Query Result Aggregation on Multiple Search Engines

Chih-Lu Lin and Hung-Yu Kao
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
{p7693105, hykao}@mail.ncku.edu.tw

ABSTRACT

As the rapid development of the network environment in recent years, we could get more and more Web resources, however some problems happened as followed, e.g., Lacking of the effective method of finding the Web resources. This problem is solved as the birth of the search engines, but there are some other problems and issues needed to be resolved.

For some examples that will be mentioned in this paper: (1) for a short query, it is difficult for search engines to understand what users' goal of the Web search. As a result, search engines are difficult to provide Web resources that are related to users' search goal. (2) There is no effective method for helping the users to find their information need among search engines' enormous indexes. Therefore, this paper will focus on continuing and improving the previous work about clustering on search engine returned results, and also try to study the suitability of pre-deciding that whether the query should be clustered or not, in order to avoid additional overheads both the search engines and users.

1. Introduction

As the improvement of the information and network techniques, we can find almost all of the information and data, which we need, from the network resources. This advantage of network contains enormous data also makes some problems occurred for finding out the desired data on the contrast. Fortunately, problem of searching the needed data and information had been solved by well-developed technique of search engine, such as "Google" [18] is an excellent solution.

However, the large amount of results from search engines are in general unreadable or redundant for users. Besides, for general users, they are used to type short terms for querying. This is an important issue that most search engines are well-designed for provided the best results, and the best results are usually decided by the search engines' own ranking algorithm. Getting the idea of the user's goal is really a tough work for designers of the ranking algorithm. Especially under the circumstances of just the short terms were provided as judging information, the works that provide suitable results for users are getting more and more harder. This paper also focuses on this issue, and we want to provide a method for search engine that gets rid of knowing the users' goal by presenting their results in several clusters to users. Actually, there are some real cases for explanation. We take the search engine "Vivisimo" [20] as an example in Figure 1.

Figure 1. An example query in VIVISMO[20]

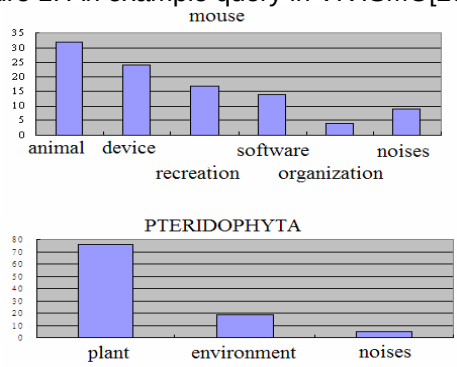

Figure 2. A human-labeled cluster's static graph of the two queries

In order to solve those problems that are mentioned above, we choose to improve the method of clustering and took it as our primary solution for such problem. In our research, we first concentrated on finding method to determine that whether a query is suitable for clustering or not from different important features of Web resources. For making it more clearly, we take Figure 2 as example to explain. Figure 2 shows that answer set of the human evaluators' categorizing results from 100 Google's search results for each query. Obviously, it is easier for user to find wanted search result in query

"PTERIDOPHYTA", and we think query "mouse", which has more clusters, that should be clustered more. So, we want to find the relation between query, like "mouse", that should be clustered and features of query in our first step.

Another important point, that we should mention here, is the original clustering method. We take the clustering method, which is proposed in the paper of Zheng [15], as our basic method for further improvement. Not only improving the method by adding new measuring features, but also combining new method to original method. Like what we mentioned above that checked the suitability of clustering for query is one of the new methods and an experiment of researching in DF (Document Frequency) of N-GRAM ($1 \leqq N \leqq 3$).

By following the original flow of the proposed method of Zheng [15], we do a similar experiment for checking whether the accuracy of taking Chinese search results as input data is good as the English search results, like he showed in paper [15], or not.

By observing the experiment results, apparently, it works not well in the results of the Chinese searches. As we known, it is a hard work to extract reasonable phrases of the Chinese sentences, and this influences on calculating values of selected features. This may be the reason that low accuracy occurred under the environment of the Chinese search. In order to solve this problem, we select new features for improving the accuracy. By observing the structural information of URL, we define two features that are TUD (Total URL Depth) and TDC (Total Domain Count). Intuitively, important keywords or phrases are usually placed in upper level of the website.

