Improved Web Searching through Neural Network Based Index Generation

Xiaozhe Wang, Damminda Alahakoon, and Kate A. Smith

School of Business Systems, Faculty of Information Technology,
Monash University, Clayton, Victoria 3800, Australia
{catherine.wang,damminda.alahakoon,
kate.smith}@infotech.monash.edu.au

Abstract. In this paper we propose a method to improve web search results in search engines. The Self Organizing Map is used for clustering query logs in order to identify prominent groups of user query terms for further analysis. Such groups can provide meaningful information regarding web users' search interests. Identified clusters can further be used for developing an adaptive indexing database for improving conventional search engine efficiency. The proposed hybrid model which combines neural network and indexing for web search applications can provide better data filtering effectiveness and efficiently adapt to the changes based on the web searchers' interests or behaviour patterns.

1 Introduction

Understanding the needs, interests and knowledge of the web users has grown to be an important research area in the recent past [9]. Web mining is a popular method which is being used for this purpose. Web mining includes web usage mining, web content mining and web structure mining [2,12]. Web log files are the basic data source for extracting useful information in web usage mining, and web documents are used in the web content mining process [11]. Our previous work has described several useful techniques of web mining using soft computing techniques [14,15]. This paper describes our current work towards developing an integrated hybrid system which can make use of web mining results to provide improved functionality to web users.

With the hybrid approach, we propose to build an adaptive web searching system by combining neural network based clustering, cluster analysis and a dynamic indexing technique. The proposed system makes use of the discovered knowledge from data clusters of web query logs to develop an index for more efficient web searching. The novelty of this method is that the index would be updated routinely when the web query patterns change. In this paper we demonstrate the cluster generation and analysis techniques with experiments. In addition we propose an algorithm for identifying change (or shift) in web search patterns by recognizing *movement* and *shift* in data

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2659, pp. 151–158, 2003.

clusters. Such change identification can then be used to update the index with current trends. The development of this complete adaptive system is presented as on-going work.

The rest of the paper is organised as follows. Section 2 describes the proposed hybrid system for adaptive query searching through web log mining. Experimental results on web query term clustering is provided and discussed in Sect. 3. Section 4 concludes the paper with details of ongoing and future work.

2 A Hybrid System for Improved Web Searching

The hybrid approach to developing the adaptive web searching system is as follows. The query log data are initially collected and pre-processed to convert to a suitable format for clustering and analysis. The pre-processing is described in Sect. 2.1. The pre-processed data are then clustered using the Self Organising Map (SOM) algorithm. The SOM is used due to its unsupervised nature and also its ability to handle high dimensional and large data sets [3]. The clustering is used to identify web searchers' query interest patterns from web query log data and further to create pattern files identifying different characteristics from clustered data results [8]. Results are used to develop a web documents indexing database. Since the web users search interests can change over time the index needs to adapt with the changes. We propose a technique for identifying change in data (Sect. 2.3) and such change can then be used to update the current user search terms index. The approach proposed in our research is represented in Fig. 1. As such the system could be dynamic instead of static with the new data and changes.

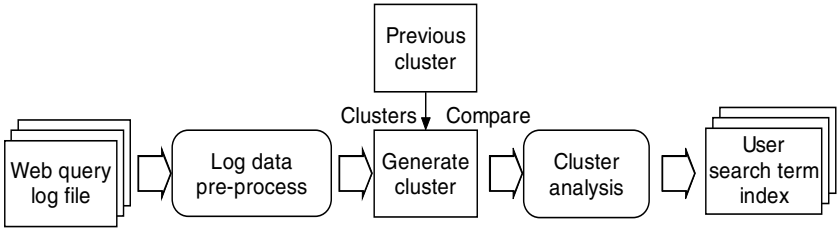

Fig. 1. Dynamic web searching system

2.1 Data Collection and Pre-processing

Query Log File Data Collecting. Experiments described in this paper were conducted with data acquired from Monash University's web query log files located at the main server in Melbourne [7]. The query logs record all past search query entries requested by web searchers on a weekly basis.

Query logs used cover the time range from 00:00:49 14 July 2002 to 00:00:04 25 August 2002 that represent 6 weekly log files in text format. The size of the weekly query log files varies from 2.6 megabytes to 4.1megabytes and consist of 35322 to 61745 query entries. Each entry in the initial log file records a single query, the *date and time stamp*, *number of documents retrieved* and *original query input*. The unprocessed logs contain data of the following form as shown in Table 1:

Table 1. Format of unprocessed query entries in web log files

Date	Time	Result	Query entry
2002/08/18	00:00:36	6	'http://www.adm.monash.edu.au/sss/handbook'
2002/08/18	00:00:39	75	'clayton map CSE2002'
2002/08/18	00:01:39	7854	'oversea fee'

Query Term Data Set Pre-processing. In order to construct a training set for the clustering algorithm, the data from the original log files is passed through the following three stages of pre-processing: *session identification*, *term identification* and *data set identification*.

