Web Usage Mining Based on Probabilistic Latent Semantic Analysis

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

ABSTRACT

The primary goal of Web usage mining is the discovery of patterns in the navigational behavior of Web users. Standard approaches, such as clustering of user sessions and discovering association rules or frequent navigational paths, do not generally provide the ability to automatically characterize or quantify the unobservable factors that lead to common navigational patterns. It is, therefore, necessary to develop techniques that can automatically identify the users' underlying navigational objectives and to discover hidden semantic relationships among users as well as between users and Web objects. Probabilistic Latent Semantic Analysis (PLSA) is particularly useful in this context, since it can uncover latent semantic associations among users and pages based on the co-occurrence patterns of these pages in user sessions. In this paper, we develop a unified framework for the discovery and analysis of Web navigational patterns based on PLSA. We show the flexibility of this framework in characterizing various relationships among users and Web objects. Since these relationships are measured in terms of probabilities, we are able to use probabilistic inference to perform a variety of analysis tasks such as user segmentation, page classification, as well as predictive tasks such as collaborative recommendations. We demonstrate the effectiveness our approach through experiments performed on several real-world data sets.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

Keywords

Web usage mining, Latent variable models, User profiling, PLSA

KDD'04, August 22–25, 2004, Seattle, Washington, USA.

1. INTRODUCTION

Web users exhibit different types of behavior depending on their information needs and their intended tasks. These tasks are captured implicitly by a collection of actions taken by users during their visits to a site. For example, in a dynamic application-based e-commerce Web site, user tasks may be reflected by sequences of interactions with Web applications to search a catalog or to make a purchase. On the other hand, in an information intensive site, such as a portal or an online news source, user tasks may be reflected in a series of user clicks on a collection of Web pages with related content.

The identification of intended user tasks can shed light on various types of user navigational behaviors. For example, in an e-commerce site, there may be many user groups with different (but overlapping) behavior types. These 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 user tasks and behavior types may, for example, allow a site to distinguish between those who show a high propensity to buy versus whose who don't. This, in turn, can lead to automatic tools that can tailor the content of pages for those users accordingly.

Web usage mining techniques [7, 34], which capture Web users' navigational patterns, have achieved great success in various application areas such as Web personalization [22, 24, 25, 27], link prediction and analysis [20, 29], Web site evaluation or reorganization [31, 33], Web analytics and e-commerce data analysis [13, 19], Adaptive Web sites [26, 21], and Web pre-fetching [30, 28]. 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 user historical 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. A common approach for capturing the

latent or hidden semantic associations among co-occurring objects is Latent semantic analysis (LSA) [10]. It is mostly used in automatic indexing and information retrieval [2], where LSA usually takes the (high dimensional) vector space representation of documents based on term frequency as a starting point and applies a dimension reducing linear projection, such as Singular Value Decomposition (SVD) to generate a reduced latent space representation.

Probabilistic latent semantic analysis (PLSA) models, proposed by Hofmann [14, 16], provide a probabilistic approach for the discovery of latent variables which is more flexible and has a more solid statistical foundation than the standard LSA. The basis of PLSA is a model often referred to as the *aspect model* [17]. Assuming that there exist a set of hidden factors underlying the co-occurences among two sets of objects, PLSA uses Expectation-Maximization (EM) algorithm to estimate the probability values which measure the relationships between the hidden factors and the two sets of objects. Due to its great flexibility, PLSA has been widely and successfully used in variety of application domain, including information retrieval [15], text learning [3, 4, 12, 18], and co-citation analysis [5, 6].

In this paper, we propose a Web usage mining approach based on the PLSA model. In the Web usage scenario, as in information retrieval, we have co-occurrence data which, in this case, is comprised of Web users and Web objects. In this paper, we refer to the hidden factors that represent the latent relationships among these entities as *tasks*. This is to emphasize the fact that these factors generally represent the navigational objectives of users in a Web site, as reflected in their interaction with the Web objects.