Furthermore, we evaluate the effect of taking the search results of different search engine as source data of clustering. In this paper, we focus on improving the accuracy of taking Chinese search results as input data of the proposed method of Zheng, predicting whether a query should be clustered or not and providing a better way for taking search results of different search engines.

2. Related work

Like we mentioned in Section 1, a well-developed search engine would be proud of the number of webpage. It is reasonable that the more crawled webpage does a search engine has, the more information and Web resources can it provide in point of search engine's view. But this will induce a situation that may make traditional simple way to show results directly in a webpage no longer suitable for general users.

Some recent researches interested in studying the "goal" behind a user's Web query. At first, researchers focus on the observation of the search log [1] [14] and the concept of the query type was proposed in [5], and then user goals were proposed in [12] [13]. We could also find the application of user goal in analyzing anchor text in [3] [6].

Most users are trying to restart the query with another similar short query terms after viewing 20 results at most. Short query is a tough problem for search engines that causes search engines to get no idea of user's goal. Search engine are easily caused to provide redundant or non-related results without getting the idea of user's goal. Moreover, information of a short query is too few for judging the user's goal. So we would consider the possibility of improving present existed method for solving this problem.

Besides, the Zheng's research [15] was a new idea of transforming the original content-related clustering problem to another problem of finding the salient phrases. It should not to be considered as a drawback of ignoring the property of content-related, because it would not affect the search behavior, said by Zheng. We agree with what Zheng said, so we try to improve Zheng's method and apply our proposed method to the Chinese search.

2.1. Overall introduction of previous methods

There were a lot of researches about this topic, and major representative methods of them actually could be categorized as followed:

2.1.1. Traditional methods. Methods of this kind primarily focus on finding documents [1] [4] [9] [10]. Basically, this method generates clusters first according to the similarity of each document in data source. For making it more clearly, we should emphasize that all documents would finally belong to one of those generated clusters. Causing this method isn't applied to cluster Web search results at the beginning, the researches of generating a more readable seems to be ignored. But a readable cluster name, for users, is very important for search engine. As a result, it's not the best choice for us to follow.

2.1.2. Extended version of traditional methods. Adding concept from other researches, such as topics finding and text trend analysis [7] [11], to traditional methods. Basically, methods, which belong to this kind, mainly focus on improving the correctness of clustering.

2.1.3. Suffix Tree Clustering (STC). By using the suffix tree structure, this method first analyzes and generates the phrases that contain same suffix terms [16] [17]. Clusters will be generated according to those generated phrases.

2.1.4. Clustering under the network environment or present search engine. Actually, clustering methods, which belong to this kind, can not be categorized to each of above methods along. Because each of them are not enough to deal with the issues under this situations, such as data source is set of snippets instead of documents, generated cluster name should be able to be readable for search engine users and etc.

Fortunately, if we mixed those present methods' advantages then we could be able to solve those existed issues properly. Although this method is suitable to cluster the Web search results, new issues show up. The

new issue is that some element of data source may belong to none of the final generated clusters, which would be provided to users, such as method of Zheng [15]. From experiments' results of Zheng, it showed that the average coverage ratio was sixty percent.

2.2. Description of Zheng's method [15]

Our method is followed the thought of his proposed method, so we would briefly describe his method. Zheng chooses the result of MSN search [21] as source data of cluster. For each query, he would crawl 200 results. From his experiment, 200 results are the most proper number for fetching back. Zheng selected five features, i.e., TF-IDF, Phrase Length, Intra-Cluster Similarity, Cluster Entropy and Phrase Independence, and calculation soon after document parsing. Zheng then want to find out the relation between x = (TFIDF, Length, ICS, PI, CE) and y. Finding the most similar relation was the destination of this task. Zheng's work showed that the transformed cluster problem was basically linear. So, we choose the SVR with linear kernel as our ranking tool for finding salient word. Another reason for choosing SVR as our ranking method is that SVR is simple to use but effective.

3. Experiment

The system flow in the Figure 3 is the architecture graph of the proposed system. The bold blocks and objects are our work on this topic. We first mentioned the experiment of DF analysis, and then focus on the clustering experiment.

3.1. Pre-clustering analysis

We hope to conclude a method for deciding whether a query is suitable for being clustering or not by analyzing the DF value of each N-GRAM term. However, we are not able to find a better method with highly precision of predicting, for the reason that it is not an easy work that generating an objective answer set of enough number for English query for Chinese evaluators.

