History Based User Interest Modeling in WWW Access

Shuang Han, Wenguang Chen, and Heng Wang

School of Electronics Engineering and Computer Sciences, Peking University
hs@graphics.pku.edu.cn

Abstract. WWW cache stores user's browsing history, which contains large amount of information that may be accessed again but not yet added to user's favorite page folder. The existed www pages can be used to abstract user's interest and predicts user interaction. By that means, a model that describes user's interest is needed. In this paper, we discuss two methods about www-cache, data mining and user interest: simple user interest model and real time two-dimensional interest model. Moreover, the latter is described in detail and applied to user interest modeling. An experiment is performed on 20 users' interest data sets, which shows real time two-dimensional interest model is more effective in www cache modeling.

Keywords: www cache, user interest, interest model, data mining.

1 Introduction

As World Wide Web becomes more common in people's lives, the continuously increasing amount of information poses challenges for Web users to find useful information from web data lack of regular order. Sometimes an individual user needs to discover valid information from history web pages over again. However, the disorder storage of web information makes it rather hard. Therefore, it is necessary to bring a method to get user's favorite web pages in history. A user interest model that abstracts user interest and predicts user interaction is constructed based on log record. In this paper, we propose to obtain user's access preference by history based user interest modeling. Two methods are performed in our experiment on 20 users' interest data sets. Then the two methods are measured in consistent way, which shows real time two-dimensional interest model is more effective in www cache modeling.

2 History Based User Interest Modeling in WWW Access

2.1 Method Overview

There are several methods about www-cache, data mining and user interest. Traditional www-cache methods store the most highly accessed web pages. To abstract user interest, *simple user interest model* [1] described user interest through the definition (term, weight), in which weight represents the importance of term. The main idea of building *simple user interest model* is to calculate weight of term. Early *simple user interest model* takes the frequency of term in hyper text as weight of term

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 308–312, 2007.

[2][3][4]. Improved *simple user interest model* defines the term set T as $\{t_1, t_2, ..., t_m\}$, the text set in www cache D as $\{d_1, d_2, ..., d_n\}$. The improved calculation of weight of term is given as

$$Node(t_i).weight = idf_i \sum_{j=1}^{n} stf_{ij} . \quad (1)$$

Where idf_i represents reverse-text frequency of term t_i in D while df_i represents the text frequency of term t_i in D (only one count for each text), and stf_{ij} represents the frequency of term t_i in d_j, considering both places and tags of t_i in d_j.

Simple user interest model method is text-based data mining method [2][5]. However, *simple user interest model* ignored the important relationship among interests. To solve this problem, a *real time two-dimensional interest model* [6] was proposed.

The property of real time in *real-time two-dimensional interest model* can show the user's current interest states. And the inferential relations between interests are well considered in the model. This model is not the simple extension of the simple interest model, but the round improvement of the model and its related algorithm.

Our main work is to get user's favorite web pages in history through *real time two-dimensional interest model*, so as to make it more convenient for user to find useful information.

2.2 Real-Time Two-Dimensional Interest Model

User interest can be described both one-dimensionally and two-dimensionally, respectively related to "how important a single interest is in the interest set of a user" and "the successive relationship between two interests". We take the former as Interest Node, the latter as Interest Rule. *Real-time two-dimensional interest model* is mainly based on them.

We collected 20 users' interest data sets provided by 20 different users as experimental materials. For each user, we build real-time two-dimensional interest model to predict his past browsing interest. Therefore, we need to get both Interest Node and Interest Rule for each user.

Exactly like the one in *simple user interest model*, *Interest Node* is a binary group (term, weight), where weight represents the importance of interest term. Weights of *Interest Nodes* are calculated through mining of text information in the www cache, so as to obtain the term that shows user's interest. Then the favorite web pages will be picked out. Primary calculation of weight of *Interest Node* follows the expression (1).

However, every newly visited web page will cause the recalculation of all *Interest Nodes*. Therefore, *Interest Nodes* need to be updated in real-time. Also, the relationship among one-dimensional interests is significant for interest prediction. As a result, *Interest Rules*, which keep passages between interests, are proposed. To obtain *Interest Rules*, we need the existing dependence among web pages in www cache: the existing passages from one page to another. The turning from current web page to another is what we call "browsing trends", according to which a Trend Matrix

can be generated, by which we create the *Interest Library* that keeps all *Interest Rules*, with their weights calculated using the *Trend Matrix*. Assume at time *t*, user is browsing page S_i ; at the next time *t+1*, user might choose:

① Keep browsing page S_i
② Click the addresses on page S_i then turn to one or more other pages
③ Click the "go back" button so as to return to the page last visited
④ Enter new address or open a new page in favorite folder

Pages in www-cache can be represent by directed graph *G=(V,E)*, in which pages are abstracted as nodes, the hyperlink relationship among pages as directed edges. We take α , β , χ and δ to represent the 4 trends user might choose as described above, where they meet $0 < \alpha, \beta, \chi, \delta < 1$, $\alpha + \beta + \chi + \delta = 1$. In this case, the *Trend Matrix Q* can be defined as follows:

$$Q = (q_{ij})_{n\times n},\ q_{ij} = \begin{cases} \alpha, & i = j \\ \beta, & (v_i, v_j) \in E \\ \chi, & (v_j, v_i) \in E \\ \delta, & others \end{cases}. \tag{2}$$

With the generation of *Trend Matrix*, *Interest Library* for each user can be built up. As a result, *Interest Nodes* will be updated in real-time. The *Interest Library* can be updated every other time, which is up to the user.

