Discovering User Access Pattern Based on Probabilistic Latent Factor Model

Guandong Xu, Yanchun Zhang, Jiangang Ma
School of Computer Science and Mathematics
Victoria University
PO Box 14428, VIC 8001, Australia
{xu,yzhang,ma}@csm.vu.edu.au

Xiaofang Zhou
School of Information Technology & Electrical Engineering
University of Queensland
Brisbane QLD 4072 Australia
zxf@itee.uq.edu.au

Abstract

There has been an increased demand for characterizing user access patterns using web mining techniques since the informative knowledge extracted from web server log files can not only offer benefits for web site structure improvement but also for better understanding of user navigational behavior. In this paper, we present a web usage mining method, which utilize web user usage and page linkage information to capture user access pattern based on Probabilistic Latent Semantic Analysis (PLSA) model. A specific probabilistic model analysis algorithm, EM algorithm, is applied to the integrated usage data to infer the latent semantic factors as well as generate user session clusters for revealing user access patterns. Experiments have been conducted on real world data set to validate the effectiveness of the proposed approach. The results have shown that the presented method is capable of characterizing the latent semantic factors and generating user profile in terms of weighted page vectors, which may reflect the common access interest exhibited by users among same session cluster.

Keywords: Web usage mining, web linkage information, user profile, probabilistic latent semantic model

1 Introduction

World Wide Web has become very popular recently and brought us a powerful platform to disseminate, retrieve information as well as conduct business. Generally, users are usually performing their interest-oriented actions by clicking or visiting one or more functional web items. They may exhibit different types of access interests associated with their tasks during their surfing period. For example, there may be many types of user groups with different interest involved in an E-commerce application. One type of users shows particular interest in browsing specific category goods such as sports products, while another just pays more attention to purchase the special products rather than specific category. In this manner, different clickstream of web pages will be recorded in the web log files. Thus, capturing the different web user access interest or pattern can, not only provide help for web site structural improvement, but also for better understanding common user navigational behavior from the same customer group. This, furthermore, can lead to recommend or predict tailored and personalized web contents to users, who have already exhibited similar navigational interests to one specific user group, and provide benefits of obtaining more preferred information and reducing waiting time as well.

In order to characterize access patterns from web server log files, many web usage mining techniques have been developed by researchers in a variety of application areas, and many works have been published to present their great success achieved in such fields as web personalization and recommendation systems (Lieberman 1995; Joachims, Freitag et al. 1997; Ngu and Wu 1997; Mobasher, Cooley et al. 1999), web system improvement (Cohen, Krishnamurthy et al. 1998), web site modification or redesign (Perkowitz and Etzioni 1998; Perkowitz and Etzioni 1999), and business intelligence and e-commerce (Buchner and Mulvenna 1998).

In the context of web usage mining, one important goal is to extract the informative knowledge from web user log file and identify the underlying user functional interests related to common navigational activities. With the benefit of great progress in data mining research community, many data mining techniques, such as web user or page clustering (Han, Karypis et al. 1998; Perkowitz and Etzioni 1998; Mobasher, Dai et al. 2002), association rule mining (Agrawal and Srikant 1994; Agarwal, Aggarwal et al. 1999) and sequential pattern mining technique (Agrawal and Srikant 1995) are adopted in current web usage mining methods and have been achieved great success as well. In most cases, web usage mining techniques are principally based on web pageview information, such as viewing time and frequency, visiting order etc., and conventional similarity measures are employed as well (Shahabi, Zarkesh et al. 1997). Common interest among Web users can be measured directly from these observation data using different similarity-based criteria (Xiao, Zhang et al. 2001). The web users, in turn, will be able to be classified into different categories depending on their different interests. Although these usage patterns can reveal the user access pattern explicitly, they, however, do not capture the intrinsic characteristics of Web user navigational activities, nor uncover the underlying and unobservable factors associated with specific usage goals exhibited from their web transaction histories. For example, such discovered usage patterns provide little knowledge of the underlying reasons why such web pages and user sessions are grouped together. Therefore, there has been an increased demand for developing techniques, that can extract the user underlying

access patterns and discover the latent semantic relationships among users as well as between users and web objects, to better understand those factors which are associated with common or specific navigational interests.

Latent Semantic Analysis (LSA) model is an approach to capture the latent or hidden semantic relationships among co-occurrence activities (Baeza-Yates and Ribeiro-Neto 1999). Generally, in order to reveal such deeply latent information, the relationships between co-occurrence objects are first mapped into a high dimensional matrix which is usually called as latent semantic space. Then, a dimensional reducing algorithm such as Single Value Decomposition (SVD) or Primary Component Analysis (PCA) algorithm, is applied to perform a linear projection of the original relationship space to generate a reduced latent space (Deerwester, Dumais et al. 1990; Dumais 1995; Baeza-Yates and Ribeiro-Neto 1999). LSA has been widely used in Web information indexing and retrieval applications (Deerwester, Dumais et al. 1990; Berry, Dumais et al. 1995), for example, by applying LSA method, the latent semantic model can be discovered from web linkage information, which will lead to find relevant web pages and improve web searching efficiency and effectivity (Hou and Zhang 2002; Hou and Zhang 2003). In addition, Factor analysis technique has been used in web mining research area as an alternative LSI model recently. Based on Principal Factor Analysis (PFA) model especially, a well-known statistical analysis model, the latent factors associated with user access interests can be derived from the usage data, and used for clustering users or web pages (Zhou, Jin et al. 2004).