By applying the PLSA model, we can effectively identify and characterize these hidden factors, thus quantitatively measuring the relationships between Web users and tasks, as well as between Web objects and tasks. These relationships are measured in terms of probabilities, which, in turn, allow for the discovery of a variety of usage patterns by using probabilistic inference. In this way, the model enables different types of analysis including the characterization of a task by a group of most related pages; the identification of prototypical users who perform a certain task; the identification of underlying tasks present in a specific user's activity; and the characterization of user groups (or segments) that perform a similar set of tasks.

The primary contributions of this paper are two-fold: first, we develop a general framework for discovery and analysis of Web navigational patterns based on the PLSA model. Secondly, we show, in detail, how this model can be used to generate various usage pattern, such as those described above, and point out possible applications, including a specific approach for Web personalization based on the discovered user segments. Furthermore, we illustrate many of these usage patterns by providing several illustrative examples based on real Web usage data, and we quantitatively evaluate the effectiveness of derived user segments.

The paper is organized as follows. In Section 2 we provide an overview of Probabilistic Latent Semantic Analysis model as applied to Web usage data. We present the details of deriving various usage patterns based on the PLSA model in Section 3. Finally, we present our experiments and interpretation of the result in Section 4 and conclude the paper in Section 5.

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 may then be analyzed and interpreted for use in such applications as Web personalization.

The usage data preprocessing phase [8, 32] results in a set of n pageviews, $P = \{p_1, p_2, \ldots, p_n\}$ and a set of m user 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 an $m \times n$ session-pageview binary matrix $UP = [w(u_i, p_j)]_{m \times n}$, where $w(u_i, p_j)$ represents the weight of pageview p_j in a user session u_i. The weights can be binary, representing the existence or non-existence of the pageview in the session, or they may be a function of the occurrence or duration of the pageview in that session.

PLSA is a latent variable model which associates hidden (unobserved) factor variable $Z = \{z_1, z_2, ..., z_l\}$ with observations in the co-occurences data. In our context, each observation corresponds to an access by a user to a Web resource in a particular session which is represented as an entry of the $m \times n$ co-occurrence matrix UP.

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. pick a latent factor z_k with probability $Pr(z_k|u_i)$,
3. 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

$$Pr(p_j|u_i) = \sum_{k=1}^{l} Pr(p_j|z_k) \bullet Pr(z_k|u_i),$$

summing over all possible choices of z_k from which the observation could have been generated. Using Bayes' rule, it is straightforward to transform the joint probability into:

$$P(u_i, p_j) = \sum_{k=1}^{l} Pr(z_k) \bullet Pr(u_i|z_k) \bullet Pr(p_j|z_k).$$

Now, in order to explain a set of observations (U, P), we need to estimate the parameters $Pr(z_k)$, $Pr(u_i|z_k)$, $Pr(p_j|z_k)$, while maximizing the following likelihood $L(U, P)$ of the observations,

$$L(U,P) = \sum_{i=1}^{m}\sum_{j=1}^{n} w(u_i,p_j)\log Pr(u_i,p_j).$$

Expectation-Maximization (EM) algorithm [11] is a well-known approach to performing 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, based on the current estimates of the parameters, (2) a maximization (M) step, re-estimate the parameters in order to maximize the expectation of the complete data likelihood.

The EM algorithm begins with some initial values of $Pr(z_k)$, $Pr(u_i|z_k)$, and $Pr(p_j|z_k)$. In the expectation step we compute:

$$Pr(z_k|u_i,p_j) = \frac{Pr(z_k) \bullet Pr(u_i|z_k) \bullet Pr(p_j|z_k)}{\sum_{k'=1}^{l} Pr(z_{k'}) \bullet Pr(u_i|z_{k'}) \bullet Pr(p_j|z_{k'})}.$$

In the maximization step, we aim at maximizing the expectation of the complete data likelihood $E(L^C)$,

$$E(L^C) = \sum_{i=1}^{m}\sum_{j=1}^{n} w(u_i,p_j)\sum_{k=1}^{l} P(z_k|u_i,p_j)\log Pr(u_i,p_j)$$

while taking into account the constraints, $\sum_{k=1}^{l} Pr(z_k) = 1$, on the factor probabilities, as well as the following constraints on the two conditional probabilities:

