Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 6713-6721

http://www.ripublication.com

Mathematical User Profiling Algorithms for Web Recommendation Based on PLSA and LDA Model

[1] Padma Priya. G, and [2] Dr. M. Hemalatha

[1]PhD Research Scholar, [2]Research Supervisor,
[1, 2] Bharathiar university, Tamilnadu, India.

Abstract

Web transaction data between Web visitors and Web pages usually convey user task-oriented behavior patterns. We have proposed a mathematical user profiling algorithm for Web recommendation based on PLSA and LDA model. With the discovered usage knowledge from Web usage mining via various latent semantic analysis models, we construct a set of usage access patterns. By measuring the similarities between the active user and the discovered usage patterns, we choose the most similar user profile as the candidate usage pattern. The recommended page list is generated by incorporating the chosen user profile with the top-N weighted scoring scheme for Web recommendation. We have developed two mathematical Web recommendation algorithms with Probabilistic Semantic Latent Analysis(PLSA) and Latent Dirichlet Allocation (LDA) model respectively.

Keywords: Mathematical Algorithms, PLSA, LDA, Semantic analysis

1. INTRODUCTION

Web recommendation approach by identifying the user's task-oriented navigational distribution and incorporating it into the top-N weighted scoring scheme. Experiments conducted on the real world data sets have evaluated the proposed algorithm in terms of recommendation accuracy. The main idea of this approach is the use of the weights of pages within the dominant task space; however, it doesn't take the historical visits of other Web users into consideration. As a consequence, we aim to develop a Web recommendation algorithm via collaborative filtering techniques in this chapter. In particular, we propose two Web recommendation algorithms, which are called user

profiling approaches based on two latent semantic analysis models. The rest of this chapter is organized as follows: we first present two usage-based Web recommendation mathematical algorithms based on PLSA and LDA model respectively. we give Web recommendation performance analysis in terms of recommendation accuracy from experiments conducted on the real world usage datasets, with the proposed recommendation algorithms . In order to evaluate the effectiveness of the proposed algorithms, comparison studies are carried out against existing Web recommendation algorithms.

2. RELATED WORK

Web recommendation research has become a hot topic in the context of Web data management in the last decade despite of the fact that recommender systems have been well studied in machine learning and information retrieval areas. To-date, there are two kinds of approaches and techniques commonly used in Web recommendation, namely content-based filtering and collaborative filtering systems [3, 4]. Content-based filtering systems such as WebWatcher [5] and client-side agent Letizia [6] generally generate recommendations based on the pre-constructed user profiles by measuring the similarity of Web contents to these profiles, while collaborative filtering systems make recommendations by utilizing the rating of the current user for objects via referring to other users' preferences that is closely similar to the current one. In addition, Web usage mining has been proposed as an alternative method for not only revealing user access patterns, but also making Web recommendations recently [1]. With the benefit of great progress in data mining research communities, many data mining techniques, such as collaborative filtering based on the k-Nearest Neighbour algorithm (kNN) [7-9], Web user or page clustering [1, 10, 46], association rule mining and sequential pattern mining technique [19] have been adopted in the current Web usage mining methods. With the development of Web usage mining techniques contributed by academics and researchers in a variety of application areas, the application fields are broadened widely and deepened throughout. For example, Liu et al [16] proposed a framework for forming communities in a peer-to-peer communication environment by analysing the client-side Web browsing history. This framework is based on a statistics-based approach. In [17], Bose et al incorporated the ontology in the form of concept hierarchy into usage based recommendation systems to reinforce recommendations by taking semantics into account, whereas [18] combined model-based and memory-based CF algorithms into a hybrid system to improve the recommendation performance without incurring high computational costs. And Jin et al [20] introduced a Maximum Entropy algorithm into the recommendation scoring algorithm to achieve better recommendation. Consequently, many efforts have been contributed and great achievements have been made in such research fields as Web personalization and recommendation systems [5, 6], Web

system improvement , Web site modification or redesign, and business intelligence and e-commerce [5].