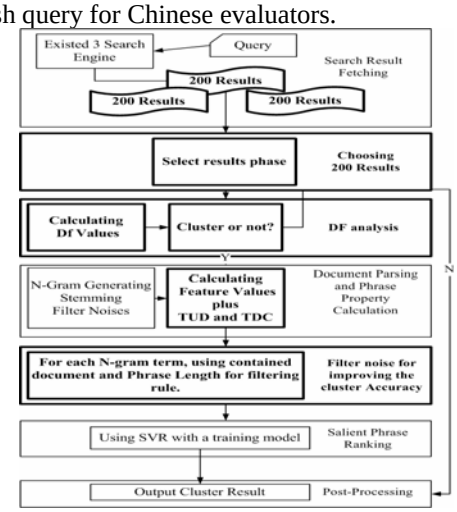

Figure 3. The system flow

We simply use number to represent the document id and cluster id. In general case, a result contains title and description, and about 100 N-Gram terms would be generated after the procedure of generating N-Gram terms. We will also use a human-labeled answer set for comparing the difference of the DF distribution curve between the real query and the simulated one.

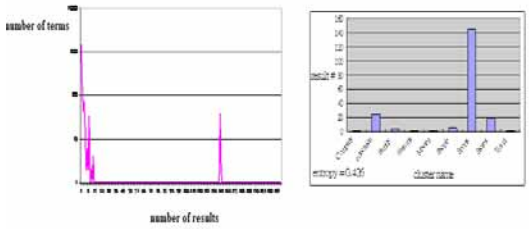

Figure 4. Compare with the real searching data

In real case, like Figure 4, one more peak occurs. This is not the same with the result of the ideal situation. After checking the experiment, we though there are two possible way that could be lead us to this situation. One is the answer set and the other is the DF value itself. First, the answer set is totally English and may be too tough for labeling the correct answer. As for taking the other feature value for curve analysis might need some more experiment for evaluating.

A query log from the dataset of the KDD CUP CONTEST 2005, and the query log are composed of 800'000 queries. We first calculate the TF (Term frequency) of each single word and select Top 30 high frequency keywords by ranking the TF as a high frequency term set. After the high frequency term set is generated, we will choose 30 queries from the query log of the KDD CUP CONTEST 2005. Then we fetch the results from Google as experiments' data set for each of the 30 queries. We use the matrix C to calculate the diversity.

$$C = \sum_{X \in Df} \left(\log\left(X_{i+1}^2\right) - \log\left(X_i^2\right) \right)$$

