-WEB 2000 Conference*, Lecture Notes in Computer Science (LNCS) 1875, 165–176. Springer.

Mobasher, B.; Cooley, R.; and Srivastava, J. 2000. Automatic personalization based on web usage mining. *Communications of the ACM* 43(8):142–151.

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

Nasraoui, O.; Krishnapuram, R.; Joshi, A.; and Kamdar, T. 2002. Automatic web user profiling and personalization using robust fuzzy relational clustering. In Segovia, J.; Szczepaniak, P.; and Niedzwiedzinski, M., eds., *Studies in Fuzziness and Soft Computing*. Springer-Verlag.

Pazzani, M. 1999. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review* 13(5-6):393–408.

Pierrakos, D.; Paliouras, G.; Papatheodorou, C.; and Spyropoulos, C. 2003. Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction* 13:311–372.

Sarukkai, R. 2000. Link prediction and path analysis using markov chains. In *Proceedings of the 9th International World Wide Web Conference*.

Spiliopoulou, M. 2000. Web usage mining for web site evaluation. *Communications of the ACM* 43(8):127–134.

Srikant, R., and Yang, Y. 2001. Mining web logs to improve website organization. In *Proceedings of the 10th International World Wide Web Conference*.

Srivastava, J.; Cooley, R.; Deshpande, M.; and Tan, P. 2000. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations* 1(2):12–23.