3. USER PROFILING ALGORITHMS FOR WEB

we present two usage-based user profiling algorithms for Web recommendation based on PLSA and LDA model respectively. we have derived user access patterns and user profiles via a probability inference algorithm. In the following parts, we aim to incorporate the discovered usage knowledge with the collaborative filtering algorithm into the Web recommendation algorithm.

3.1 Recommendation Algorithm based on PLSA Model

As discussed in the previous chapter, Web usage mining will result in a set of user session clusters { } SCL SCLSCL SCL= _ , where each , ,1 2 kSCL is a collection of user I sessions with similar access preferences. And from the discovered user session clusters, we can then generate their corresponding centroids of the user session clusters, which are considered as usage profiles, or user access patterns. The complete formulation of usage profiling algorithm is expressed as follows:

Given a user session cluster SCL , the corresponding usage profile of the cluster is I represented as a sequence of page weights, which are dependent on the mean weights of all pages engaged in the cluster where the contributed weight,ijii ii n up w w w= _ (7.1) , ,( )1 2w , of the page p within the user profile jup is: iAnda is the element weight of the page select the most significant pages for recommendation, we can use filtering method to choose a set of dominant pages with weights exceeding a certain value as an expression of user profile, that is, we preset a thresholdμ and filter out those pages with weights greater than the threshold for constructing the user profile. Given This process performs repeatedly on each user session cluster and finally generates a number of user profiles, which are expressed by the weighted sequences of pages. These usage patterns are then used into collaborative recommendation operations. Generally, a Web recommendation is to predict and customize Web presentations in a user preferable style according to the interests exhibited by individual or groups of users. This goal is usually carried out in two ways. On the one hand, we can take the current active user's historic behaviour or pattern into consideration, and predict the preferable information to this specific user. On the other hand, by finding the most similar access pattern to the current active user from the learned usage models of other users, we can recommend the tailored Web content. The former one is sometimes called memory-based approaches, whereas the latter one is called model-based recommendations, respectively. In this work, we adopt the model-based technique in our Web recommendation framework. We consider the

usage-based user profiles generated in section 4.3 as the aggregated representations of common navigational behaviours exhibited by all individuals in the same particular user category, and utilize them as a usage knowledge base for recommending potentially visited Web pages to the current user. Similar to the method proposed in [1] for representing user access interest in the form of a n-dimensional weighted page vector, we utilize the commonly used cosine function to measure the similarity between the current active user session and discovered usage patterns. We, then, choose the best suitable profile, which shares the highest similarity with the current session, as the matched pattern of current user. Finally, we generate the top-N recommendation pages based on the historically visited probabilities of pages by other users in the selected profile. The detailed procedure is described as follows:

[Algorithm 1]: MathematicalUser profiling algorithm for Web recommendation based on PLSA

[Input]: An active user session s and a set of user profiles { }aup up= . matmat PLSA a j j

[Output]: Top-N recommendation pages( ) { , 1, 2, , }Step 1: The active session ajREC s p pP j N= ∈ = …. S and the discovered user profile sup are viewed as n-dimensional vectors over the page space within a site, i.e. jiw is the significant weight contributed by the page a aaa ns w wStep 2: Measure the similarities between the active session and all derived usage profiles, and choose the maximal one out of the calculated similarities as the most matched pattern

Step 3: Incorporate the selected profile recommendation score ( )ia mat a jjup with the active session rs p for each page Thus, each page in the profile will be assigned a recommendation score between 0 and 1. Note that the recommendation score will be 0 if the page is already visited in the current session,

Step 4: Sort the calculated recommendation scores obtained in step 3 in a descending order, i.e. mat matmatnrs w w w= _, and select the N pages with the highest