$$\sum_{k=1}^{l}(\sum_{i=1}^{m} Pr(u_i|z_k) - 1) = 0,$$

and

$$\sum_{k=1}^{l}(\sum_{j=1}^{n} Pr(p_j|z_k) - 1) = 0.$$

Through the use of Lagrange multipliers (see [16] for details), we can solve the constraint maximization problem to get the following equations for re-estimated parameters:

$$Pr(z_k) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w(u_i,p_j)Pr(z_k|u_i,p_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k'=1}^{l} w(u_i,p_j)Pr(z_{k'}|u_i,p_j)}$$

$$= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w(u_i,p_j)Pr(z_k|u_i,p_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} w(u_i,p_j)}$$

$$Pr(u_i|z_k) = \frac{\sum_{j=1}^{n} w(u_i,p_j)Pr(z_k|u_i,p_j)}{\sum_{i'=1}^{m}\sum_{j=1}^{n} w(u_{i'},p_j)Pr(z_k|u_{i'},p_j)}$$

$$Pr(p_j|z_k) = \frac{\sum_{i=1}^{m} w(u_i,p_j)Pr(z_k|u_i,p_j)}{\sum_{i=1}^{m}\sum_{j'=1}^{n} w(u_i,p_{j'})Pr(z_k|u_i,p_{j'})}.$$

Iterating the above computation of expectation and maximization steps monotonically increases the total likelihood of the observed data $L(U,P)$ until a local optimal solution is reached.

The computational complexity of this algorithm is $O(mnl)$, where m is the number of user sessions, n is the number of pageviews, and l is the number of factors. Since the usage observation matrix is, in general, very sparse, the memory requirements can be dramatically reduced using efficient sparse matrix representation of the data.

3. DISCOVERY AND ANALYSIS OF USAGE PATTERN WITH PLSA

One of the main advantages of PLSA model in Web usage mining is that it generates probabilities which quantify relationships between Web users and tasks, as well as Web pages and tasks. From these basic probabilities, using probabilistic inference, we can derive relationships among users, among pages, and between users and pages. Thus this framework provides a flexible approach to model a variety of types of usage patterns. In this section, we will describe various usage patterns that can be derived using the PLSA model.

As noted before, the PLSA model generates probabilities $Pr(z_k)$, which measures the probability of a certain task is chosen; $Pr(u_i|z_k)$, the probability of observing a user session given a certain task; and $Pr(p_j|z_k)$, the probability of a page being visited given a certain task. Applying Bayes' rule to these probabilities, we can generate the probability that a certain task is chosen given an observed user session:

$$Pr(z_k|u_i) = \frac{Pr(u_i|z_k)Pr(z_k)}{\sum_{d=1}^{l} Pr(u_i|z_d)Pr(z_d)}$$

and the probability that a certain task is chosen given an observed pageview:

$$Pr(z_k|p_j) = \frac{Pr(p_j|z_k)Pr(z_k)}{\sum_{d=1}^{l} Pr(p_j|z_d)Pr(z_d)}$$

In the following, we discuss how these models can be used to derive different kinds of usage patterns. We will provide several illustrative examples of such patterns, from real Web usage data, in Section 4.

3.1 Characterizing Tasks by Pageviews or by User Sessions

Capturing the tasks or objectives of Web users can help the analyst to better understand these users' preferences and interests. Our goal is to characterize each task, represented by a latent factor, in a way that is easy to interpret. One possible approach is to find the "prototypical" pages that are strongly associated with a given task, but that are not commonly identified as part of other tasks. We call each such page a *characteristic page* for the task, denoted by p_{ch}. This definition of "prototypical" has two consequences, first, given a task, a page which is seldom visited cannot be a good characteristic page for that task. Secondly, if a page is frequently visited as part of a certain task, but is also commonly visited in other tasks, the page is not a good characteristic page. So we define characteristic pages for a task z_k as the set of all pages, p_{ch}, which satisfy:

$$Pr(p_{ch}|z_k)Pr(z_k|p_{ch}) \geq \mu,$$

where μ is a predefined threshold.

By examining the characteristic pages of each task, we can obtain a better understanding of the nature of these tasks.

Characterizing tasks in this way can lead to several applications. For example, most Web sites allow users to search for relevant pages using keywords. If we also allow users to explicitly express their intended task(s) (via inputting task descriptions or choosing from a task list), we can return the characteristic pages for the specified task(s), which are likely to lead users directly to their objectives.

