

Using Probabilistic Latent Semantic Analysis for Web Page Grouping

Guandong Xu^{1,3}, Yanchun Zhang¹, Xiaofang Zhou²

¹*School of Computer Science and Mathematics
Victoria University, PO Box 14428, VIC 8001, Australia
{xu,yzhang}@csm.vu.edu.au*

²*School of Information Technology & Electrical Engineering
University of Queensland, Brisbane QLD 4072, Australia
zxf@itee.uq.edu.au*

³*School of Computer Science & Engineering
Wenzhou University, Wenzhou 325003, China
xgd@wznc.zj.cn*

Abstract

The locality of web pages within a web site is initially determined by the designer's expectation. Web usage mining can discover the patterns in the navigational behaviour of web visitors, in turn, improve web site functionality and service designing by considering users' actual opinion. Conventional web page clustering technique is often utilized to reveal the functional similarity of web pages. However, high-dimensional computation problem will incur in due to taking user transaction as dimension. In this paper, we propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) model. An iterative algorithms based on maximum likelihood principle is employed to overcome the aforementioned calculation shortcoming. The web pages are classified into various groups according to user access patterns. Meanwhile, the semantic latent factors or tasks are characterized by extracting the content of "dominant" pages related to the factors. We demonstrate the effectiveness of our approach by conducting experiments on real world data sets.

1. Introduction

With the massive influx of information onto World Wide Web, Internet has become a platform to facilitate information dissemination, information retrieval and business conducting, especially e-commerce as well. By efficiently utilizing the functionalities provided by the web sites, web users are able to achieve their interest-oriented tasks. For example, imagine a web site

devoted to online service regarding sports goods. Generally, there exist various types of user groups associated with different interests while they were visiting such a E-commerce website. One type of customers intends to browse specific category products, for example footwear, by browsing various content pages representing different companies, while another just exhibits more interests in purchasing any products belonging to one particular famous brand, such as "Nike", rather than one specific sports product category. In this manner, different types of clickstream data will be recorded in web access log files and may convey usage pattern information.

On the other hand, in order to improve the click rate of customer, web designer must anticipate the users' interests initially and construct the site appropriately. However, due to customers having vastly differing views of the site's information and shifting their needs frequently, the exhibited usage pattern may violate the author's initial opinion [1, 2]. As a sequence, there is an increasing need of improving Web site organization and presentation to facilitate user's navigation of a Web site by learning user access patterns. Capturing the characteristics of web user usage pattern can provide for, not only better understanding user navigational behavior but also for improving web functionality and service.

Related work: Recently, web mining has become a hot topic and attracted a variety of research communities, who have addressed these topics and achieved great success as well [3-6],[1, 7, 8], [9]. With the benefits of great progress in data mining research, many techniques, such as clustering [10-12], association rule mining [13, 14] and sequential pattern mining [15] are widely adopted in current web mining. Among these, the clustering-based

approach is one of the most commonly used approaches recently. Basically, there are two kinds of clustering methods in the context of web usage mining, which are associated with the objects of performing: user session clustering and web page clustering [16]. One successful application of web page clustering is adaptive web site. For example, an algorithm called PageGather [12] is proposed to synthesize index pages that are not existing initially, based on finding web page segment sharing common semantic similarity. The generated index pages are conceptually representing the various access tasks of web users. Mobasher et al. [11] utilize web user session and page clustering techniques to characterize user access pattern for web personalization based on web usage mining. Generally, web page clusters can be resulted from applying clustering process on the transpose of the session-page matrix. However, the conventional clustering techniques such as distance-based similarity methods are not capable of tackling this type high-dimensional matrix. This is mainly because that there is usually tens to hundreds of thousands sessions in web log files. Consequently, the high computational difficulty will be incurred in when we utilize sessions as dimensions rather than pages, on which we will employ clustering technique. An alternative approach for web page clustering is proposed to overcome this type clustering by [10]. This so-called Association Rule Hypergraph Partitioning (ARHP) exploits associate rule mining and graph-based technique to classify web pages into a set of clusters efficiently.