Although LSA has achieved great success in some applications, it still has some shortcomings due to its unsatisfactory statistical foundation (Hofmann 1999). Probabilistic latent semantic analysis (PLSA) model is a probabilistic variant of LSA just as its name indicates. Although they share similar concept of latent analysis, but there still exist distinct differences between them on the level of theoretical basis. Due to the sound solid foundation PLSA model may overcome the conventional LSI algorithm such as SVD in some particular applications. Recently, approaches based on PLSA have been successfully applied in collaborative filtering (Hofmann 2004) web usage and content mining (Jin, Zhou et al. 2004), text learning and mining (Cohn and Chang 2000; Hofmann 2001), co-citation analysis (Cohn and Hofmann 2001) and related topics.

In this paper, we present a method for discovering user access patterns and semantic usage factors based on probabilistic latent semantic analysis (PLSA) model. From web log files or user clickstream records, web user navigational observation data are collected and user sessions are generated in terms of pageview-weight pairs. In addition, we consider the web linkage information derived from web site map as web structural data and utilize web page transitivity to enhance data preprocessing as well. The combination usage data with linkage information may provide benefits for supporting usage mining and improving scalability of user access pattern discovery. On the other hand, the probabilistic model is constructed to infer the latent semantic factors and user access patterns from the integrated usage data. In order to derive the user access patterns from web access observation data efficiently, a specific probabilistic model analytical algorithm, i.e. EM algorithm, is employed. Based on the discovered probabilistic information, we propose a web user clustering algorithm to classify user sessions into different web user groups, which reflect similar interests during their visiting histories, according to the calculated conditional probability distribution. We, in turn, generate the aggregated user profiles in terms of weighted pages for representing user access pattern. Meanwhile, the latent usage factors are uncovered by clustering web pages and characterizing "theme" from the page clusters. Furthermore, the discovered usage patterns can be provided for making web recommendation to an active user or helping user to personalize his/her needs. The experiments on real world data set are implemented to evaluate the proposed method. The experimental results have shown the scalability of inferring latent usage factors, generating user access patterns or user profiles as well.

The rest of the paper is organized as follow: firstly, the data preparation process is discussed in the following section, particularly focused on how to integrate the usage data with linkage data information. In Section 3, we introduce the principle of probabilistic latent semantic analysis model and propose a probabilistic model analysis algorithm (i.e. EM algorithm) based on web usage data and linkage information. Section 4 presents the algorithms for clustering Web user sessions and producing aggregated user profiles. The preliminary experimental results conducted on real data set to validate the proposed probabilistic model as well as interpret the discovered user access patterns are given in Section 5. Finally, Conclusion and some future work are discussed in Section 6.

2 Data preparation and integration

In web usage mining application, data preprocessing stage is essential and crucial for discovering user access pattern due to the demand for eliminating "noise" data such as pages with too low visited frequency or sessions with too short size existing in web log files and generating sectionalized usage data. Generally, three steps will be involved in the process of web usage mining (Cooley, Mobasher et al. 1999). They are, namely, data preprocessing, access pattern extraction, and pattern analysis. In first stage, raw data in web log files are preprocessed and transformed into transaction data that can be processed for the purpose of usage mining in next step. Secondly, a variety of data mining techniques, such as web user or page clustering, association rule mining, and sequential pattern discovery can be employed to the transformed data generated in first stage to extract the web user access patterns. The mined patterns should then be analyzed and prepared for further pattern analysis applications, e.g. web page recommendation by matching the discovered pattern with the active user transaction. The primary data source used in web usage mining is access log files, which record the user access history to web server or application server. In addition, other types of data sources, such as content data, structure data as well as user data, are also needed to integrate for data preparation and

pattern discovery (Cooley, Mobasher et al. 1999; Srivastava, Cooley et al. 2000; Jin, Zhou et al. 2004)

2.1 Usage data

In web usage data preprocessing stage, the goal is to generate the user session data in terms of page views. Basically, according to W3C definition, a web view can be viewed as a visual rendering of a Web page. In this way, the user access interest exhibited may be reflected by the varying degree of visits in different web pages during one session. Thus, we can represent a user session as a collection of page visits in the period of time interval. In other words, the user sessions can be expressed in terms of pageview vectors. From such viewing point, we can generate the following user session expression. Given n web pages in a web site and m web users visiting the web site during a period of time, after data preprocessing, we can built up a set of n pages as $P = \{p_1, p_2, ..., p_n\}$ and a set of m user sessions as $S = \{s_1, s_2, \ldots, s_m\}$. Alternatively, these kinds of data transforms can also be called as page identification and user sessionization process respectively (Cooley, Mobasher et al. 1999). Conceptually, modelling of user session in a collection of pages defined by the so-called web page dictionary, which consists of all items visited by whole users, is similar to modelling a document in terms of word frequencies by using a word dictionary in text IR. In other words, each user session can be, in turn, expressed as a set of weight- pageview pairs, $s_i = \{<p_1,a_{i1}>, <p_2,a_{i2}>, \ldots <p_n,a_{in}>\}$. By simplifying the above expression in terms of pageview vectors, each user session can be considered as an n-dimensional vector $s_i = \{a_{i1}, a_{i2}\ldots, a_{in}\}$, where a_{ij} denotes the weight for pageview p_j in s_i user session.