Session Identification. Initially, non-null queries were extracted from the original log data and the data set session used for processing was identified. Time series data is a common feature for web log data. A session of the data set can be identified based on different time frames according to the difference in the analysis purpose. Since the university runs on a semester basis, and within each semester, the schedule is arranged on a weekly basis, in this research, a week is identified as the session time frame for data set determination. A frequency rule also was used to ensure that normal patterns could be obtained such as the distribution of data frequency on a weekly basis.

Term Identification. The text terms entered by web searchers represent their search topics. Each single term in the log data needs to be identified to process further analysis procedures. From the initial sectionalized log data, each text term and its frequency (number of appearances) is extracted, and the frequency is used as a weight to build up a vector for each term during the clustering process. For example, from a weekly log of 51474 query entries, 15735 text terms were extracted with occurrence values differing from 1 to 5084. From the huge amount of terms, there were only a limited number of terms representing common interests from the majority of web searchers. A high-frequency rule was applied in term identification by comparing it with a predefined threshold. If we denote the *total query number* as Q_t, the *null query number* (queries with no results) as Q_n and the *identifying parameter* as IP, the value of is experimentally set in the range $IP \in [0.0001, 0.01]$, so that 0.001 means that a term must appear in more than 0.1% of the query entries. In the experi-

ments, by using a *frequency threshold* $T_f = (Q_t - Q_n) * IP$, for instance, if a log file has Q_t and Q_n numbers are 51474 and 48 respectively, the frequency threshold was 51 by using $IP = 0.1\%$. As a result, 271 terms survived after this selection.

After data pre-processing, each session had quite similar number of text terms with a stable trend after different steps as shown in Fig. 2, which reveals that there are a certain number of common query terms requested by most of web searchers all the time. It would be interesting to see whether same procedures could apply for other web environments as well.

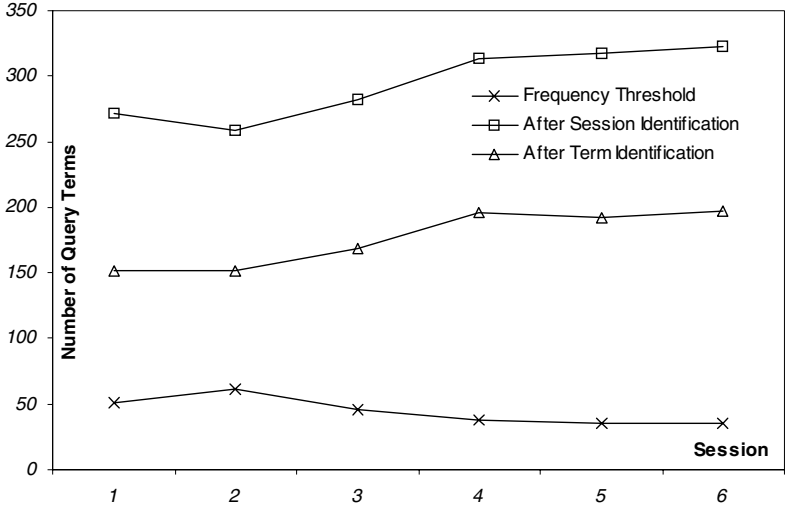

Fig. 2. Comparison of number of terms selected in different steps of data pre-processing

Data Set Identification. Since only the single text terms are extracted for analysis, noisy data including URLs and different versions of the same word are removed before forming the data set for further processing. From the experiments results, different sessions had a different number of query terms. After deciding the number of sessions used for forming a processing data set, all the terms were merged as a group in the identified data set. For instance, the number of query terms after the term identification step in 6 sessions of logs covered from 14 July 2002 to 18 August 2002 varies from 152 to 197. After the merging step, there were finally 218 terms selected in the identified data set. Then the query terms were represented as high-dimensional vectors of sessional *weights* w. The value of dimension was decided by the number of sessions used in data set. If n sessions of data were used in the data set, the *query term* (T_q)'s occurrence value in each session was w_i, and the query term vectors

represented as $T_q = (w_1, w_2, ..., w_n)$. The value of the dimension increases when more sessions of data are used for the process.