On the other hand, these aforementioned techniques can only reveal the user tasks explicitly, they, however, do not capture the intrinsic characteristics of web users' navigational activities, nor can they reveal the underlying factors associated with the usage goals. For example, such discovered usage patterns provide little knowledge of the underlying reason why such web pages are grouped together. Therefore, it is necessary to develop techniques, which can not only reveal the usage-based web user session or page clusters, but also discover the latent semantic relationships among web objects.

Latent Semantic Analysis (LSA) model is an approach to capture the latent or hidden semantic relationships among co-occurrence activities [17], and has been widely used in web information management. For example, based on LSA method, the latent semantic relationships among web pages can be discovered from linkage information, which will lead to find relevant web pages and improve web searching efficiency and effectivity [18, 19]. Factor analysis technique has become another LSA-based web usage mining approach recently, for example, Principal Factor Analysis (PFA) model is proposed to extract web user session or web page clusters and reveal the latent factors associate with user access patterns [20].

Probabilistic latent semantic analysis (PLSA) model is a probabilistic variant of LSA that was proposed for text mining by [21]. This model is originated from classical statistical analysis model and adopts maximum likelihood principle to estimate the probabilities of co-occurrences, in turn, generate web page segments with similar semantic. Recently, approaches based on PLSA has been successfully applied in collaborative filtering [22] web mining [23], text learning and mining [24, 25], co-citation analysis [26] and other related topics.

Our approach: In this paper, we address the issue of web page grouping and latent factor discovering through exploiting probabilistic latent semantic analysis (PLSA) model. By combining the usage and linkage information, we adopt the PLSA to model the relationships among user session, web page and latent factor. An iterative algorithm is employed to estimate the probability values. Based on the discovered probabilistic information, we propose a modified k-means clustering algorithm to classify web pages into different groups according to their functional similarity. It is inferred that web pages within same group may either share single similar "theme" or possess similar task-oriented "overlapping" of "themes". Meanwhile, we characterize the latent semantic factors by semantically interpreting the content of "dominant" pages whose probabilities are exceeding a predefined threshold. Furthermore, the discovered web page groups or task patterns can be utilized to advise web designer creating so-called "task list" functionality in the original web page for recommending user preferred navigational information. The experiments on two real world data sets are conducted to validate the proposed method. The experimental results have shown the effectiveness of discovering latent factors and topic-oriented web page groups as well.

The rest of the paper is organized as follow: In section 2, we introduce the principle of probabilistic latent semantic analysis model as well as integrating web linkage information into usage mining. We propose the algorithms for clustering web pages and characterizing semantic latent factors based on PLSA model in section 3. Preliminary experiments are conducted on two real data sets to evaluate the effectiveness of the proposed technique, and experimental analyses are made as well in section 4. Finally, conclusion and some future work are outlined in section 5.

2. Probabilistic Latent Semantic Analysis (PLSA) model

2.1. Data sessionization process

Prior to introduce the principle of PLSA model, we discuss briefly the issue with respect to usage data

sessionization process. In general, the user access interests 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 weighted page vector visited by users during a period of time. After data preprocessing, we can build up a page set of size n as $P = \{p_1, p_2, \dots, p_n\}$ and user session set of size m as $S = \{s_1, s_2, \dots, s_m\}$. The whole procedures are called page identification and user sessionization respectively. By simplifying user session in the form of page vector, each session can be considered as an n -dimensional page vector $s_i = \{a_{i1}, a_{i2}, \dots, a_{in}\}$, where a_{ij} denotes the weight for page p_j in s_i user session.

(page set)	p_1	p_2	...	p_j	...	p_n
s_1						
s_2						
...						
s_i				a_{ij}		
...						
s_m						

Figure 1. Organization of session-page data expression

1) Main Movies: 20sec Movies News: 15sec NewsBox: 43sec Box-Office Evita: 52sec News Argentina:31 sec Evita: 44sec
2) Music Box: 1lsec 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 box: 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 2. A usage snapshot and its normalized session-page matrix expression

As a result, the user session data can be generated to form web usage data represented by a session-page matrix $SP_{m \times n} = \{a_{ij}\}$ (Figure 1 illustrates the organization of session-page matrix). The entry in the session-page matrix, a_{ij} , is the 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 pages in same user session in order to eliminate the influence caused by the amount difference of visiting time durations or hit numbers. The session normalization is able to capture the relative significance of a page within one user session with respect to others pages accessed by same user. The

figure 2 depicts an example of usage data snapshot and its corresponding session-page matrix in the form of normalized weight expression from [27, 28].