1) Main Movies: 20sec *Movies* News: 15sec *News*Box: 43sec *Box-Office* Evita: 52sec *News* Argentina:31 sec *Evita: 44sec*

2) Music Box: llsec *Box-Office* Crucible: 12sec *Crucible* Book: 13sec *Books: 19sec*

3) Main Movies: 33sec *Movies* Box: 21sec *Boxoffice* Evita: 44sec *News* Box: 53sec *Box-office* Evita: 61 sec *Evita : 31sec*

4) Main Movies: 19sec *Movies* News: 21sec *News b*ox: 38sec Box-*Office* Evita:61 sec *News* Evita:24sec *Evita* News: 31 sec *News* Argentina: 19sec *Evita: 39sec*

5) Movies Box: 32sec *Box-Office* News: 17sec *News* Jordan: 64sec *Box-Office* Evita: 19sec *Evita: 50sec*

6) Main Box: 17sec *Box-Office* Evita: 33sec *News* Box: 41 sec *Box-Office* Evita: 54sec *Evita* News: 56sec *News: 47sec*

$$SP_{ex} = \begin{bmatrix} 9.76 & 7.32 & 36.1 & 25.4 & 21.5 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 21.8 & 0.00 & 20.0 & 23.6 & 34.6 \\ 13.6 & 8.64 & 21.8 & 43.2 & 12.8 & 0.00 & 0.00 & 0.00 \\ 7.54 & 8.33 & 32.1 & 34.2 & 27.8 & 0.00 & 0.00 & 0.00 \\ 0.00 & 17.6 & 35.2 & 19.8 & 27.5 & 0.00 & 0.00 & 0.00 \\ 6.85 & 0.00 & 35.5 & 35.1 & 22.6 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

Figure 1: A usage snapshot and its session-page matrix expression

As a result, the whole user session data can be utilized to form web usage data represented by a session-pageview matrix $SP_{m\times n}=\{a_{ij}\}$. The element value in the session-page matrix, a_{ij}, can be represented by a weight associated with the page p_j in the user session s_i, which is usually determined by the number of hit or the amount time spent on the specific page. Generally, the weight a_{ij} associated with page p_j in the session s_i should be normalized across pageviews in same user session in order to eliminate the influence caused by the relative amount difference of visiting time durations or hit numbers. The so-called session normalization implementation is capable of capturing the relative significance of a page within one user session with respect to others pages accessed by same user. The figure 1 illustrates an example of usage data snapshot, in which the names (in *italic*) are the titles of web pages followed by the links and corresponding link times (underlined), and its corresponding session-page matrix in terms of normalized weight forms from (Shahabi, Zarkesh et al. 1997; Xiao, Zhang et al. 2001) as well.

2.2 Linkage data

On the other hand, the structure data is another important data source representing the latent relationships among the web pages for web mining. The structure data, represented through hyperlink by inter-page or intra-page linkage structure among pages often reflects the content organization relationship within the web site developed by the designer (Cooley, Mobasher et al. 1999). In most cases, linkage (or hyperlink) can reveal semantic information between web pages due to the fact that the web designer always want to create links to other pages which are considered to be relevant to the linking ones. If a hyperlink is reasonable, it may more or less reveal their mutual semantic relationship among the web pages. In this manner, if two users visited two web pages with linked or linking information each other, they could be viewed to be exhibiting similar interests during their visiting website. For example, in a virtual e-commerce web site *http://www.virtualecom.com*, two users have visited two different pages with respect to distinct sport products within the website: "*www.virtualecom.com/sportproducts/nike.htm*" and "*www.virtualecom.com/sportproducts/adidas.html*" respectively. Although these two web pages are quite different from their titles, we can find that they indeed have common semantic relationship in the contents, which are all considered to be associated with sport products category, if they are linked or linking each other directly. In short, such co-citation analysis in hyperlink can measure web page similarity in content. Similar to finding relevant pages in web searching, web user access interests can also be captured partly from the mutual linkage information. Taking linkage information into account will result in improving the scalability of discovering user access pattern rather than using usage data standalone during web usage mining. Based on the above discussion, these two users in above example are likely to be all considered as being interested in "sport goods" and should be classified into one user category that showing similar access pattern with interests on "sport goods". Besides that web usage analysis can discover the user access pattern directly from web log files, web page linkage is capable of utilizing to reveal the intrinsic relationship among pages, which will lead to assist clustering user sessions more reasonable.

Similar to the work in (Hou and Zhang 2003), we can utilize hyperlink transitivity expression (i.e. correlation matrix) as linkage information, and take it as web structural data to integrate into the usage data expression. In the context of hyperlink transitivity, the linkage information between pages is expressed by correlation matrix. For measuring the page correlation, we firstly introduce the following definitions adopted from (Hou and Zhang 2003).

Definition 1: If there is a direct link from page A to page B, then the *length of path* from page A to page B is 1, denoted as $l(A,B) = 1$. If page A has a link to page B via n other pages, then $l(A,B) = n+1$. The *distance* from page A to page B, denoted as $sl(A,B)$, is the shortest path length from A to B, i.e. $sl(A,B) = min(l(A,B))$. The length of path from a page to itself is zero, i.e. $l(A,A) = 0$. If there are no links from page A to page B (direct or indirect), then $l(A,B) = \infty$.

Definition 2: *Correlation factor*, denoted as F, $0<F<1$, is a constant that measures the correlation coefficient between two page with direct link, i.e. if page A has a direct link to page B, then the correlation rate from page A to page B is F.

Definition 3: The *correlation* degree from page i to page j, denoted as c_{ij}, is defined as

$$c_{ij} = F^{sl(i,j)}$$

where F is the correlation factor for all page pairs, $sl(i,j)$ is the distance from page i to page j.