3.2 Recommendation Algorithm Based on LDA Model

In this section, we present a user profiling algorithm for Web recommendation based on LDA generative model. LDA is one of the generative models, which is to reveal the latent semantic correlation among the co-occurent activities via a generative procedure. Similar to the Web recommendation algorithm proposed in the previous section, we, first, discover the usage pattern by examining the posterior probability estimates derived via LDA model, then, measure the similarities between the active user session and the usage patterns to select the most matched user profile, and eventually make the collaborative recommendation by incorporating the usage patterns with collaborative filtering, i.e. referring to other users' visiting preferences,

who have similar navigational behaviours. Likewise, we employ the top-N weighted scoring scheme algorithm into the collaborative recommendation process, to predict the user's potentially interested pages via referring to the page weight distribution in the closest access pattern. In the following part, we explain the details of the algorithm. Given a set of user access models and the current active user session, the algorithm of generating the top-N most weighted pages recommendation is outlined as follows:

[Algorithm 2]: Mathematical User profiling for Web recommendation based on LDA model

[Input]: An active user sessionand a predefined threshold, the computed session-topic preference distribution θ arec reclda a j j

[Output]: Top-N recommendation pages( ) { | , 1, 1}

Step 1: Treat the active session a i w = , if page 1iaREC s p pP j N= ∈ = -_. sas a n-dimensional vectors: p is already clicked, and otherwise

Step 2: For each latent topicjaiw=z, choose all user sessions user session aggregation R and compute the usage pattern as s Rjiszisup∑∈θ·=Rij0, ia a aa ns w w w= _ , where

Step 3: Measure the similarities between the active session and all learned user access models, and choose the maximum one out of the calculated similarities as the most closely matched access pattern

Step 4: Refer to the page weight distribution in the most matched access pattern and calculate the recommendation score ( ) RS p for each page I recii a rec ( ) (,) pi: RS p w sim s up= × (7.11) Thus, each page in the matched pattern will be assigned a recommendation score between 0 and 1. Note that the recommendation score will be 0 if the page is already visited in the current session,

Step 5: Sort the calculated recommendation scores obtained in step 4 in a descending order, i.e. rec recrecnRS w w w= _, and select the N pages with the top-N highest (, , ,)1 2recommendation scores to construct the top-N recommendation set: rec recrecLDA a j jjREC s p RS p RS p j N( ) { | ( ) ( ), 1, 2, 1}

4. EXPERIMENTS AND RESULTS

In order to evaluate the effectiveness of the proposed methods based on PLSA and LDA model, we conduct experiments on two real world usage data sets, and conduct comparisons with existing recommendation algorithms. We present the results of the recommendation performance in this section. The Web usage dataset used in experiments is the same as those used in the previous chapters. We briefly describe the datasets as follows. The data set is from a university's department Web log file,

which consists of 13710 sessions and 683 pages, and each entry up , rec of the usage matrix corresponds to the amount of time (in seconds) spent on pages during a given session. For convenience, we refer to this data as "CTI data".

4.2 Evaluation Metric of Web Recommendation Accuracy

Here we use the evaluation metric of Web recommendation accuracy described in the previous chapter. This metric is called hit precision [1], which is used to assess the effectiveness of the recommendation algorithm in the context of top-N recommendation. In order to compare our approach with other existing methods, we implement a baseline method that is based on the clustering technique [1].

4.3 Experimental Results

Figure 1 depicts the comparison results of recommendation accuracy in terms of hitpparameter using PLSA-based and clustering-based recommendation algorithm respectively with CTI dataset. From the Figure 1, it is shown that the proposed PLSA-based technique consistently overweighs the standard clustering-based algorithm in terms of hit precision parameter. In this scenario, it can be concluded that our approach is capable of making Web recommendation more accurately and effectively against the conventional methods. For another user, we can find that the user is mainly conducting two tasks, i.e. task #4 and task #13. Referring to the derived tasks in Table 6-2, we can further identify that task #4 represents prospective students searching for admission information, such as requirement, orientation etc, whereas task #13 reflects the activity of those students who are particularly interested in postgraduate programs in IT disciplines. Unlike the first user, the second user clearly exhibits the cross-interest as the difference between the two corresponding probabilities of the tasks is not quite significant.