2.2. PLSA model

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 discover the latent usage pattern. For the given aspect model, suppose that there is a latent factor space $Z = \{z_1, z_2, \dots, z_k\}$ and each co-occurrence observation data (s_i, p_j) is associated with the factor $z_k \in Z$ by varying degree to z_k .

According to the viewpoint of aspect model, thus, it can be inferred that there are existing different relationships among web users or pages related to different factors, and the factors can be considered to represent the user access pattern. For example, for an academic website, we can predefine that there exist k latent factors associated with k navigational behavior patterns, such as z_1 standing for admission applying of international students, z_2 for particular interests on postgraduate programs, and $z_3, z_4 \dots$ etc.. In this manner, each usage data (s_i, p_j) can convey the user navigational interests by mapping the observation data into the k -dimensional latent factor space. The degree to which such relationships are “explained” by each factors are derived from the factor-conditional probabilities. In this work, we adopt PLSA model to model the relationships among web pages and reveal latent semantic factors as well.

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, we construct probability of an observed pair (s_i, p_j) by adopting the latent factor variable z_k as:

$$P(s_i, p_j) = P(s_i) \cdot P(p_j | s_i) \quad (1)$$

Where,

$$P(p_j | s_i) = \sum_{z \in Z} P(p_j | z) \cdot P(z | s_i) \quad (2)$$

By applying Bayes's rule and substituting equation (2), equation (1) is re-parameterized as

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

Following the likelihood principle, the total likelihood L_i is determined as

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

where $m(s_i, p_j)$ is the element of the session-page matrix corresponding to session s_i and page p_j .

In order to maximize the total likelihood, we make use of *Expectation Maximization (EM)* algorithm to perform maximum likelihood estimation in latent variable model [29]. Generally, two steps are needed to implement in this algorithm 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 describe the whole procedure in details:

- (1) Firstly, 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:

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

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

$$P(p_j | z_k) = \frac{\sum_{s_i \in S} m(s_i, p_j) \bullet P(z_k | s_i, p_j)}{\sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k | s_i, p_j)} \quad (6)$$

$$P(s_i | z_k) = \frac{\sum_{p_j \in P} m(s_i, p_j) \bullet P(z_k | s_i, p_j)}{\sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k | s_i, p_j)} \quad (7)$$

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

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 executing of E-step and M-step is repeating until L_i is converging to a local optimal limit, which means the estimated results can represent the final probabilities of observation data.

It is easily found that the computational complexity of this algorithm is $O(mnk)$, where m is the number of user session, n is the number of page, and k is the number of factors.

2.3. Integrating Web linkage into usage mining

On the other hand, the structural 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 within the web site that is developed by the designer. 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. In short, the co-citation analysis of hyperlink can measure web page similarity in content. Therefore, considering linkage information will result in reinforcing the scalability of discovering usage pattern rather than using usage data standalone during web usage mining.

In order to model linkage information, we take hyperlink transitivity technique[30], in which the linkage information between pages is expressed by correlation matrix into account. In this way, we set up linkage matrix in the form of correlation as follows: $LP_{n \times n} = \{b_{ij}\}$. Multiplying the session-page matrix with correlation matrix will result in integrating linkage information into usage mining. More details of constructing hyperlink transitivity matrix can be found in [30, 31].

3. Characterizing semantic latent factor and discovering usage-based web page group

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, we assume that there is a task-oriented web page group associated with it. From this viewing point, we, thus, can utilize the class-conditional probability estimates generated by the PLSA model and clustering algorithm to partition web pages into various usage-based groups. Meanwhile, we can infer the latent factors by interpreting the meaning of web pages whose probabilities are exceeding the predefined threshold. Furthermore, we can a task list in the form of a collection of pages related to the specific web page group.