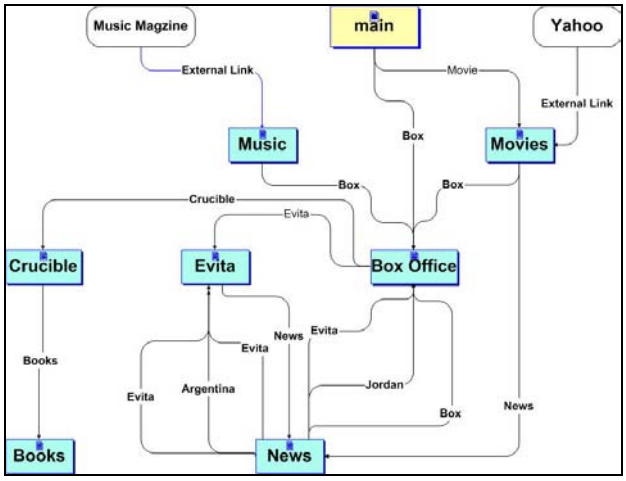

Figure 2: A simplified example of website graph

$$\begin{pmatrix} 1 & F & F^3 & F & F^2 & 0 & F^2 & F^3 \\ 0 & 1 & F^2 & F & F^2 & 0 & F^2 & F^3 \\ & 0 & 1 & F & F & 0 & F^2 & F^3 \\ 0 & 0 & F^2 & 1 & F & 0 & F & F^2 \\ 0 & 0 & F & F^2 & 1 & 0 & F^3 & F^4 \\ 0 & 0 & F^3 & F & F^2 & 1 & F^2 & F^3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & F \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 3: The corresponding correlation matrix

Upon the above definitions, we can find that the correlation degree is determined by the correlation factor F, the distance between the two pages, and the correlation matrix can represent the relevant relationship between different pages. By referring the work in (WEISS, VÉLEZ et al. 1996), we can limit F in the extent of 1/4 - 1/2 for the following linkage integration implementation. Figure 2 depicts an example of simplified website graph in (Shahabi, Zarkesh et al. 1997). The following corresponding correlation matrix (in figure 3) illustrates transitivity relationships between pages, which represent the structural information within the website in a certain extent.

2.3 Usage and linkage data integration

As we discussed above, the integration of usage, linkage and other source data is crucial for providing the ability to analyze and reason about the discovered patterns, derive more actionable knowledge, and produce reasonable recommendation further. Conceptually, the collected web usage and linkage data can be viewed as m×n session-pageview matrix in which each row is a weighted n-dimensional vector over the pageview space, and n×n page-correlation matrix which represents the correlation between pages as well, respectively. A direct approach for integration usage and linkage data for web mining task is to transform original user sessions into hyperlink-enhanced sessions containing semantic features of the underlying pageviews by multiplying the session-pageview matrix with correlation matrix. This transform will result in a new matrix $SP' = \{s_1', s_2', \ldots, s_m'\}$ where each S_i' is an n-dimensional vector over the pageview space. The transformed hyperlink-enhanced session-pageview matrix, thus, can combine not only the usage information but also the web linkage information, which is helpful to discover the underlying user access patterns.

In order to discuss how the integrated usage data can reflect the common visit interests from mutual linking, let's consider the simplest example mentioned in section 2.2, describing the two distinct web pages with mutual linking visited by two users. The two corresponding sessions can be expressed as two simple page vectors in expression of binary weight: $s_1=(1,0)$ and $s_2=(0,1)$. By simply applying the well-known cosine coefficient to measure the similarity of two sessions, it is found that these two sessions have no common interest due to their similarity equals to 0. On the other hand, from the correlation matrix, based on the linkage information, we can obtain the linkage-enhanced session-pageview matrix by multiplying the original matrix with correlation matrix as:

$$SP' = SP \times CR = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & F \\ F & 1 \end{pmatrix} = \begin{pmatrix} 1 & F \\ F & 1 \end{pmatrix}$$

Analogously, we can also determine the cosine similarity between s_1 and s_2 to be 0.45 when F is considered to be 1/2 in above equation. The increased similarity value means the greater common visiting interest exhibited by the two users. In other words, the integrated usage data can reveal the deeply navigational behavior from user usage data and web site linkage information, which will lead to classify user sessions and generate user profiles efficiently.

3 Probabilistic latent semantic analysis model (PLSA)

The PLSA model has been firstly presented and successfully applied in text mining by (Hofmann 1999). As discussed above, PLSA is based on maximum likelihood principle, which is derived from statistic principle, while LSA utilizes the L_2 or Frobenius norm as an optimization criterion.

Basically, the PLSA model is based on a statistic model called aspect model, which can be utilized to identify the hidden semantic relationships among general co-occurrence activities. Similarly, we can conceptually view the user sessions over web pages space as co-occurrence activities in the context of web usage mining to infer the latent usage pattern. Given the aspect model over user access pattern in the context of web mining, it is first assumed that there is a latent factor space $Z=\{z_1,z_2,...,z_l\}$ and each co-occurrence observation data (s_i, p_j) (i.e. the visit of page p_j in user session s_i) is associated with the factor $z_k \in Z$ by varying degree to z_k. According to the viewpoint of aspect model, it can be inferred that there do exist different relationships among web users or pages corresponding to different factors. Furthermore, the different factors can be considered to represent the corresponding user access pattern. For example, during the web usage mining process on e-commerce website, we can define that there exist k latent factors associated with the k kinds of navigational behavior patterns, such as z_1 factor standing for having interests in sports-specific product category, z_2 for sale product interest and z_3 for browsing through a variety of product pages in different categories and z_4 … etc,. In this manner, each co-occurrence observation data (s_i, p_j) may convey the user navigational interest by mapping the observation data into the k-dimensional latent factor space. The degree to which such relationships are "explained" by each factors is derived from the conditional probabilistic factors associated with web usage data. The goal of PLSA is to determine the conditional probabilities, in turn, to reveal intrinsic relationships among web users or pages based on the computed semantic probabilities. In our work, the PLSA model is constructed to infer user navigational behavior in a semantic web-space and improve the availability and scalability of latent semantic analysis based on the integrated usage data rather than usage data standalone.