2.2 Clustering of Web Query Terms

Clustering-based algorithms have gained a lot of attention lately and have been used in a variety of applications. SOM has been successfully used in a number of web mining projects [4][6] as one of the most appropriate techniques for the Web documents clustering process because of its strength in both flexibility of grouping and visualization for the clusters [3]. With web data, the related transaction entries are grouped into the same cluster and the relationship between different clusters is explicitly shown on the map.

2.3 Data Shift and Movement Identification Technique

Since the proposed system requires adaptation to user trend changes, the following method is suggested to identify such changes in data. It is proposed that any changes in data will be reflected in the subsequent clusters and as such an algorithm [1] for comparing clusters is made use of.

A measure called the cluster error (ERR_{cl}) between two clusters in two SOMs is defined as: $ERR_{cl}(Cl\_j(MAP1), Cl\_k(MAP2)) = \sum | A\_i(Cl\_j) - A\_i(Cl\_k) |$ where Cl_j and Cl_k are two clusters belonging to $MAP1$ and $MAP2$ respectively and $A\_i(Cl\_j)$, $A\_i(Cl\_k)$ are the attribute of clusters Cl_j and Cl_k .

We define two clusters similar when the condition $ERR_{cl}(Cl\_j, Cl\_k) \le T_{ce}$ is satisfied, T_{ce} is the threshold of cluster similarity ($0 < T_{ce} < D/2$, where D is the dimension of the data) and has to be provided by the data analyst depending on the level of similarity required. If complete similarity is required, then $T_{ce} = 0$.

Since the similarity between two clusters depends on the T_{ce} value, we define a new indicator I_s called 'the measure of similarity' which indicates the amount of similarity when two clusters are considered to be similar. Then I_s is calculated as the fraction of the actual cluster error to the maximum tolerable error for two clusters to be considered similar $I_s = 1 - ERR_{cl}(Cl\_j, Cl\_k) / Max(T_{ce})$.

3 Experimental Results on Query Log Analysis

We used the Viscovery SOMine [13] to simulate the SOM. Data set contains 6 sessional frequencies were used as weights to form attributes for query term vectors from the pre-processing step was fed in to SOM algorithm [5] and clustering results were obtained after unsupervised learning.

3.1 Web Users' Query Interest Pattern Analysis

After feeding the pre-processed data set into the SOM, the most natural clustering results were discovered with a stable number of clusters. The 4 clusters were separated based on their own characteristics and the pattern of each cluster used to analyse for obtaining a better understanding of Web searchers' behavior (the fundamental knowledge to further build up online adaptive searching). Based on the degree of the query entry frequency and the time frame, we found interesting information representing each cluster, as illustrated in Fig. 3.

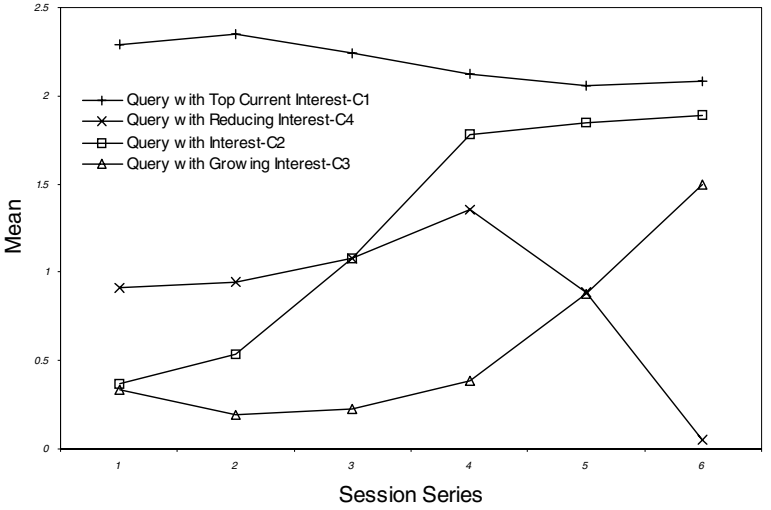

Fig. 3. Searching query term movement pattern of 4 categories from data set's clustering result

By computing the frequency degree and the time based trend change, the web searcher's interest patterns were categorized as 1) *Query with top in current interest*, 2) *Query with interest*, 3) *Query with growing interest* and 4) *Query with reducing interest*.

In the group 'Query with top in current interest', all terms have the same common feature that they remain the highest request rate most of the time within the whole period. These clustered terms directly reveal the search entries with highest frequency

requested through the embedded search engine, and also represent the web users' current interests contents for the web site. Using this information, terms in this cluster with the highest priority can be used to generate the frequently requested areas index documents for searching.