3.1. Characterizing latent factor

First, we discuss how to capture the latent factor associated user navigational behavior. This aim is to be achieved by characterizing the "dominant" pages. Note

that $P(p_j|z_k)$ represents the conditional occurrence probability over the page space corresponding to a specific factor, whereas $P(z_k|p_j)$ represents the conditional probability distribution over the factor space corresponding to a specific page, which is expressed in the form of

$$P(z_k | p_j) = \frac{P(p_j | z_k) \cdot P(z_k)}{\sum_{z_k \in Z} P(p_j | z_k) \cdot P(z_k)} \quad (13)$$

In such expression, we may consider that the pages whose conditional probabilities $P(z_k|p_j)$ and $P(p_j|z_k)$ are both greater than a predefined threshold μ can be viewed to contribute to one similar functionality related to the latent factor. In this manner, we choose all pages satisfying aforementioned condition to form ‘‘dominant’’ page base sets. By exploring and interpreting the content of these pages, we may characterize the semantic meaning of each factor. In section 4, we will present some examples of latent factors derived from two real data sets. The algorithm to characterize the task-oriented semantic latent factor is described as following:

Algorithm 1 characterize latent factor

Input: $P(p_j|z_k)$ and $P(z_k|p_j)$, predefined threshold μ
Output: A set of characteristic page base sets $LF = \{LF_1, LF_2, \dots, LF_k\}$

1. $LF_1 = LF_2 = \dots = LF_k = \Phi$
2. For each z_k , choose all pages $p_j \in P$
If $P(p_j|z_k) \geq \mu$ and $P(z_k|p_j) \geq \mu$ then
 $LF_k = LF_k \cup p_j$
Else go back to step 2
3. If there are still pages to be classified, go back to step 2
4. Output: $LF = \{LF_k\}$

3.2. Building web page groups

Note that the set of $P(z_k|p_j)$ is conceptually representing the probability distribution over the latent factor space for a specific web page p_j , we, thus, construct the page-factor matrix based on the calculated probability estimates, to reflect the relationship between web pages and latent factors, which is expressed as follows:

$$vp_j = (c_{j,1}, c_{j,2}, \dots, c_{j,k}) \quad (14)$$

Where $c_{j,s}$ is the occurrence probability of page p_j on factor z_s . In this way, the distance between two page vectors may reflect the functionality similarity exhibited by them. We, therefore, define their similarity by applying well-known cosine similarity as:

$$sim(p_i, p_j) = (vp_i, vp_j) / (\|vp_i\|_2 \cdot \|vp_j\|_2) \quad (15)$$

where

$$(vp_i, vp_j) = \sum_{m=1}^k c_{i,m} c_{j,m}, \quad \|vp_i\|_2 = \left(\sum_{l=1}^k C_{i,l}^2 \right)^{1/2}$$

With the page similarity measurement (15), we propose a modified k -means clustering algorithm to partition web pages into corresponding groups. The detail of the clustering algorithm is described as follows:

Algorithm 2 web page grouping

Input: the set of $P(z_k|p_j)$, predefined threshold μ
Output: A set of web page clusters $PCL = \{PCL_1, PCL_2, \dots, PCL_p\}$

1. Select the first page p_1 as the initial cluster PCL_1 and the centroid of this cluster: $PCL_1 = \{p_1\}$ and $Cid_1 = p_1$.
2. For each page p_i , measure the similarity between p_i and the centroid of each existing cluster $sim(p_i, Cid_j)$
3. If $sim(p_i, Cid_t) = \max_j (sim(p_i, Cid_j)) > \mu$, then insert p_i into the cluster PCL_t and update the centroid of PCL_t as
 $Cid_t = 1/|PCL_t| \cdot \sum_{j \in PCL_t} vp_j$ (16)

where $|PCL_t|$ is the number of pages in the cluster. Otherwise, p_i will create a new cluster itself and is the centroid of the new cluster.

4. If there are still pages to be classified into one of existing clusters or a page that itself is a cluster, go back to step 2 iteratively until it converges (i.e. all clusters’ centroid are no longer changed)
5. Output $PCL = \{PCL_p\}$

Similarly, the probabilistic distribution over the factor space of a user $P(z_k|s_i)$ can reflect the specific user’s access tendency over the whole latent factor space, in turn, may be utilized to uncover usage pattern. More detailed work with respect to generating user access pattern is described in [31].