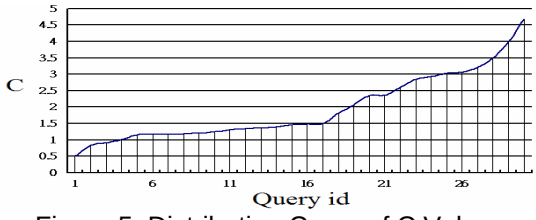

Figure 5. Distribution Curve of C Value

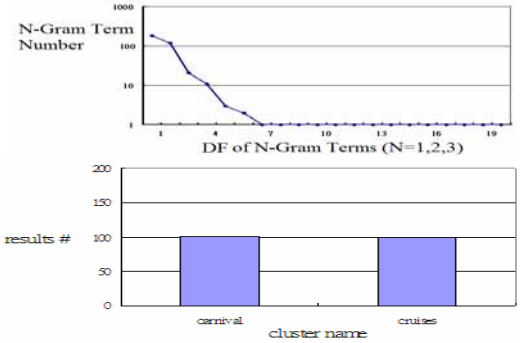

Figure 6. DF Curve and Human-Labeled Answer of Query 1

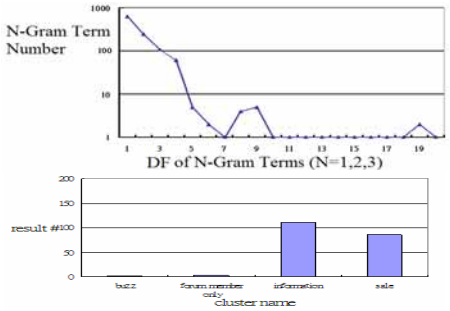

Figure 7. DF Curve and Human-Labeled Answer of Query 15

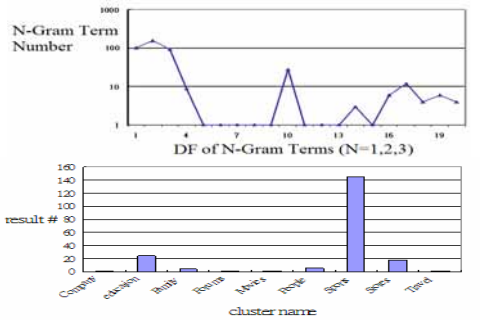

Figure 8. DF Curve and Human-Labeled Answer of Query 29

From Figure 5~Figure 8, we could say that the more the query near the left side of the curve figure in the top, the more smooth the query's DF distribution curve would be. But some query having too much terms, and this maybe a reason that their corresponding DF distribution curve looks roughly.

3.2. The proposed clustering method

According to some previous researches [15], it was effective that applied support vector regression, which is a useful and common technique in the domain of machine learning, to the issue of clustering the Web search results. By adopting effective features and taking multi-sourced search results as dataset, we hoped to improve not only the accuracy of prediction but also the practicability of clustering the search results in Chinese. In order to test the efficiency of my method, two kinds of experiment were done.

By using the weekly ranking list of search keywords, which was provided by the search engine YAHOO, we randomly took 20 keywords from the lists as query terms. For each query, the corresponding results at search engine Google, MSN and YAHOO were crawled separately.

In research of Zheng, it is pointed that about 200 results could be a better choice for consideration of accuracy. So, for each search engine, top 200 results were crawled. As we know, a result contained title, description, URL and some other information. After the step of crawling results for each query, the step of generating N-GRAM ($2 \leqq N \leqq 5$) terms would be processed for each result. Each generated N-GRAM ($2 \leqq N \leqq 5$) term was a candidate cluster name of corresponding results.

Table 1. The chose query for Chinese search

apple	中華電信	台鐵	無名小站
foxy	天堂官方網站	台灣論壇	魔獸世界
linex	巴哈姆特	平水相逢	楓之谷
Nba	史萊姆	勁舞團	
小遊戲	史萊姆好玩遊戲區	統一發票	

In order to evaluate the accuracy of my method, an answer set of 18 real queries from 3 search engines (Yahoo [19], MSN [21] and Google [18]) must be generated. Table 1 was all query keywords, and these 18 keywords were chosen from weekly ranking hot query list of Yahoo's query log. In average, 30 meaningful keywords were chosen per keyword. In addition, a rule for filtering noise was that TF (Term Frequency) > 1 and length (e.g.: Length of"中華電信" was 4) > 1.

3.2.1. URL structure. For general websites, the index pages are often placed some important messages or information. Based the idea, we proposed that terms at upper level of website were important than those at lower level. In order to make sure the influence of the URL's hierarchy, a relative experiment was setting for checking the importance of the level of its URL hierarchy. Specifically speaking, if the URL of a result was Http://www.google.com/language/index.htm, then the level of this URL, was defined as the number of the symbol of "/" in this URL , was 3.

In general situation, the deeper the result page was and the more specific content did the result page. In order to test the influence of this feature, I prepared an experiment to make sure the chose feature was effective.

3.2.1.1. Total URL Depth (TUD). This is the total sum up of each document's URL depth for corresponding N-Gram terms.

E.g., http://www.ncku.edu.tw/acad/index.htm and http://www.ncku.edu.tw/acad.htm are belonged to N-Gram "A". UD of "http:// www.ncku.edu.tw/acad /index.htm" = the number of "/", after removing the prefix "http://", plus 1.= 3. UD of "http:// www.ncku.edu.tw /acad.htm" = the number of "/", after removing the prefix "http://", plus 1.= 2, Therefore, TUD of A= 3 + 2 = 5.

3.2.1.2. Total Domain suffix Count (TDC). Simply speaking, for each N-gram terms, we count the number of how many different suffixes (E.g.: com, edu, and etc.)

We observed that search results of a query might contain same URL in different search engines. Intuitively, those important results should be provided by the well-designed search engines. So, we assumed that those duplicated URLs are more important than others. For our method, it is a simply way to apply this observation by adding a new feature. We give those search results of duplicated URL high score, and do same experiments by just replacing TDC or TUD with this feature. Unfortunately, the experiment results could not improve the precision.

