Representing Personal Web Information Using a Topic-Oriented Interface*

Zhigang Hua[2], Hao Liu[3], Xing Xie[1], Hanqing Lu[2], Wei-Ying Ma[1]

[1]Microsoft Research Asia
5/F Sigma Center, No. 49, Zhichun
Road, Beijing, 100080, P.R. China
+86-10-62617711
{xingx,wyma}@microsoft.com

[2]Institute of Automation
Chinese Academy of Sciences
Beijing, 100080, P.R. China
+86-10-62542971
{zghua, luhq}@nlpr.ia.ac.cn

[3]Dept of Information Engineering
Chinese University of Hong Kong
Shatin, Hong Kong
+852-98359692
hliu@cuhk.edu.hk

ABSTRACT

Nowadays, Web activities have become daily practice for people. It is therefore essential to organize and present this continuously increasing Web information in a more usable manner. In this paper, we developed a novel approach to reorganize personal Web information as a topic-oriented interface. In our approach, we proposed to utilize anchor, title and URL information to represent content information for the browsed Web pages rather than the content body. Furthermore, we explored three methods to organize personal Web information: 1) top-down statistical clustering; 2) salience phrase based clustering; and 3) support vector machine (SVM) based classification. Finally, we conducted a usability study to verify the effectiveness of our proposed solution. The experimental results demonstrated that users could visit the pages that have been browsed previously more easily with our approach than existing solutions.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications – *Information browsers*;

General Terms: Design, Human Factors

Keywords: Personal Web information, clustering, topic classfication, user information mining, user interface

1. INTRODUCTION

Nowadays, people have easy access to the World Wide Web and Web activities go pervasively into daily life. Consequently, the information on the past Web activity continues increasing. The personal Web information is valuable for individuals, as is shown by many studies that there exist common needs for users to revisit a site or a page that has been visited previously [3].

Revisitation support is commonly provided by the current Web browsers such as Microsoft Internet Explorer, Opera Explorer and so on, in forms of history, bookmarks, URL auto-completion, address bar menu, and back and forward buttons [3]. However, they either require manual maintenance or lack a convenient interface. This is far from being satisfactory for users in some circumstances. For example, users want to search for similar pages from past Web information with a specific topic. Such a task becomes more tedious when the Web information volume increases with daily Web activities. The issue is then raised on how to represent such a large volume of personal Web information in a more usable interface. To our knowledge, few existing works have involved the issue of organizing the large volume of personal Web information using a topic-oriented interface.

In this paper, we developed a novel scheme which reorganized Web information as a topic-oriented interface. We investigated three methods to represent users' Web information: 1) top-down statistical clustering; 2) salience phrase based clustering; and 3) support vector machine based classification. We conducted a usability study to verify the effectiveness of our new approach.

2. A TOPIC-ORIENTED APPROACH

In this section, we first present our method to represent the content information of Web pages, namely anchor, title and URL. We then describe two representation styles to organize the personal Web information as a topic-oriented interface, namely the clustering and classification approaches.

2.1 Content Information

There exist multiple types of representations for a Web document as shown in [4]. These representations typically contain titles, anchor texts, URLs, and main body texts. A title provides the main idea and the brief explanation of a Web document. An anchor text provides the description of linked Web documents and files. It was pointed out that anchor text often provides more accurate description of a Web document than the document itself.

We don't adopt content body as a kind of source information in our method. The reasons are multifold: 1) content body may contain privacy or security information, such as e-commerce account information, email login information, etc; 2) content body of a Web page often contains noise information, such as advertisements, etc. Although there exist many studies involving noise elimination such as page segmentation, they usually cause processing overload and increase time complexity; 3) content body contains vast texts that will increase processing load; and 4) Web objects with *multimedia* or *others* type have no text body besides binary stream. Above all, we believe anchor, title and URL information are sufficient to describe content information.

2.2 Representation Metrics

In this sub-section, we propose three methods, i.e. SVM based classification, top-down statistical clustering, and salience phrase ranking based clustering to organize personal Web information.

[*] This work was done when Z. Hua and H. Liu were interns at Microsoft Research Asia.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2.2.1 *SVM-based Classification (SVM)*

SVM is a powerful learning method [2]. It is well founded in terms of computational learning theory and has been successfully applied to text categorization [2]. In our implementation, the feature of each document is selected as the word weight vector that is specified by the mixture model described above. The topic classifier employs the linear support vector machine to assign each page a probability of being in each category. We took the most likely category as the topic of the page. We collected the category from the *Yahoo! Directory* page, which is available in the URL of http://dir.yahoo.com/.

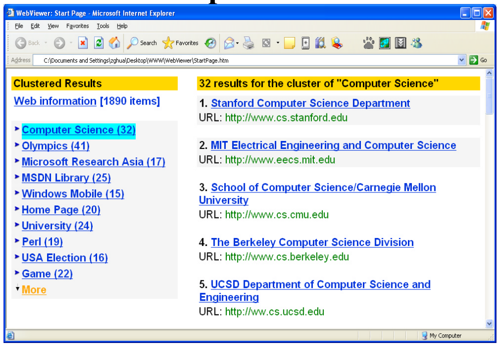
2.2.2 *Clustering*

Clustering methods don't require pre-defined categories as in classification methods. Thus, they are more adaptive for various queries. It has been proved that organizing Web search results into clusters facilitates users' quick browsing through search results. Since it enable users to identify their required group at a glance. In our work, we explored two clustering methods: 1) top-down statistical clustering; and 2) salience phrase based ranking.