Upon the characteristics of probabilistic distribution and clustering process, it can be inferred some generated web page groups exhibit with of single-peak probabilistic distribution, whereas the others are mostly with of multi-peak probability distribution over latent factor space. In other words, the former represent the single main ‘‘theme’’ property whereas latter ones reflect the multi-purpose characteristics derived from usage data.

4. Preliminary experimental results

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

4.1. Data sets

The first data set we used is downloaded from KDDCUP (www.ecn.purdue.edu/KDDCUP/). After data preparation, we have setup an evaluation data set including 9308 user sessions and 69 pages, where every session consists of 11.88 pages in average. We refer this data set to “KDDCUP data”. In this data set, the entries in session-page matrix associated with the specific page in the given session are determined by the numbers of web page hits by the given user.

The second data set is from a academic website log files[16]. The data is based on a 2-week web log file during April of 2002. After data preprocessing stage, the filtered data contains 13745 sessions and 683 pages. The entries in the table correspond to the amount of time (in seconds) spent on pages during a given session. For convenience, we refer this data as “CTI data”. The linkage information is derived from the web site maps and integrated into web usage mining. By considering the number of web pages and the content of the web site carefully and referring the selection criteria of factors in [23, 25], we choose 12 factors and 20 factors for KDDCUP data and CTI data sets respectively.

4.2. Latent factors based on PLSA model

We conduct the experiments on the two data sets to extract the latent factors and group web pages. Firstly, we present the experimental results of the derived latent factors derived from two real data sets based on PLSA model respectively. Table 1 illustrates the results extracted from the KDDCUP data set, whereas Table 2 presents the results of CTI data set. From these tables, it is shown the titles of latent factors are characterized by some “prominent” pages whose probabilistic weights are exceeding one predefined threshold. This work is done by interpreting the content of corresponding pages since these “dominant” pages contribute greatly to the latent factors. With the derived characteristic factors, we may semantically discover usage-based task pattern.

4.3. Examples of web page groups

At this stage, we utilize aforementioned clustering algorithm to partition the web page into various clusters. By analyzing the discovered clusters, we may conclude that many of clusters do really reflect the single user access task; whereas others may cover two or more tasks, which may be relevant in nature. As indicated above, the former can be considered to correspond to the intuitive latent factors, and the latter may reveal the “overlapping” relationships in content among web pages.

Table 1: Titles of factors from KDDCUP

Factor #	Title	Dominant Page #
1	Department search	6, 7
2	Product information of Legwear	4
3	Vendor service info	10,36,37,39
4	Freegift, especially legcare	1,9,33
5	Product information of Legcare	5
6	Online shopping process	27,29,32,42,44,45,60
7	Assortment of various lifestyle	3,26
8	Vendor2’s Assortment	11,34
9	Boutique	2
10	Replenishment info of Department	6,25,26,30
11	Article regarding Department	12,13,22,23
12	Home page	8,35

Table 2: Titles of factors from CTI

Factor #	Title	Factor #	title
1	specific syllabi	11	international_study
2	grad_app_process	12	Faculty-search
3	grad_assist_app	13	postgrad_program
4	admission	14	UG_scholarship
5	advising	15	tutoring_gradassist
6	program_bachelor	16	Mycti_stud_profile
7	syllabi list	17	schedule
8	course info	18	CS_PhD_research
9	jobs	19	specific news
10	calendar	20	Home page

In Table 3, we list three web page groups out of total generated groups from KDDCUP data set, which is expressed by top ranked page information such as page numbers and their relative URLs as well. It is seen that each of these three page groups reflects sole usage task, which is consistent with the corresponding factor depicted in Table 1. Table 4 illustrates two web page groups from CTI data set as well. In this table, the upper row lists the top ranked pages and their corresponding content from one of the generated page clusters, which reflect the task regarding searching postgraduate program information, and it is easily to conclude that these pages are all contributed to factor #13 displayed in Table 2. On the other hand, the listed significant pages in lower row in the table involve in the “overlapping” of two dominant tasks, which are corresponding to factor #3 and #15 depicted in Table 2.

Note that with these either generated web page groups or web session groups; we may make use of these intrinsic relationships among web pages or web users to improve web organization or functionality, for example,

the instrumental and suggestive task list based on the discovered page groups can be added into the original web page as the means of web recommendation, to provide better service to users.