Firstly, let's introduce the following probability definitions:

- $P(s_i)$ denotes the probability that a particular user session s_i will be observed in the occurrences data,
- $P(z_k|s_i)$ denotes a user session-specific probability distribution on the unobserved class factor z_k explained above,
- $P(p_j|z_k)$ denotes the class-conditional probability distribution of pages over a specific latent variable z_k.

Based on these definitions, the probabilistic latent semantic model can be expressed in following way:

- Select a user session s_i with probability $P(s_i)$,
- Pick a hidden factor z_k with probability $P(z_k|s_i)$,
- Generate a page p_j with probability $P(p_j|z_k)$;

As a result, we can obtain probability of an observed pair (s_i, p_j) by adopting the latent factor variable z_k. Translating this process into a probability model results in the expression:

$$P(s_i, p_j) = P(s_i) \bullet P(p_j \mid s_i) \tag{1}$$

where,

$$P(p_j \mid s_i) = \sum_{z \in Z} P(p_j \mid z) \bullet P(z \mid s_i) \tag{2}$$

By applying Bayes's rule, a re-parameterized version will be transformed based on (1) and (2) as

$$P(s_i, p_j) = \sum_{z \in Z} P(z) \bullet P(s_i \mid z) \bullet P(p_j \mid z) \tag{3}$$

Following the likelihood principle, we can determine the total likelihood Li as

$$Li = \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet \log P(s_i, p_j) \tag{4}$$

where $m(s_i,p_j)$ corresponds to the entry of the linkage-enhanced session-pageview matrix associated with session s_i and pageview p_j.

In order to maximize the total likelihood, we need to generate repeatedly the conditional probabilities $P(z)$, $P(s_i|z)$, $P(p_j|z)$ by utilizing the usage observation data. By the knowledge of statistics, we are told that Expectation Maximization (EM) algorithm is a standard procedure to perform maximum likelihood estimation in latent variable model (Dempster, Laird et al. 1977). Generally, two steps are needed to implement in the procedure alternately: (1) Expectation (E) step where posterior probabilities are calculated for the latent factors based on the current estimates of conditional probability, and (2) Maximization (M) step, where the estimated conditional probabilities are updated and used to maximize the likelihood based on the posterior probabilities computed in the previous E step.

We now discuss the whole procedure in details:

(1) Firstly, for given the randomized initial values of $P(z_k)$, $P(s_i|z_k)$, $P(p_j|z_k)$

(2) Then, in the E-step, we can simply apply Bayes' formula to generate following variable based on usage observation:

$$P(z_k \mid s_i, p_j) = \frac{P(z_k) \bullet P(s_i \mid z_k) \bullet P(p_j \mid z_k)}{\sum_{z_k \in Z} P(z_k) \bullet P(s_i \mid z_k) \bullet P(p_j \mid z_k)} \tag{5}$$

(3) Furthermore, in M-step, we can compute:

$$P(p_j \mid z_k) = \frac{\sum_{s_i \in S} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j)}{\sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s\ , p_j)} \tag{6}$$

$$P(u_i \mid z_k) = \frac{\sum_{p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j)}{\sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j)} \tag{7}$$

$$P(z_k) = \frac{1}{R} \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j) \tag{8}$$

where

$$R = \sum_{s_i \in S, p_j \in P} m(s_{i,}, p_j), \quad (9)$$

Basically, substituting equation (6)-(8) into (3) and (4) will result in the monotonically increasing of total likelihood L_i of the observation data. The iterating implement of E-step and M-step is repeating until L_i is converging to a local optimal limit, which means the calculated results can represent the final probabilities of observation data. The computational cost and efficiency analysis may be referred in (Hofmann 1999).

By now, we have obtained the conditional probability distribution estimates $P(z_k)$, $P(s_i|z_k)$, $P(p_j|z_k)$ by performing the E and M step alternatively. The estimated probability distribution which is corresponding to local maximum likelihood *Li* contains the useful information for inferring semantic usage factors, performing web user sessions clustering and generating the aggregated user profiles in next section.

4 Clustering User session, generating user profile and characterizing latent factor algorithm

As we discussed in section 2, we note that each latent factor z_k do really represent specific aspect associated with co-occurrence in nature. In other words, for each factor, there is a task-oriented user access pattern associated with it. We, thus, can utilize the class-conditional probability estimates generated by the PLSA model to produce aggregated user profiles for characterizing user navigational behaviours. Conceptually, each aggregated user profile will be expressed as a collection of pages, which are accompanied by their corresponding weights indicating corresponding contributions to such user group by those pages. Furthermore, analyzing the generated user profile can lead to reveal common user access interests, such as dominant or secondary "themes" by sorting the page weights.

4.1 Clustering User session

Firstly, we begin with the probabilistic variable $P(z_k|s_i)$, which represents the probability of a latent class factor z_k exhibited by a given user session s_i. Basically, this probabilistic distribution over the factor space of a user can reflect the specific user access tendency over the whole latent factor space, therefore, may be utilized to uncover the dominant factors by distinguishing the top probability values.

For each user session s_i, we can compute a set of probabilities $P(z_k|s_i)$ corresponding to different latent class factors as following:

$$P(z_k \mid u_i) = \frac{P(u_i \mid z_k) \bullet P(z_k)}{\sum_{z_k \in Z} P(u_i \mid z_k) \bullet P(z_k)} \quad (10)$$

Actually, the set of probabilities $P(z_k|s_i)$ is tending to be "sparse", for a given s_i typically only few entries are significant different from predefined threshold. We can classify the user into corresponding cluster based on these probabilities greater than the given threshold. Since each user session can be expressed as a pages vector in the original n-dimensional space, we can create a mixing representation of the collection of user sessions associated with the factor z_k in terms of a collection of weighted pageviews. The algorithm for clustering user session is described as following.