A similar approach can be used to identify "prototypical" user sessions for each task. We believe that a user session involving only one task can be considered as the characteristic session for the task. So, we define the characteristic user sessions, u_{ch}, for a task, z_k, as sessions which satisfy

$$Pr(u_{ch}|z_k)Pr(z_k|u_{ch}) \geq \mu,$$

where μ is a predefined threshold.

When a user selects a task, returning such exemplar sessions can provide a guide to the user for accomplishing the task more efficiently. This approach can also be used in the context of collaborative filtering to identify the closest neighbors to a user based on the tasks performed by that user during an active session.

3.2 User Segments Identification

Identifying Web user groups or segments is an important problem in Web usage mining. It helps Web site owners to understand and capture users' common interests and preferences. We can identify user segments in which users preform common or similar task, by making inferences based on the estimated conditional probabilities obtained in the learning phase.

For each task z_k, we choose all user sessions with probability $Pr(u_i|z_k)$ exceeding a certain threshold μ to get a session set C. Since each user sessions, $\vec{u}$, can also be represented as a pageview vector, we can further aggregate these users sessions into a single pageviews vector to facilitate interpretation. The algorithm of generating user segments is as follows:

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

$$\vec{v}_k = \frac{\sum \vec{u}_i \bullet Pr(u_i|z_k)}{|C|}.$$

4. For each factor z_k, output page vector $\vec{v}_k$. 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 weights so that the top items in the list correspond to the most frequently visited pages for the user segment.

These user segments provide an aggregate representation of all individual users' navigational activities in the a particular group. In addition to their usefulness in Web analytics, user segments also provide the basis for automatically generating item recommendations. Given an active user, we compare her activity to all user segments and find the most similar one. Then, we can recommend items (e.g., pages) with relatively high weights in the aggregate representation of the segment.

In Section 4, we conduct experimental evaluation of the user segments generated from two real Web sites.

3.3 Identifying the Underlying Tasks of a User Session

To better understand the preferences and interests of a single user, it is necessary to identify the underlying tasks performed by the user. The PLSA model provides a straightforward way to identify the underlying tasks in a given user session. This is done by examining $Pr(task|session)$, which is the probability of a task being performed, given the observation of a certain user session.

For a user session u, we select the top tasks z_k with the highest $Pr(z_k|u)$ values, as the primary task(s) performed by this user.

For a new user session, u_{new}, not appearing in the historical navigational data, we can adopt a "folding-in" method as introduced in [16] to generate $Pr(task|session)$ via the EM algorithm. In the E-step, we compute

$$Pr(z|u_{new}, p) = \frac{Pr(p|z)Pr(z|u_{new})}{\sum_{z'} Pr(p|z')Pr(z'|u_{new})},$$

and in the M-step, we fix $Pr(p|z)$ and only update $Pr(z|u_{new})$:

$$Pr(z|u_{new}) = \frac{\sum_{p'} w(u_{new}, p')Pr(z|u_{new}, p')}{\sum_{z'}\sum_{p'} w(u_{new}, p')Pr(z'|u_{new}, p')}.$$

Here, $w(u_{new}, p)$ represents the new user's visit frequency on the specified page p. After we generate these probabilities, we can use the same method to identify the primary tasks for the new user session.

The identification of the primary tasks contained in user sessions can lead to further analysis. For example, after identifying the tasks in all user sessions, each session u can be transformed into a higher-level representation,

$$\langle (z_1, w_1), \ldots, (z_l, w_K) \rangle$$

where z_i denotes task i and w_i denotes $Pr(z_i|u)$. This, in turn, would allow the discovery and analysis of task-level usage patterns, such as determining which tasks are likely to be visited together, or which tasks are most (least) popular, etc. Such higher-level patterns can help site owners better evaluate the Web site organization.

3.4 Integration of Usage Patterns with Web Content Information

Recent studies [23, 1, 9, 13] have emphasized the benefits of integrating semantic knowledge about the domain (e.g., from page content features, relational structure, or domain ontologies) in the Web usage mining process. 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. 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. The PLSA model described here also provides an ideal and uniform framework for integrating content and usage information.