Table 3: Examples of web page groups from KDDCUP

Page	Content	Page	Content
10	main/vendor	38	articles/dpt_payment
28	articles/dpt_privacy	39	articles/dpt_shipping
37	articles/dpt_contact	40	articles/dpt_returns
27	main/login2	50	account/past_orders
32	main/registration	52	account/credit_info
42	account/your_account	60	checkout/thankyou
44	checkout/expresCheckout	64	account/create_credit
45	checkout/confirm_order	65	main/welcome
47	account/address	66	account/edit_credit
12	dpt_about	20	dpt_affiliate
13	dpt_about_mgmtteam	21	new_security
14	dpt_about_boarddirectors	22	new_shipping
15	dpt_about_healthwellness	23	new_returns
16	dpt_about_careers	24	dpt_terms
17	dpt_about_investor	57	dpt_about_the_press
18	dpt_about_pressrelease	58	dpt_about_advisoryboard
19	dpt_refer		

Table 4: Examples of web page groups from CTI

Page	Content	Page	Content
386	/News	588	/Prog/2002/Gradect2002
575	/Programs	590	/Prog/2002/Gradis2002
586	/Prog/2002/Gradcs2002	591	/Prog/2002/Gradmis2002
587	/Prog/2002/Gradds2002	592	/Prog/2002/Gradse2002
65	/course/internship	406	/pdf/forms/assistantship
70	/course/studyabroad	666	/program/master
352	/cti/.../applicant_login	678	/resource/default
353	/cti/.../assistantship_form	679	/resource/tutoring
355	/cti/.../assistsubmit		

5. Conclusion and future work

In this paper, we have presented a probabilistic latent semantic analysis (PLSA) model, which can discover the hidden semantic factors and web page groups from the usage and linkage observation data. The data from two different sources, namely, web log server files (i.e. usage data) and web site map (i.e. linkage information) are combined into web mining. Two clustering algorithms have been proposed to estimate the probabilistic values for discovering the intrinsic relationships among web pages and user sessions. The preliminary experiments on two real world data sets have been conducted to evaluate the effectiveness of the proposed method. The experimental results have shown that the latent factors can be inferred by semantic interpretation of “dominant” pages. Meanwhile, task-based web page groups have been generated according to user access patterns. Furthermore, the uncovered page groups and latent factors can be

instrumental for task-oriented web recommendation or personalization, which will provide user more useful functionality and service. The future work will be focused on how to make recommendation and predict user needed content to current active user session by combining the aggregated page segment.

6. Acknowledgement

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

7. Reference

1. Perkowitz, M. and O. Etzioni. *Adaptive web sites: Automatically synthesizing web pages*. in *Fifteenth National Conference on Artificial Intelligence*. 1998. Madison, WI.
2. Perkowitz, M. and O. Etzioni, *Adaptive Web sites*. *Communications of the ACM*, 2000. **43**(8): p. 152 - 158.
3. Joachims, T., D. Freitag, and T. Mitchell. *Webwatcher: A tour guide for the world wide web*. in *The 15th International Joint Conference on Artificial Intelligence (ICJAI'97)*. 1997. Nagoya, Japan.
4. Lieberman, H. *Letizia: An agent that assists web browsing*. in *Proc. of the 1995 International Joint Conference on Artificial Intelligence*. 1995. Montreal, Canada: Morgan Kaufmann.
5. Mobasher, B., R. Cooley, and J. Srivastava. *Creating adaptive web sites through usage-based clustering of URLs*. in *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*. 1999: IEEE Computer Society.
6. Ngu, D.S.W. and X. Wu. *Sitehelper: A localized agent that helps incremental exploration of the world wide web*. in *Proceedings of 6th International World Wide Web Conference*. 1997. Santa Clara, CA: ACM Press.
7. Cohen, E., B. Krishnamurthy, and J. Rexford. *Improving end-to-end performance of the web using server volumes and proxy lters*. in *Proc. of the ACM SIGCOMM '98*. 1998. Vancouver, British Columbia, Canada: ACM Press.
8. Perkowitz, M. and O. Etzioni. *Adaptive web sites: Conceptual cluster mining*. in *Proc. of 16th International Joint Conference on Artificial Intelligence*. 1999. Stockholm, Sweden: Morgan Kaufmann.
9. Alex Buchner and Maurice D Mulvenna., *Discovering internet marketing intelligence through online analytical web usage mining*. *SIGMOD Record*, 1998. **27**(4): p. 54-61.
10. Han, E., et al., *Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results*. *IEEE Data Engineering Bulletin*, 1998. **21**(1): p. 15-22.
11. Mobasher, B., et al., *Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization*. *Data Mining and Knowledge Discovery*, 2002. **6**(1): p. 61-82.