Algorithm 1 Clustering user session

Input: $P(z_k|s_i)$, user session-page matrix SP_{ij}, threshold μ.
Output: A set of clusters $SCL=(SCL_1, SCL_2, \ldots, SCL_K)$
Begin

Step 1: $SCL_1=SCL_2=\ldots=SCL_k=\Phi$
Step 2: For each $s_i \in S$, select $P(z_k|s_i)$, if $P(z_k|s_i) \geqslant \mu$, then $SCL_k=SCL_k \cup s_i$
Step 3: If there are still users sessions to be clustered, go back to step 2
Step 4: Return clusters $SCL=\{SCL_k\}$

Algorithm 2 generating user profiles

Input: session cluster set $SCL=\{SCL_k\}$
Output: user profiles $PF=\{\overrightarrow{PF_k}\}$

Step 1: for each factor z_k, choose all candidate sessions in SCL_k
Step 2: represent each session $\vec{s_i}$ as a pageview vector and compute their *centroid* pageview vector as:

$$\overrightarrow{PF_k} = \frac{\sum_i \vec{s_i} \bullet P(z_k \mid s_i)}{|R|} \quad (11)$$

where |R| denotes the total number of session in the cluster
Step 3: if there are still user session clusters not to be processed, go back to step1
Step 4; output the *centroid* pageview vector as the aggregated user profile corresponding to each factor z_k

By now, we classify the user sessions into corresponding clusters which can be considered to represent user navigational behaviours based on the calculated conditional probability distributions from PLSA model and characterize the representations of the user profiles in terms of weighted page vectors as well. As discussed above, it can be seen that a particular user session does not belong to just one, but several different groups associated with different latent factors. For example, a user session may exhibit different interests (with different probabilities) on two aspects z_1 and z_2. This can be "explained" as that a user may, indeed, perform different tasks during the same session and really reflect the user access pattern in real world. It can be implied, in turn, the PLSA model partitions user session-page pairs, which is different from clustering either users sessions or pages or both. In other words, the user session-page probabilities in PLSA model reflect "overlay" of latent factors, while the conventional clustering model assumes there is just one cluster-specific distribution contributed by all user sessions in the cluster (Hofmann 1999).

4.2 Characterizing the factor

As mentioned in section 3.1, the core of PLSA model is hidden factor space. From this point of view, how to characterize the factor space or explain the semantic meaning of factors is the crucial issue in PLSA model. Similarly, we can also utilize PLSA model to identify the

semantic meaning of factors by clustering the web pages into corresponding categories associated with the latent factors based on another derived probability distribution $P(p_j|z_k)$.

For each hidden factor z_k, we may consider that the pages whose conditional probabilities $P(p_j|z_k)$ are greater than a predefined threshold can be viewed to provide similar functional operation corresponding to the latent factor. In this way, we can select all pages with probabilities exceeding a certain threshold to form an aspect-specific cluster. By analyzing the pages and their weights derived from the conditional probabilities, which are associated with the specific factor, we may characterize and explain the semantic meaning of each factor. In section 4, we will present examples with respect to the discovered latent factors. The algorithm to generate the aspect-oriented web page cluster is briefly described as following:

Algorithm 3 clustering web pages

1. Input: $P(p_j|z_k)$, predefined threshold μ
2. For each z_k, choose all pages with $P(p_j|z_k) \geq \mu$, construct $PCL_k=\{p_j| P(p_j|z_k) \geq \mu\}$
3. If there are still pages to be classified, go back to step 2
4. Output: $PCL=\{PCL_k\}$

5 Preliminary experimental evaluation

In order to evaluate the effectiveness of the proposed method based on PLSA model and explore the discovered latent semantic factors, we have conducted preliminary experiments on real world data set.

5.1 Data set

The data set we used is downloaded from KDDCUP website (http://www.ecn.purdue.edu/KDDCUP/), which is provided for a yearly competition in data mining started in 1997. The data set is common-used data resource to test and compare prediction algorithm, clustering approaches for data mining purpose. Data preprocessing has been performed on the raw data set since there are some short user sessions existing in the data set, which mean they are of less contribution for the purpose of data mining. Data filtering technique is used to eliminate these user sessions, leaving only sessions with at least 4 pages. After data preparation, we have setup an evaluation data set including 9308 user sessions and 69 pageviews, where the average session length is 11.88 pageviews. The linkage information is derived from the web site map and integrated into the usage data to generate the linkage-enhanced usage data for web usage mining. In this data set, the entries in session-pageview matrix are determined by the numbers of web page hits since the numbers of a user coming back to a specific page is a good measure to reflect the user interest in the page.

By considering the number of web pages and the content of the web site carefully and referring the selection criteria of factors in (Hofmann 2001; Jin, Zhou et al. 2004), we choose 13 factors for experiment.

5.2 Examples of latent factors

We conduct the experiments on the data set to extract the latent factors and generate user profiles. Firstly, we present the example of latent factors derived from real data set by using proposed PLSA model.