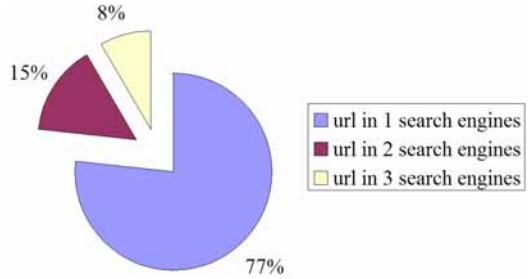

Figure 9. URL distribution of the Chinese search results in the dataset

By observing the experiment's dataset, we can get the idea of the URL distribution. From Figure 9, we can find that most URLs are not duplicated ones. Besides, some of those duplicated search results' content contains too many information to represent a cluster or concept well.

3.2.2. About filtering method of Chinese search. The Chinese Search that we focus here means that the query contains Chinese and English search results. Without using the advanced technique of NLP, we started to think methods that could easily filter most noises in Chinese search. By observing the feature value of those human-labeling training data, we find that some candidate n-gram terms have almost all the same feature values. This situation would have a badly influence to the SVR. In Chinese search, there is a useful heuristic observation that, for choosing N-Gram terms, longer term is better.

3.2.3. Combing Results of Multiple Search Engines. Some researches pointed that even Google just crawled and indexed a small part of the enormous and growing-rapidly existed Web pages. Actually, the search results that we could access from different search engines were amazingly having a low ratio of result-overlapping.

By using results of different search engines, there must were some weighting rules for selecting useful and effective result in consideration of clustering results better.

3.2.4. Experiment Result and Discussion. We use data of each N-Gram terms' feature as training data to train different model with SVR-L (Linear kernel) and combine the features with SVR to get a corresponding single value that represents each N-Gram terms' importance. Then we sort them in a descending order, and the terms with high value are the generated cluster name. By comparing the top 5, top 10 and top 20 generated cluster names with the human-labeled answer set, we could get the accuracy of prediction.

Four features are chosen to find the salient keywords, and the four features are PI, CE, TFIDF and Length. From Figure 10, our proposed method is almost as good as Zheng's. From Figure 11, obviously, TFIDF, CE and PI are a better indicator for ranking salient phrase in Chinese Search. As for the reason of the length seems not to be a good feature in the Chinese search from the Figure 11, are mainly because we use length in our filtering method for Chinese search. In our proposed filtering method, the length is used to decide which N-Gram should be filtered when N-Gram terms contain same documents. By comparing Figure 11 and Figure 12, we both have a better evaluating result in phrase independence.

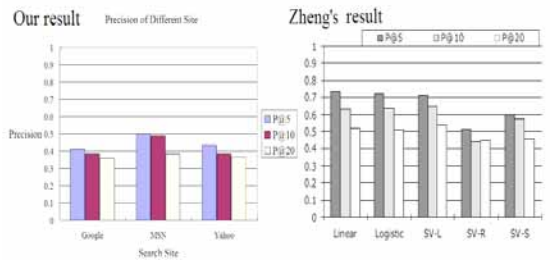

Figure 10. Precision compare of different search engines

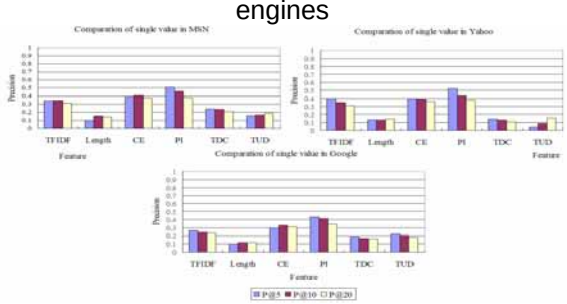

Figure 11. Compare of the single feature in Chinese Search

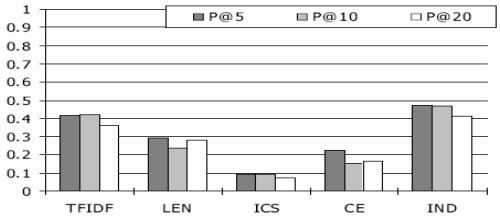

Figure 12. Zheng's experiment result for compare of single feature

In order to compare Precision Compare of Multi-Sourced search, we use six features that are TFIDF, Length, CE, PI, TUD and TDC and our proposed filtering method for Chinese search. By using results of different search engines, we could raise the accuracy of the Chinese search. It is a pity that we could not say which one of the three methods is better, because all the experiment results of the three methods are similar. But we could say that use different results of search engines are good for improving accuracy of our method.