12. Perkowit, M. and O. Etzioni. *Adaptive Web Sites: Automatically Synthesizing Web Pages*. in *Proceedings of the 15th National Conference on Artificial Intelligence*. 1998. Madison, WI: AAAI.
13. Agarwal, R., C. Aggarwal, and V. Prasad, *A Tree Projection Algorithm for Generation of Frequent Itemsets*. *Journal of Parallel and Distributed Computing*, 1999. **61**(3): p. 350-371.
14. Agrawal, R. and R. Srikant. *Jorge B. Bocca and Matthias Jarke and Carlo Zaniolo*. in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*. 1994. Santiago, Chile: Morgan Kaufmann.
15. Agrawal, R. and R. Srikant. *Mining Sequential Patterns*. in *Proceedings of the International Conference on Data Engineering (ICDE)*. 1995. Taipei, Taiwan: IEEE Computer Society Press.
16. Mobasher, B., *Web Usage Mining and Personalization*, in *Practical Handbook of Internet Computing*, M.P. Singh, Editor. 2004, CRC Press.
17. Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*. 1999, Sydney: Addison Wesley.
18. Hou, J. and Y. Zhang. *Constructing Good Quality Web Page Communities*. in *Proc. of the 13th Australasian Database Conferences (ADC2002)*. 2002. Melbourne, Australia: ACS Inc.
19. Hou, J. and Y. Zhang, *Effectively Finding Relevant Web Pages from Linkage Information*. *IEEE Trans. Knowl. Data Eng.*, 2003. 15(4): p. 940-951.
20. Zhou, Y., X. Jin, and B. Mobasher. *A Recommendation Model Based on Latent Principal Factors in Web Navigation Data*. in *Proceedings of the 3rd International Workshop on Web Dynamics*. 2004. New York: ACM Press.
21. Hofmann, T. *Probabilistic Latent Semantic Analysis*. in *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*. 1999. Berkeley, California: ACM Press.
22. Hofmann, T., *Latent Semantic Models for Collaborative Filtering*. *ACM Transactions on Information Systems*, 2004. **22**(1): p. 89-115.
23. Jin, X., Y. Zhou, and B. Mobasher. *A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content*. in *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*. 2004. San Jose.
24. Cohn, D. and H. Chang. *Learning to probabilistically identify authoritative documents*. in *Proc. of the 17th International Conference on Machine Learning*. 2000. San Francisco, CA: Morgan Kaufmann.
25. Hofmann, T., *Unsupervised Learning by Probabilistic Latent Semantic Analysis*. *Machine Learning Journal*, 2001. **42**(1): p. 177-196.
26. Cohn, D. and T. Hofmann, *The missing link: A probabilistic model of document content and hypertext connectivity*, in *Advances in Neural Information Processing Systems*, T.G.D. Todd K. Leen, and Tresp, V., Editor. 2001, MIT Press.
27. Shahabi, C., et al. *Knowledge discovery from user web-page navigational*. in *Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97)*. 1997: IEEE Computer Society.
28. Xiao, J., et al. *Measuring similarity of interests for clustering web-users*. in *Proceedings of the 12th Australasian Database conference (ADC2001)*. 2001. Queensland, Australia: ACS Inc.
29. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. *Journal Royal Statist. Soc. B*, 1977. **39**(2): p. 1-38.
30. Hou, J. and Y. Zhang. *Utilizing Hyperlink Transitivity to Improve Web Page Clustering*. in *Proceedings of the 14th Australasian Database Conferences (ADC2003)*. 2003. Adelaide, Australia: ACS Inc.
31. Xu, G., et al. *Discovering User Access Pattern Based on Probabilistic Latent Factor Model*. in *Proceeding of 16th Australasian Database Conference*. 2004. Newcastle, Australia: ACS Inc.