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.

As before, we assume there exists a set of hidden factors $z \in Z = \{z_1, z_2, ..., z_l\}$, each of which represents a "semantic" group of pages. They can be a group of pages which have similar functionalities for users performing a certain task, or a group of pages which contain similar content information or semantic attributes. However, now, in addition to the set of pageviews, P, and the set of user sessions, U, we also specify a set of t semantic attributes, $A = \{a_1, a_2, \ldots, a_t\}$. To model the user-page observations, we use

$$P(u_i, p_j) = \sum_{k=1}^{l} Pr(z_k) \bullet Pr(u_i|z_k) \bullet Pr(p_j|z_k),$$

and to model the attribute-page observation, we use

$$P(a_q, p_j) = \sum_{k=1}^{l} Pr(z_k) \bullet Pr(a_q|z_k) \bullet Pr(p_j|z_k).$$

These models can then be combined based on the common component $Pr(p_j|z_k)$. This can be achieved by maximizing the following log-likelihood function with a predefined weight α.

$$L = \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha \log Pr(u_i, p_j) + \sum_{q=1}^{t}\sum_{j=1}^{n} (1-\alpha) \log Pr(a_q, p_j)$$

where α is used to adjust the relative weights of two observations. The EM algorithm can again be used to generate estimates for $Pr(z_k)$, $Pr(u_i|z_k)$, $Pr(p_j|z_k)$, and $Pr(a_q|z_k)$. By applying probabilistic inferences, we can measure the relationships among users, pages, and attributes, thus we are able to answer questions such as, "What are the most important attributes for a group of users?", or "Given an Web page with a specified set of attributes, will it be of interest to a given user?", and so on.

4. EXPERIMENTS WITH PLSA MODEL

In this section, we use two real data sets to perform experiments with our PLSA-based Web usage mining framework. We first provide several illustrative examples of characterizing users' tasks, as introduced in the previous section, and of identifying the primary tasks in an individual user session. We then perform two types of evaluations based on the generated user segments. First we evaluate individual user segments to determine the degree to which they represent activities of similar user. Secondly, we evaluate the effectiveness of these user segments in the context of generating automatic recommendations. In each case, we compare our approach with the standard clustering approach for the discovery of Web user segments.

In order to compare the clustering approach to the PLSA-based model, we adopt the algorithm presented in [24] for creating "aggregate profiles" based on session clusters. In the latter approach, first, we apply a multivariate clustering technique such as k-means to user-session data in order to obtain a set of user clusters $TC = \{c_1, c_2, ..., c_k\}$; then, an aggregate representation, pr_c, is generated for each cluster c as a set of pageview-weight pairs:

$$pr_c = \{\langle p, weight(p, pr_c)\rangle | p \in P, weight(p, pr_c) \geq \mu\}$$

where the significance weight, $weight(p, pr_c)$, is given by $weight(p, pr_c) = (1/|c|) \sum_{u \in c} w(p, u)$ and $w(p, u)$ is the weight of pageview p of the user session $u \in c$. Thus, each segment is represented as a vector in the pageview space. In the following discussion, by a user segment, we mean its aggregate representation as a pageview vector.

4.1 Data Sets

In our experiments, we use 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. This Web site provide various functionalities to different types of Web users. For example, prospective students can obtain program and admissions information or submit online applications. Current students can browse course information, register for courses, make appointments with faculty advisors, and log into the Intranet to do degree audits. Faculty can perform student advising functions online or interact with the faculty Intranet. After data preprocessing, we identified 21,299 user sessions (U) and 692 pageviews (P), with each user session consisting of at least 6 pageviews. This data set is referred to as the "CTI data."

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 to 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. During preprocessing, we recorded each user-property pair and the corresponding visit frequency. Finally, the data was filtered to limit the final data set to those users that had visited at least 3 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."

Each data set was randomly divided into multiple training and test sets to use with 10-fold cross-validation.

By conducting sensitivity analysis, we chose 30 factors in the case of CTI data and 15 factors for the Realty data. To avoid "overtraining", we implemented the "Tempered EM" algorithm [14] to train the PLSA model.