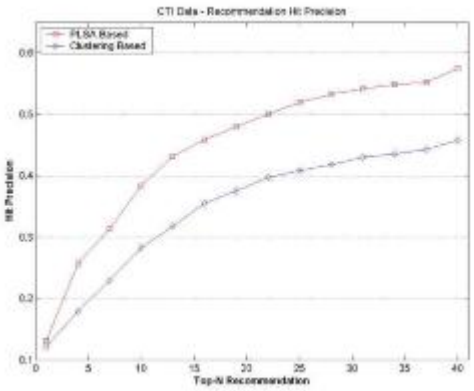

Figure 1: Web recommendation evaluation uponhitpcomparison for CTI dataset

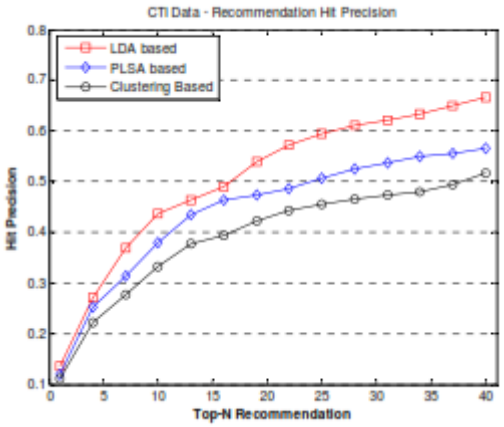

Figure 2: Hit precision comparison of Web recommendation on CTI dataset

The experimental results in terms of hit precision with LDA model are shown in Figure 2. In order to compare the proposed approach with other methods, we also carry out experiments on CTI dataset with the conventional clustering-based and the PLSA-based approaches. In a similar manner, the usage-based session clusters by performing the k-means clustering and the probability inference with PLSA model [1, 15] are constructed to aggregate user sessions with similar access preferences, and the centroids of clusters are derived as the aggregated user access patterns.

The results demonstrate that the proposed LDA-based technique consistently outperforms the standard clustering-based and the PLSA-based algorithms in terms of hit precision parameter, the standard clustering-based algorithm always generates the least accurate recommendation precision, and the recommendation performance of the PLSA-based algorithm is in the middle of the other two. From this comparison, it can be concluded that the proposed recommendation approaches based on latent semantic analysis models are capable of making Web recommendation more accurate and effective against the conventional recommendation methods. In addition to the advantage of high recommendation accuracy, these approaches are also able to identify the latent semantic factors why such user sessions or Web pages are grouped together in the same category.

5. CONCLUSION

Web transaction data between Web visitors and Web pages usually convey user task-oriented behaviour patterns. As a result, there is an increasing demand to develop techniques that can not only discover user task-oriented access patterns, but also characterize the underlying relationships among Web users, user access tasks and Web pages.

In this chapter, we have proposed a unified user profiling algorithm for Web recommendation based on PLSA and LDA model. With the discovered usage knowledge from Web usage mining via various latent semantic analysis models, we construct a set of usage access patterns (i.e. user profiles). By measuring the similarities between the active user and the discovered usage patterns, we choose the most similar user profile as the candidate usage pattern. The recommended page list is generated by incorporating the chosen user profile with the top-N weighted scoring scheme for Web recommendation. We have developed two Web recommendation algorithms with PLSA and LDA model respectively. Experimental results that conducted in comparison with other existing recommendation algorithms have shown that latent semantic analysis based recommendation algorithms are able to make recommendations accurately and efficiently. In addition to the high recommendation precision, the latent semantic analysis models have the capability of revealing the latent factor space associated with the discovered usage knowledge.

REFERENCES

[1] Mobasher, B., et al., Discovery and Evaluation of Aggregate Usage Profiles forWeb Personalization. Data Mining and Knowledge Discovery, 2002. 6(1): p. 61-82.