Top-Down Statistical Clustering (Top). We adopted Crabtree and Soltysiak's vector space clustering model [1], where each document is represented as a vector of term weights. The weight of a term *i* in a document *j* is calculated according to the mixture model. In our work, we sum the $w_{i,j}$ scores of all documents in each cluster and adopt the top twelve most highly scoring terms to present these words as the "essence" of interests. Cluster vectors use the summing $w_{i,j}$ scores of top 12 terms, and the same clustering technique is applied on the cluster vectors to produce "theme vectors". These represent the overall themes of interests that a user might have, and probably align reasonably well with keywords that a user might give to present his/her interests.

Salience Phrase Based Clustering (Sal). We also take a new method called salience phrase clustering proposed by Zeng [5] to organize the browsed Web pages. This method generates short and readable cluster names, which enable users to quickly identify the topic of a specified cluster. Further, the clusters are ranked according to their salience scores, thus the more likely a cluster required by users, the higher it ranks.

2.3 Document Group Visualization


WebViewer: Start Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
C:\Documents and Settings\zghua\Desktop\WWW\WebViewer\StartPage.htm
Go
My Computer

Clustered Results

Web information [1890 items]

- Computer Science (32)
- Olympics (41)
- Microsoft Research Asia (17)
- MSDN Library (25)
- Windows Mobile (15)
- Home Page (20)
- University (24)
- Perl (19)
- USA Election (16)
- Game (22)
- More

32 results for the cluster of "Computer Science"

1. Stanford Computer Science Department
URL: http://www.cs.stanford.edu

2. MIT Electrical Engineering and Computer Science
URL: http://www.eecs.mit.edu

3. School of Computer Science/Carnegie Mellon University
URL: http://www.cs.cmu.edu

4. The Berkeley Computer Science Division
URL: http://www.cs.berkeley.edu

5. UCSD Department of Computer Science and Engineering
URL: http://ww.cs.ucsd.edu

Figure 1. An example for a start page

We apply our approach to present a start page in Web browser to facilitate user's navigation and revisitation. Incidentally, we take the topic-oriented groups generated by the salience phrase based clustering method as an example in the following description. As shown in Figure 1, each of document groups is formed as a set of nested <frame>, which includes a name that describes the group, a set of items and the number of these items. For each item in a group, we use its anchor text (or title if anchor text is not available) and URL to represent it.

3. EXPERIMENTAL EVALUATIONS

We set the methods that organize Web information according to the visiting time (*time*) and Web site (*site*) as the baseline, to evaluate the various representations generated by our topic-oriented approach. In our study, we added an additional pane so as to elicit feedback of different methods whenever the user viewed the representations on a simple rating principle, reflecting how "pleased" the user was with the representation results, ranging from 1, i.e. "not pleased at all" up to 5, i.e. "very pleased". We average the users' feedback in Table 1. The results are also straightforward: 1) the baselines that generate time-sorted or site-oriented representation is lowly rated with an average score of 2.79 and 2.32 respectively; 2) two clustering representations, *sal* and *top*, are rated with a higher score, 3.13 and 3.35 respectively; 3) *svm* is ranked with the lowest score of 2.08, because most of a users' Web information tends to fall into one or several fixed categories. The users' qualitative evaluations on our new approach indicated that users felt more pleased with our method.

Table 1. The evaluation of various representation methods

Methods	score
Site	2.32
Time	2.79
SVM	2.08
Sal	3.13
Top	3.35

4. CONCLUSIONS

We developed a novel approach to reorganize personal Web activity information in a more usable topic-oriented interface. We investigated three methods to represent users' Web information: 1) top-down statistical clustering; 2) salience phrase based clustering; and 3) support vector machine (SVM) based classification. Experimental results demonstrated that personal Web information can be better organized in our solution. Users felt that topic-oriented clustering representation can facilitate the access to past Web information in comparison with the existing solutions.

5. REFERENCES

[1] Foner L. A multi-agent referral system for matchmaking. Proc. of the 1st Int. Conf. on the Practical Applications of Agents and Multi Agent Systems, London, UK, 1996.

[2] Joachims T. Text categorization with support vector machines: learning with many relevant features. Proc. of the 10th ECML, Chemnitz, Germany, April 1998.

[3] Milic-Frayling N., Jones R., Rodden K., Smyth G., Blackwell A. F. and Sommerer, R. Smartback: supporting users in back navigation. Proc. of the 13th International World Wide Web Conference, New York, USA, May 2004.

[4] Westerveld T., Kraaij W., and Hiemstra D. Retrieving Web pages using content, links, urls and anchors. In Text REtrieval Conference (TREC-10), pages 663–672, 2001.

[5] Zeng H.J., He Q.C., Chen, Z., Ma, W.Y., and Ma. J. Learning to Cluster Web Search Results. Proc. Of the 27th Annual International ACM SIGIR Conference, Sheffield, UK, July 2004.