With 'Query with interest', the frequency rate starts from a very low point and grows very fast to nearly close to the top level. By analysing detailed terms in the cluster, we found that this cluster could be also categorized as 'Query with growing interest'. Terms representing a growing interest trend from the web users for a longer period of time suggest a high possibility for them to move in to upgrade cluster in the coming period.

In the patterns for categories of 'Query with growing interest' and 'Query with reducing interest' are very similar. In the first half of time period, the frequency rate shows very minor change, but after the middle point, both the growing and reducing movements become significant until the end. For the growing interest trend, requests grow to almost equal to half of the top level, and the reducing trend moves to almost 0. The web documents' topics associated with those terms can be indexed in the temporary indexing database based on the movement of web user's interest change.

3.2 Adaptive Web Query Searching

To achieve dynamic and adaptive web query searching [10], 2 components are included in the hybrid system proposed in our research. Offline web query term pattern files construction and online web document indexing. From the clustered data, different features and characteristics are demonstrated as shown in Fig. 3. As such the experimental results demonstrate the offline part of the system and the associated possibilities.

Generating web document index is considered as the online part, since to be useful, the index has to dynamically adapt to the changes in the user search patterns. This online system uses the cluster comparison and change identification technique presented above and experiments are currently being carried out to ascertain the advantages of the method.

4 Current Work and Future Research Direction

In this paper we have proposed a system for analysing and improving web searching by using web query log data. The system described is in two parts as an offline data analysis part and an online dynamic index building and searching part. Experimental results were presented to demonstrate the potential of the system after detailed descriptions of the pre-processing of data has been presented. It is required to run similar experiments on several consecutive time (session) log files to confirm the apparent potential from this method. The terms in each cluster will need to be compared with terms in the clusters in the next session to identify the trends such as current interest terms losing interest over time.

References

1. Alakakoon, L. D.: Data Mining with Structure Adapting Neural Networks. PhD thesis, Monash University (2000)
2. Chang, G., Healey, M. J., McHugh, J. A. M., Wang, J. T. L.: Web Mining, Mining the World Wide Web. Kluwer Academic Publishers, Chapter 7 (2001) 93–104
3. Flakes, G. W., Lawrence, S., Giles, C. L., Coetzee, F. M.: Self-organization and Identification of Web Communities. IEEE Computer, vol. 35, no. 3 (2002) 66–71
4. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: WEBSOM – Self-organizing Maps of Document Collections. Proceedings of Workshop on Self-Organizing Maps (WSOM'97), Espoo, Finland (1997) 310–315
5. Kohonen T.: Self-Organizing Maps. 2nd edition, Springer, Heidelberg (1997)
6. Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., Saarela, A.: Self Organization of a Massive Documents Collection. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, vol.11, no.3 (2000) 574–585
7. Monash University Web Site, <http://www.monash.edu.au>
8. Ng, A., Smith, K. A.: Web Page Clustering Using A Self-Organizing Map of User Navigation Patterns. Smart Engineering System Design: Neural Networks, Fuzzy, Logic, Evolutionary Programming, Data Mining, and Complex Systems, Missouri, USA (2000)
9. Paliouras, G., Papatheodorou, C., Karkaletsisi, V., Spyropoulous, C. D.: Clustering the Users of Large Web Sites into Communities. Proceedings of the 17th International Conference on Machine Learning (ICML'00), Stanford University (2000)
10. Perkowitz, M., Etzioni, O.: Towards Adaptive Web Sites: Conceptual Framework and Case Study. Computer Networks, Amsterdam, Netherlands, Artificial Intelligence (2000)
11. Pirolli, P., Pitkow, J., Rao, R.: Silk From a Sow's Ear: Extracting Usable Structures from the Web. Proceedings on Human Factors in Computing Systems (CHI'96), ACM Press (1996)
12. Srivastava, J., Cooley R., Deshpande, M. Tan, P. N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. 1, no. 2 (2000) 12–23
13. Viscovery SOMine, <http://www.eudaptics.com/technology/somin4.html>
14. Wang, X., Abraham, A., Smith, K. A.: Soft Computing Paradigms for Web Access Pattern Analysis. Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1 (2002) 631–635
15. Wang, X., Smith, K. A.: Clustering Web User Interests Using Self Organising Maps. Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Soft Computing Systems: Design, Management and Applications, IOS Press, the Netherlands (2002) 843–852