4.2 Examples Usage Patterns Based on the PLSA Models

Figure 1 depicts an example of the characteristic pages for a specific discovered task in the CTI data. The first 6 pages have the highest $Pr(page|task)*Pr(task|page)$ values, thus are considered as the characteristic pages of this task. Observing these characteristic pages, we may infer that this

Page Name	Pr(page|task)	Pr(task|page)	Pr(page|task)*Pr(task|page)
Online application-start	0.1075	1	0.1075
Online application-step1	0.105	1	0.105
Online application-step1	0.0949	1	0.0949
Online application-finish	0.0803	1	0.0803
Online application-submit	0.0339	1	0.0339
Online application-payment	0.0241	1	0.0241
...	...	...	...
/news/	0.0698	0.0131	1.00E-04

Figure 1: An example of the characteristic pages for the "Online Application" task in the CTI data

	ID	Price	Size	Rooms	Baths	Garage(cars)	Year	Style
Task 4	287	246900	3152	4	2.5	2	1992	2_story
	281	299900	4790	4	2.5	3	1989	2_story
	143	249500	1952	3	2.5	3	2001	1_story
	189	314900	2885	4	2.5	3	1998	1_story
	239	216900	1996	4	2.5	2	2001	2_story
	36	219500	2487	4	2	4	1999	1.5_story
Task 0	300	83500	1558	3	1.5	1	1920	bungalow
	302	85900	651	2	1	1	1949	ranch
	208	87500	1354	2	1	0	1971	1_story
	38	95900	1219	3	1	2	1962	split_level
	282	83900	1214	1	1	2	1925	1_story
	168	85000	1361	2	1.5	1	1920	bungalow
Task 5	19	121900	1480	4	1.5	2	1977	split_level
	166	164900	1616	3	1.5	2	1999	2_story
	171	142900	2003	3	1.5	2	1995	2_story
	246	121900	2051	4	1.5	2	1961	1.5_story
	294	142500	1334	3	1.5	2	1990	split_level

Figure 2: An example of the characteristic pages for three tasks in the Realty data

task corresponds to prospective students who are completing an online admissions application. Here "characteristic" has two implications. First, if a user wants to perform this task, he/she must visit these pages to accomplish his/her goal. Secondly, if we find a user session contains these pages, we can claim the user must have performed online application.

Some page may not be characteristic pages for the task, but may still be useful for the purpose of analysis. An example of such a page is the "/news/" page which has a relatively high $Pr(page|task)$ value, and a low $Pr(task|page)$ value. Indeed, by examining the at the site structure, we found that this page serves as a navigational page, and it can lead users to different sections of the site to perform different tasks (including the "online application"). This kind of discovery can help Web site designer to identify the functionalities of pages and reorganize Web pages to facilitate users' navigation.

Figure 2 identifies three tasks in the Realty data. In contrast to the CTI data, in this data set the tasks represent common real estate properties visited by users, thus reflecting user interests in similar properties. The similarities are clearly observed when property attributes are shown for each characteristic page. From the characteristic pages of each task, we infer that Task 4 represents users' interest in newer and more expensive properties, while Task 0 reflects interest in older and very low priced properties. Task 5 represents interest in properties midrange prices.

We can also identify "prototypical" users corresponding to specific tasks. An example of such a user session is depicted in Figure 3 corresponding to yet another task in the realty data which reflects interest in very high priced and large properties (task not shown here).

ID	Price	Size	Rooms	Baths	Garage(cars)	Year	Style
77	475000	5260	5	5	3	1995	2_story
189	314900	2885	4	3	3	1998	1_story
287	246900	3152	4	3	2	1992	2_story
320	279900	3298	4	4	3	1990	2_story
324	329000	2700	4	3	3	2002	2_story
327	330000	3338	5	4	3	1993	1_story
328	198900	2568	3	3	2	1999	1_story
329	264900	3320	5	4	2	1991	2_story
343	328000	3818	4	3	3	1999	2_story