[2] Madria, S.K., etal.Research Issues in Web Data Mining. inProceedings of DataWarehousing and Knowledge Discovery, First International Conference, DaWaK'18. 1189, p. 23-312, Florence, Italy.

[3] Dunja, M., Personal Web Watcher: Design and Implementation (Report). 1915, Department of Intelligent Systems, J. Stefan Institute, Slovenia.

[4] Herlocker, J.L., et al., Evaluating Collaborative Filtering Recommender Systems. ACM Transaction on Information Systems (TOIS), 2004. 22(1): p. 5 - 53

[5] Joachims, T., D. Freitag, and T. Mitchell. Webwatcher: A Tour Guide For the World Wide Web. inThe 15th International Joint Conference on Artificial Intelligence (IJCAI'16) . 1916, p. 770-777, Nagoya, Japan.

[6] Lieberman, H. Letizia: An Agent that Assists Web Browsing. inProc. of the 1914 International Joint Conference on Artificial Intelligence. 1914, p. 114-91,Montreal, Canada: Morgan Kaufmann.

[7] Herlocker, J., et al. An Algorithmic Framework for Performing Collaborative Filtering. inProceedings of the 22nd ACM Conference on Researchand Development in Information Retrieval (SIGIR'18). 1189, p. 22-237, Berkeley, CA, USA.

[8] Konstan, J., et al., Grouplens: Applying Collaborative Filtering to Usenet

News.Communications of the ACM, 1916. 5: p. 77-87.

[9] Shardanand, U. and P. Maes. Social Information Filtering: Algorithms forAutomating 'Word of Mouth'. inProceedings of the Computer-Human InteractionConference (CHI14). 1914, p. 210-217, Denver, Colorado.

[10] Han, E., et al., Hypergraph Based Clustering in High-Dimensional Data Sets: ASummary of Results. IEEE Data Engineering Bulletin, 1917. 21(1): p. 15-22.

[11] Wei, X. and W.B. Croft. LDA-Based Document Models for Ad-hoc Retrieval. in Proceedings of SIGIR'06 2006, p. 178-185, Seattle, Washington, USA.

[12] Li, F.-F. and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. inProceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005, p. 524 - 531, San Diego, CA, USA.

[13] Wang, X. and A. McCallum. Topics over Time: A Nonmarket ContinuousTime Model of Topical Trends. inProceedings of ACM SIGKDD. 2006, p. 74-83, Philadelphia, Pennsylvania, USA.

[14] Russell, S.J. and P. Norvig, Artificial Intelligence, A Modern Approach. 1914: Prentice Hall.

[15] Xu, G., Y. Zhang, and X. Zhou. A Web Recommendation Technique Based on Probabilistic Latent Semantic Analysis. inProceeding of 6th International Conference of Web Information System Engineering (WISE'2005). 2005, p. 15-28, New York City, USA: LNCS 306.

[16] Liu, K., et al., Client-side Web Mining for Community Formation in Peer-to-Peer Environments. ACM SIGKDD Explorations Newsletter, 2006. 8(2): p. 11 - 20.

[17] Bose, A., et al., Incorporating Concept Hierarchies into Usage Mining Based Recommendations. Advances in Web Mining and Web Usage Analysis, 2007: p. 110--126.

[18] Rashid, A.M., et al. ClustKNN: A Highly Scalable Hybrid Model- & Memory-Based CF Algorithm . inProceedings of WEBKDD'06. 2006, Philadelphia, Pennsylvania, USA.

[19] Jin, X., Y. Zhou, and B. Mobasher. A Maximum Entropy Web Recommendation System: Combining Collaborative and Content Features .inProceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'05). 2005, p. 612 617, Chicago.

[20] Inman, V.T., H.J. Ralston, and F. Todd, Human Walking. 1171, Baltimore, MD: Williams and Wilkins.