Until here, we can get our primary *real-time two-dimensional interest model*. However, the whole *Interest Library* will result in a huge needed disk space up to 20TB[1] that can't be afforded by normal users. Therefore, we have to roughen user interest. The set of all interest terms *T* is partitioned into disjoint union of equivalence classes according to the equivalence relation on *T*, *Interest Node* converts to *Rough Interest Node*, and *Interest Rule* converts to *Rough Interest Rule*, which greatly reduces space of *Interest Library* storage. Now the *real-time two-dimensional interest model* can be built in reality.

Finally, while building *real-time two-dimensional interest model*, all of web pages are treated equally, concealing the different importance of different web pages, which should be handled differently. The hypertext link relationship among web pages contains a large amount of underlying language meaning, helpful for the automatic analyzing of user interest. Individual web page's value can be judged, based on the defined value commonly adopted as PageRank (see reference [1]).

2.3 Obtaining User Interest

Since we've got *real-time two-dimensional interest model*, 20 users' interest can be concluded (calculated? determined by computation). With the existed *Interest Node* and *Interest Library*, at the moment user refreshes the improved history web page list, the conjectured favorite pages will be moved to top. Sequence of pages is decided by real-time-updated *Interest Nodes*. User's *Interest Nodes* are sorted by their weights. The location of every single page of each user depends on its key words' location in

sequenced *Interest Node* set. Order of page is decided by weight of page. Calculation of weight of page S_j is given by

$$S_j.weight = \sum (stf_{ij} \cdot RoughNode(C_i).weight) . \quad (3)$$

To avoid unnecessarily unimportant calculation, we don't have to consider every *Interest Node* while calculating. Only a few *Interest Nodes*, which are evidently of greater importance than others take part in calculation. Beforehand, we take these *Interest Nodes* as concerned interest set *IS*. In this case, the expression can be modified as

$$S_j.weight = \sum_{C_i \in IS} (stf_{ij} \cdot RoughNode(C_i).weight) . \quad (4)$$

2.4 Experimental Results

Considering that there is no strict preference on web pages, user's preference is usually rough. Assume that user's interest can be divided into 5 levels, 20 users manually classify their history web pages by 5 levels, which we keep as benchmark. 5 values are given to the 5 levels of user interest: 1, 2, 3, 4, 5. As for web page S_i, we take its value *L(i)*. The lower the value of a single page is, the more interest the user shows in this page.

Then we apply both *simple user interest model* and *real time two-dimensional interest model* to obtain user interest. For each model, we compare the sorted history web page sequence with pages marked by users. In the web page sequence, if one page shows in front of the other page which owns a lower value, we call it *Rank Reversal*, by which we will judge the efficiency of each model. We define *Rank Reversal* of the whole web page sequence as

$$RR = \sum_i \sum_{j<i, L(j)>L(i)} (L(j) - L(i)) . \quad (5)$$

Also, we define total *Rank Reversal* as RR_{total} . *RR* can be standardized as $\frac{RR}{RR_{total}}$. Therefore, the precision of modeling is measured by

$$P = 1 - \frac{RR}{RR_{total}} . \quad (6)$$

Fig.1 shows results of 20 users in both *simple user interest model* and *real time two-dimensional interest model*. The precisions (P) of 20 users in *real time two-dimensional interest model* are mostly higher than the ones in *simple user interest model*.

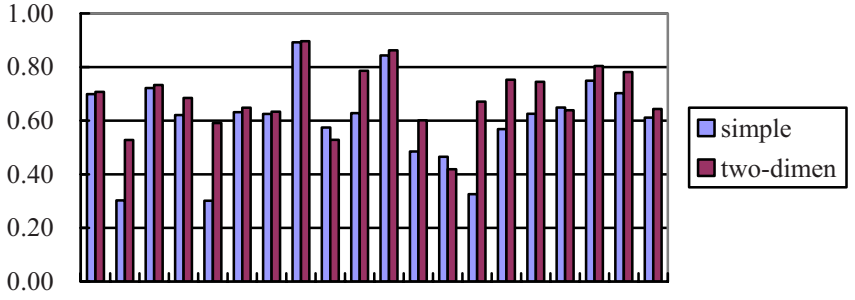

Fig. 1. Comparison of two models.

3 Conclusion

In this paper, we propose to obtain user's access preference by history based user interest modeling. Both simple user interest model and real time two-dimensional interest model are applied in our experiment on 20 users' interest data sets. Experimental results show that history based user interest modeling is helpful in obtaining user's access preference. Furthermore, real time two-dimensional interest model is more effective than simple user interest model in www cache modeling.

Acknowledgments. This study is supported by the Natural Science Foundation of China under Grant No.60473100.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual www search engine. In: Proceedings of 7th world wide www Conference (www'98), Brisbane, Australia, pp. 107–117 (1998)
2. Bao-Wen, X., Wei-Feng, Z., Chu, W.C., Hong-Ji, Y.: Application of data mining in WWW pre-fetching. In: Proceedings of IEEE MSE, Tai Wan, 2000, pp. 372–377 (2000)
3. Wei-Feng, Z., Bao-Wen, X., Chu, W.C., Hong-Ji, Y.: Data mining algorithms for WWW pre-fetching. In: Proceedings of the 1st International Conference on WWW Information Systems Engineering (WISE'2000), Hong Kong, China, pp. 34–38 (2000)
4. Wei-Feng, Z., Bao-Wen, X., Song, W., Hong-Ji, Y.: Pre-fetching WWW pages through data mining based prediction. Journal of Applied System Studies, Cambridge International Science Publishing, England, 3(2), 366–371 (2002)
5. Jia-Hui, H., Xiao-Feng, M., et al.: Research on www mining. Journal of Computer Research and Development,(in Chinese) 38(4) , 405–414 (2001)
6. Bao-Wen, X., Wei-Feng, Z.: www cache based model of users' real time two-dimensions interest. Chinese Journal of Computers 27(4), 461–470 (2004)