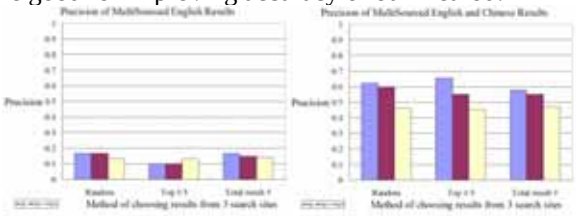

Figure 13. Compare of accuracy in multi-sourced search

4. Conclusion and Future Work

In this paper, we propose some new features and the filtering method that could have a good influence on improving the accuracy. However, there are many works that could be further researched, e.g., Pre-clustering analysis. Another important issue for our method's evaluating is the generation of the answer set. From the experiments accuracy of English search is much lower than Chinese search does. After checking the whole experiment steps, we think the quality of answer set and the amount is important. To sum up, we make it possible to apply the Zheng's proposed method to Chinese searching results by adding new features and using multi-sourced results.

In the future work, we would focus on the improvement of the pre-clustering analysis. If we could pre-decide whether a query is suitable for cluster or not, then we could know more users' goal of the Web search.

Reference

[1] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In Proceedings of ACMSIGKDD '00, 2000

[2] D. R. Cutting, D. R. Karger, and J. O. Pederson. Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections. In Proceedings of the 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), pages 125-135, Pittsburgh, PA, 1993.

[3] C. C. Chang, C. J. Lin, LIBSVM: A library for sup- port vector machines, 2001, Software available at http:// www.csie.ntu.edu.tw/?cjlin/papers/libsvm.ps.gz

[4] N. Eiron and K.S. McCurley. Analysis of anchor text for Web search. In Proceedings of ACM SIGIR '03,2003.

[5] M. A. Hearst, J. O. Pedersen. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), Zurich, June 1996.

[6] I. Kang and G. Kim. Query type classification for web document retrieval. In Proceedings of ACM SIGIR'03, 2003.

[7] R. Kraft and J. Zien. Mining anchor text for query refinement. In Proceedings of the Thirteenth Int'l.World Wide Web Conf., 2004.

[8] D. Lawrie, W. B. Croft. Finding Topic Words for Hierarchical Summarization. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), pages 349-357, 2001.

[9] B. Lent, R. Agrawal, R. Srikant. Discovering Trends in Text Databases. In Proceedings of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining (KDD'97), Newport Beach, California, August 1997.

[10] A. V. Leouski. W. B. Croft. An Evaluation of Techniques for Clustering Search Results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[11] A. Leuski and J. Allan. Improving Interactive Retrieval by Combining Ranked List and Clustering. In Proceedings of RIAO, College de France, pp. 665-681, 2000.

[12] B. Liu, C. W. Chin, and H. T. Ng. Mining Topic-Specific Concepts and Definitions on the Web. In Proceedings of the Twelfth International World Wide Web Conference (WWW'03), Budapest, Hungary, 2003.

[13] U. Lee , Z. Liu , J.H. Cho, Automatic identification of user goals in Web search, Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan

[14] D. E. Rose and D. Levinson. Understanding user goals in Web search. In Proceedings of the Thirteenth Int'l.World Wide Web Conf., 2004.

[15] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Web search engine query log. SIGIR Forum, 33(1):6 – 12, 1999.

[16] H. J. Zheng, Q. C. He, Z. Chen, W. Y. Ma, J. Ma. Learning to cluster Web search results. In Proceedings of SIGIR '04, pages 210–217, 2004.

[17] O. Zamir, O. Etzioni. Web Document Clustering: A Feasibility Demonstration, In Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'98), 46-54, 1998.

[18] O. Zamir, O. Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. In Proceedings of the Eighth International World Wide Web Conference (WWW8), Toronto, Canada, May 1999.

[19] Google, http://www.google.com

[20] Yahoo, http://tw.yahoo.com

[21] Vivisimo, http://vivisimo.com

[22] MSN search, http://www.msn.com