Table 1: latent factors and their characteristic titles

Factor #	Characteristic title
1	Department_search_results
2	ProductDetailLegwear
3	Vendor_ service
4	Freegift
5	ProductDetailLegcare
6	Shopping_cart
7	Online_shopping
8	Lifestyle_assortment
9	Assortment2
10	Boutique
11	Departmet_replenishment
12	Department_article
13	Home page

Table 2: factor examples from KDDCUP data set

Factor	Page	$P(p_j\|z_k)$	Page description
#3	10	0.865	main/vendor\.jhtml
	36	0.035	main/cust_serv\.jhtml
	37	0.021	articles/dpt_contact\.jhtml
	39	0.020	articles/dpt_shipping\.jhtml
	38	0.016	articles/dpt_payment\.jhtml
	41	1.58	articles/dpt_faqs\.jhtml
	40	1.32	articles/dpt_returns\.jhtml
#7	27	0.249	main/login2\.jhtml
	44	0.18	checkout/expresCheckout.jhmt
	32	0.141	main/registration\.jhtml
	65	0.135	main/welcome\.jhtml
	45	0.135	checkout/confirm_order\.jhtml
	42	0.045	account/your_account\.jhtml
	60	0.040	checkout/thankyou\.jhtml
#13	12	0.232	articles/dpt_about\.jhtml
	22	0.127	articles/new_shipping\.jhtml
	13	0.087	articles/dpt_about_mgmtteam
	14	0.058	articles\dpt_about_boardofdirectors
	20	0.058	articles/dpt_affiliate\.jhtml
	16	0.053	articles/dpt_about_careers
	19	0.052	articles/dpt_refer\.jhtml
	23	0.051	articles/new_returns\.jhtml

Table 1 lists the 13 extracted latent factors and their corresponding characteristics. Furthermore, Table 2 depicts 3 factor examples selected from whole factor space in terms of detailed page information of them. In this table, factor #3 indicates the concerns about vendor service message such as customer service, contact number, payment methods as well as delivery support. The factor #7 describes the specific progress which may include customer login, product order, express checkout and financial information input such steps occurred in internet shopping scenario, whereas factors #13 actually captures another character exhibited by web content,

which reveals the fact that some web users may pay more attentions to the information regarding department itself.

5.3 Examples of user profiles

Furthermore, we can generate the user profiles according to the session-specified conditional probabilities $P(z_k|s_i)$ which represent the common visit interests/ access patterns of users within the session cluster. Generally, the aggregated user access profile is expressed as a collection of web pages ranked by their associated weights, which can reflect the contributing to the profile by corresponding pages. Table 3 presents the generated user profile. For each profile, the pages are listed in a sequence ordered by their associated significance in terms of probabilistic values. It may be inferred that the greater weight of a page possesses, the more significance contributed by the page exhibits. In the other words, it is more likely to be visited by users within same user session cluster, which reflects similar visiting interest or common navigational behavior. For example, in table 3, user profile #7 represents the online-shopping activities in details, especially occurring in purchasing leg-wear products or fashion clothes, whereas user profile #13 reflects one kind of customers' concern focused on the information with regard to the department store itself. Such explanations of user profiles drawn from table 3 are consistent with the discovery from table 2.

Table 3: Example of user access profile

	Page #	Weight	Page description
Profile #7	4	1.20E-4	products/productDetailLegwear
	29	8.44E-5	main/shopping_cart
	27	6.50E-5	main/login2
	8	5.20E-5	main/home
	44	5.03E-5	checkout/expresCheckout
	65	3.86E-5	main/welcome
	2	3.77E-5	main/boutique
	7	3.75E-5	main/search_results
	6	3.75E-5	main/departments
	45	3.72E-5	checkout/confirm_order
	32	3.57E-5	main/registration
	42	1.65E-5	account/your_account
Profile #13	12	1.73E-4	articles/dpt_about
	8	9.77E-5	main/home
	13	7.17E-5	articles/dpt_about_mgmtteam
	4	5.91E-5	products/productDetailLegwear
	14	5.17E-5	articles\dpt_about_boardofdirectors
	2	4.85E-5	main/boutique
	16	4.79E-5	articles/dpt_about_careers
	22	4.55E-5	articles/new_shipping
	17	4.26E-5	articles/dpt_about_investor
	18	4.14E-5	dpt_about_pressreleases
	15	3.83E-5	dpt_about_healthwellness
	20	3.78E-5	articles/dpt_affiliate

In addition, from the discovered user profiles, we can conclude there are more than one kinds of access interests or exist "overlapping" of visiting trends involved in one user profile. But we may still distinguish the dominant or secondary pattern from others based on the corresponding weights associated with web contents (i.e. web page). Furthermore, with the discovered user access profiles, such patterns can provide benefits for further web analysis applications, for example, web recommendation or prediction.

6 Conclusion and future work

In this paper, we have presented a probabilistic latent semantic analysis (PLSA) model, which can infer the hidden semantic factors and uncover user access patterns from the session-page observation data. The data from two different sources, namely, web access log files (i.e. usage data) and web site map (i.e. linkage information) are integrated to generate linkage-enhanced usage data. The integrated usage data, in turn, are viewed as user session data in terms of pageview-weight pairs and utilized to derive the user access patterns based on the PLSA model. Consequently, the latent factor space and user profiles have been revealed by the means of pages clustering and user session clustering according to the related probabilities. The preliminary experiments on real world data set have been conducted to evaluate the effectiveness of the proposed method. The experimental results have shown that the latent factors can be visually inferred as well as the semantic interpretation can be further made. In addition, the user access patterns have also been characterized by the user profiles, which are expressed in terms of weighted page sets. The significance of page (i.e. weight) in the user profiles can be utilized to determine the main and secondary "theme" of individual and common user access pattern, which will provide useful information for further other web applications, such as web recommendation or personalization. The future work will be focused on how to use the obtained user profile to make recommendation and predict user needed content to current active user session. Meanwhile, more experiments should be implemented on more complex data sets consisting of more distinct and meaningful web pages.