Figure 3: An example of a "prototypical" user session

A real user session (pages listed in the order of being visited)	
admission mainpage	
Welcome information-Chinese version	
Admission info for international students	
Admission-requirements	
Admission-mail request	
Admission-orientation information	
Admission-F1visa and I20 information	
Application-statuscheck	
Online application-start	
Online application-step1	
Online application-step2	
Online application-finish	
Department mainpage	
Top probabilities of tasks given this user session *Pr(task|session)*	
Tasks	**Pr(task|session)**
Task 3	**0.4527**
Task 25	**0.3994**
Task 20	0.0489
Task 26	0.0458

Figure 4: An example of a identifying the prominent tasks within a given session

Our final example is this section shows how the prominent tasks contained in a given user session can be identified. Figure 4 depicts a random user session from CTI data. Here we only show the tasks IDs which have the highest probabilities $Pr(task|session)$. As indicated, the dominant tasks for this user session are Tasks 3 and 25. The former is, in fact, the "online application" task discussed earlier, and the latter is a task that represents international students who are considering applying for admissions. It can be easily observed that, indeed, this session seems to identify an international student who, after checking admission and visa requirements, has applied for admissions online.

4.3 Evaluation of User Segments and Recommendations

We used two metrics to evaluate the discovered user segments. The first is called the *Weighted Average Visit Percentage* (WAVP) [24]. WAVP allows us to evaluate each segment individually according to the likelihood that a user who visits any page in the segment will visit the rest of the pages in that segment during the same session. Specifically, 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. The weighted average similarity to the segment s over all transactions is then computed (taking both the transactions and the segments as vectors

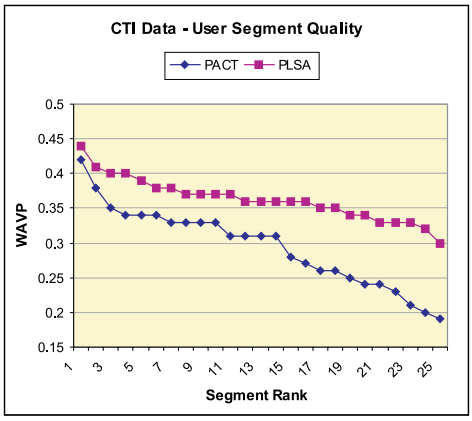

Figure 5: Comparison of user segments in the CTI site based on the Weighted Average Visit Percentage; PLSA model v. k-means clustering

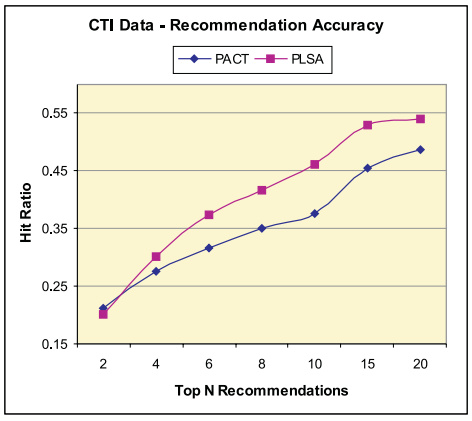

Figure 7: Accuracy of page recommendations based on PLSA segments versus k-means segments in the CTI site

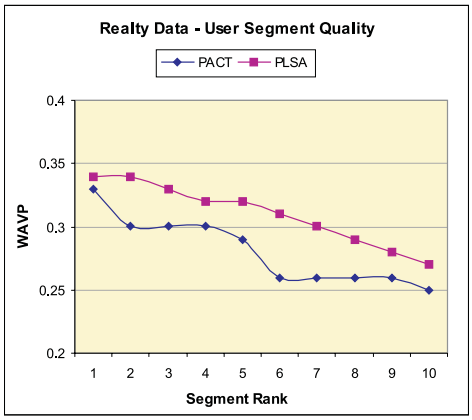

Figure 6: Comparison of user segments in the real estate site on the Weighted Average Visit Percentage; PLSA model v. k-means clustering

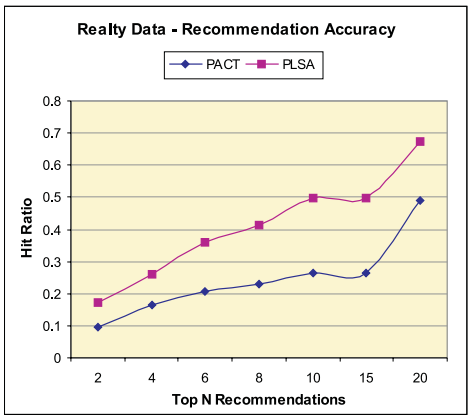

Figure 8: Accuracy of property recommendations based on PLSA segments versus k-means segments in the real estate site

of pageviews):