Acknowledgement

This research has been partly supported through ARC Discovery Project Grant (DP0345710) and National Natural Science Foundation of China (No 60403002).

7 References

Agarwal, R., C. Aggarwal, et al. (1999): A Tree Projection Algorithm for Generation of Frequent Itemsets. *Journal of Parallel and Distributed Computing* **61**(3): 350-371.

Agrawal, R. and R. Srikant (1994): Jorge B. Bocca and Matthias Jarke and Carlo Zaniolo. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, Santiago, Chile, 487-499, Morgan Kaufmann.

Agrawal, R. and R. Srikant (1995): Mining Sequential Patterns. *Proceedings of the International Conference on Data Engineering (ICDE)*, Taipei, Taiwan, 3-14, IEEE Computer Society Press.

Alex G Buchner and Maurice D Mulvenna (1998): Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record* **27**(4): 54-61.

Baeza-Yates, R. and B. Ribeiro-Neto (1999): *Modern information retrieval*. Sydney, Addison Wesley.

Berry, M. W., S. T. Dumais, et al. (1995): Using linear algebra for intelligent information retrieval. *SIAM Review* **37**(4): 573-595.

Cohen, E., B. Krishnamurthy, et al. (1998): Improving end-to-end performance of the web using server volumes and proxy lters. *Proc. of the ACM SIGCOMM '98*, Vancouver, British Columbia, Canada, 241-253, ACM Press.

Cohn, D. and H. Chang (2000): Learning to probabilistically identify authoritative documents. *Proc. of the 17th International Conference on Machine Learning*, San Francisco, CA, 167-174, Morgan Kaufmann.

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

Cooley, R., B. Mobasher, et al. (1999): Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems* **1**(1): 5-32.

Deerwester, S., S. T. Dumais, et al. (1990): Indexing by latent semantic analysis. *Journal American Society for information retrieval* **41**(6): 391-407.

Dempster, A. P., N. M. Laird, et al. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statist. Soc. B* **39**(2): 1-38.

Dumais, S. T. (1995): Latent semantic indexing (LSI): Trec-3 report. *Proceeding of the Text REtrieval Conference (TREC-3)*, 219-230.

Han, E., G. Karypis, et al. (1998): Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results. *IEEE Data Engineering Bulletin* **21**(1): 15-22.

Hofmann, T. (1999): Probabilistic Latent Semantic Analysis. *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, 50-57, ACM Press.

Hofmann, T. (2001): Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal* **42**(1): 177-196.

Hofmann, T. (2004): Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems* **22**(1): 89-115.

Hou, J. and Y. Zhang (2002): Constructing Good Quality Web Page Communities. *Proc. of the 13th Australasian Database Conferences (ADC2002)*, Melbourne, Australia, 36: 65-74, ACS Inc.

Hou, J. and Y. Zhang (2003): Effectively Finding Relevant Web Pages from Linkage Information. IEEE *Trans. Knowl. Data Eng.* **15**(4): 940-951.

Hou, J. and Y. Zhang (2003): Utilizing Hyperlink Transitivity to Improve Web Page Clustering. *Proceedings of the 14th Australasian Database Conferences (ADC2003)*, Adelaide, Australia, 37: 49-57, ACS Inc.

Jin, X., Y. Zhou, et al. (2004): A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content. *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*, San Jose.

Joachims, T., D. Freitag, et al. (1997): Webwatcher: A tour guide for the world wide web. *The 15th International Joint Conference on Artificial Intelligence (ICJAI'97)*, Nagoya, Japan, 770-777.

Lieberman, H. (1995): Letizia: An agent that assists web browsing. *Proc. of the 1995 International Joint Conference on Artificial Intelligence*, Montreal, Canada, 924-929, Morgan Kaufmann.

Mobasher, B., R. Cooley, et al. (1999): Creating adaptive web sites through usage-based clustering of URLs. *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, 19-25, IEEE Computer Society.

Mobasher, B., H. Dai, et al. (2002): Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery* **6**(1): 61-82.

Ngu, D. S. W. and X. Wu (1997): Sitehelper: A localized agent that helps incremental exploration of the world wide web. *Proceedings of 6th International World Wide Web Conference*, Santa Clara, CA, ACM Press.

Perkowitz, M. and O. Etzioni (1998): Adaptive Web Sites: Automatically Synthesizing Web Pages. *Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, WI, 727-732, AAAI.

Perkowitz, M. and O. Etzioni (1999): Adaptive web sites: Conceptual cluster mining. *Proc. of 16th International Joint Conference on Articial Intelligence*, Stockholm, Sweden, 264-269, Morgan Kaufmann.

Shahabi, C., A. Zarkesh, et al. (1997): Knowledge discovery from user web-page navigational. *Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97)*, 20-29, IEEE Computer Society.

Srivastava, J., R. Cooley, et al. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns fromWeb Data. *SIGKDD Explorations* **1**(2): 12-23.

WEISS, R., B. VÉLEZ, et al. (1996): HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link nHypertext Clustering. *Proceedings of the Seventh ACM Conference on Hypertext*, 180-193, ACM Press.

Xiao, J., Y. Zhang, et al. (2001): Measuring similarity of interests for clustering web-users. *Proceedings of the 12th Australasian Database conference (ADC2001)*, Queensland, Australia, 35: 107-114, ACS Inc.

Zhou, Y., X. Jin, et al. (2004): A Recommendation Model Based on Latent Principal Factors in Web Navigation Data. *Proceedings of the 3rd International Workshop on Web Dynamics*, New York, ACM Press.