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

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

For evaluating the recommendation effectiveness, we use a metric called *Hit Ratio* in the context of top-N recommendation. For each user session in the test set, we took the first K pages as a representation of an active session to generate a top-N recommendation set. We then compared the recommendations with the pageview $(K+1)$ in the test session, with a match being 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 (between 1 and 20) that result in good hit ratios.

In the first set of experiments we compare the WAVP values for the generated segments using the PLSA model and those generated by the clustering approach. Figures 5 and 6 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 factor factors provides a significant advantage over the clustering approach.

In the second set of experiments we compared the recommendation accuracy of the PLSA model with that of k-means clustering segments. In each case, the recommendations are generated according to the recommendation algorithm presented in Section 3.2. The recommendation accuracy is measured based on hit ratio for different number of generated recommendations. These results are depicted in Figures 7 and 8 for the CTI and Realty data sets, respectively.

Again, the results show a clear 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. Indeed, this range of values seem to repre-

sent the largest improvements of the PLSA model over the clustering approach.

5. CONCLUSIONS AND FUTURE WORK

To understand Web users' preference and interests, it's necessary to develop techniques that can automatically characterize users' objectives (tasks) and discover the semantic relationships among users, users' tasks, and Web objects (Web pages). In this paper, we have developed a unified framework for the discovery and analysis of Web navigational patterns based on PLSA. We show the flexibility of this framework in characterizing various relationships among users, user tasks and Web objects. Since these relationships are measured in terms of probabilities, we are able to use probabilistic inference to perform a variety of analysis tasks such as task identification and user segmentation, as well as predictive tasks such as collaborative recommendations. We have demonstrated the effectiveness of our approach through experiments performed on two real-world data sets.

In our future work in this area, we plan to conduct more research on using the combined PLSA framework (as introduced in Section 3.4) to discover various usage patterns which involve users, pageviews, and semantic attributes, thus capturing users' preferences and interests at a deeper semantic level.

6. REFERENCES

[1] C. Anderson, P. Domingos, and D. Weld. 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)*, Edmonton, Alberta, Canada, July 2002.

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

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

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

[5] D. Cohn and H. Chang. Probabilistically identifying authoritative documents. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, CA, June 2000.

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

[7] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, Newport Beach, CA, November 1997.

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

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

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

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

[12] E. Gaussier, C. Goutte, K. Popat, and F. Chen. A hierarchical model for clustering and categorising documents. In *Advances in Information Retrieval – Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02)*, Glasgow, UK, March 2002.

[13] R. Ghani and A. Fano. 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*, Malaga, Spain, May 2002.

[14] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, July 1999.

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

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

[17] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical report, UC, Berkeley, Berkeley, CA, 1998.

[18] Y. Kim, J. Chang, and B. Zhang. a empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. In *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03)*, Seol, Koera, April 2003.

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

[20] N. Kushmerick, J. McKee, and F. Toolan. Towards zero-input personalization: Referrer-based page prediction. In P. Brusilovsky, O. Stock, and C. Strapparava, editors, *Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems International Conference (AH 2000)*, LNCS 1892, pages 133–143. Springer, 2000.

[21] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *Proceedings of the 1999 IEEE Knowledge and*

Data Engineering Exchange Workshop (KDEX'99), Chicago, Illinois, November 1999.

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

[23] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. 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, pages 165–176. Springer, September 2000.

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

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

[26] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In *Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, WI, July 1998.

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

[28] J. Pitkow and P. Pirolli. Mining longest repeating subsequences to predict www surfing. In *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems*, Boulder, Colorado, October 1999.

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

[30] S. Schechter, M. Krishnan, and M. D. Smith. Using path profiles to predict http requests. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, April 1998.

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

[32] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web usage analysis. *INFORMS Journal of Computing - Special Issue on Mining Web-Based Data for E-Business Applications*, 15(2), 2003.

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

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