Query Clustering Using Content Words and User Feedback

Ji-Rong Wen
Microsoft Research, China
5F, Beijing Sigma Center
No.49, Zhichun Road Haidian District
Beijing, P.R.China
jrwen@microsoft.com

Jian-Yun Nie
Dept. Informatique et Recherche opérationnelle
University of Montreal
CP 6128, succursale Centre-ville
Montreal (Quebec), H3C 3J7 Canada
nie@IRO.Umontreal.CA

Hong-Jiang Zhang
Microsoft Research, China
5F, Beijing Sigma Center
No.49, Zhichun Road Haidian District
Beijing, P.R.China
hjzhang@microsoft.com

ABSTRACT

Query clustering is crucial for automatically discovering frequently asked queries (FAQs) or most popular topics on a question-answering search engine. Due to the short length of queries, the traditional approaches based on keywords are not suitable for query clustering. This paper describes our attempt to cluster similar queries according to their contents as well as the document click information in the user logs.

Keywords

Query clustering, user feedback, web search, log mining

1. INTRODUCTION

Query Clustering aims to group similar user queries together. This task is important to discover the common interests among the users, and to exploit the experience of previous users for the others. For example, if a number of users have already indicated that a document is related to a query, then this document can be relevant to a similar query of another user.

Different from document clustering where a document can be represented by a relatively large number of content words, queries submitted to a search engine usually are very short. Therefore, a clustering approach using solely the keywords in queries will not be effective.

In this paper, we propose a new approach to query clustering using user logs. In particular, we make use of the cross-references between the users' queries and the documents that the users have chosen to read. Our hypothesis is that there is a strong relationship between the queries and the selected documents (or clicked documents). Our approach is based on the following principle - If two queries lead to the same document clicks, then they are similar or related. Document clicks are comparable to user relevance feedback in a traditional IR environment, except that they denote implicit and not always valid relevance judgments. This principle is used in combination with the traditional approaches based on query contents.

2. SUMMARY OF OUR APPROACH

The current study is carried out on the Encarta encyclopedia which can be accessed on the Web (http://encarta.msn.com/).

2.1 User Sessions

The available data is a large set of user logs from which we extracted queries sessions. A session is defined as follows:

session := <query text> [clicked document]*

Each session corresponds to one query and a series of documents the user clicked on.

2.2 Clustering algorithm

Since query logs usually are very large, the clustering algorithm should be very efficient. Besides, due to the daily-changing property of the log data, it also should be incremental. For these reasons, we choose to use the density-based clustering method DBSCAN [2] and its incremental version IncrementalDBSCAN.

2.3 Similarity Functions

2.3.1 Similarity Based on Query Content Words

Content keywords are the words except function words included in a stop-list. All the keywords are stemmed using the Porter's algorithm [3]. We use the following formula to measure the content similarity between two queries.

$$Sim_{keyword}(p,q) = \frac{KN(p,q)}{Max(kn(p),kn(q))}$$

where *kn(.)* is the number of keywords in a query, *KN(p, q)* is the number of common keywords in two queries.

2.3.2 Similarity Based on User Feedback

A first feedback-based similarity considers each document in isolation. This similarity is proportional to the number of the clicked individual documents in common for two queries p and q as follows:

$$Sim_{click}(p,q) = \frac{RD(p,q)}{Max(rd(p),rd(q))}$$

where *rd(.)* is the number of clicked documents for a query and *RD(p, q)* is the number of document clicks in common.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.
Copyright 2001 ACM 1-58113-331-6/01/0009…$5.00.

In spite of its simplicity, this measure demonstrates a strong ability to cluster semantically related queries that contain different words. Below are some queries from one such cluster:

Query 1	atomic bomb
Query 2	Manhattan Project
Query 3	Hiroshima
Query 4	nuclear fission
Query 5	Japan surrender
……	……

All these queries correspond to the document "ID: 761588871, Title: Atomic Bomb" in Encarta. We observe that no keyword-based approach could create such a cluster.

Documents in Encarta are organized into a hierarchy that corresponds to a concept space. This concept hierarchy allows us to extend the previous calculation by considering a conceptual distance between documents. This distance is determined as follows: the lower the common parent node two documents have, the shorter the conceptual distance between them is. Let $s(d_i ,d_j)$ denotes the conceptual similarity between two documents, the hierarchy-based similarity is defined as follows:

$$Sim_{hierarchy}\ (p,\, q) = \frac{1}{2}\times\left(\frac{\sum_{i=1}^{m}(\underset{j=1}{\overset{n}{Max}}\ s(d_i,d_j))}{rd(p)} + \frac{\sum_{j=1}^{n}(\underset{i=1}{\overset{m}{Max}}\ s(d_i,d_j))}{rd(q)}\right)$$

2.4 Combination of Multiple Measures

A very similar study [1] has been carried out recently. However, that study rejects the use of content words and relies solely on document clicks to cluster queries. We think that both query contents words and the corresponding document clicks can partially capture the users' interests. Therefore, it is better to use both. A simple way to do it is to combine both measures linearly as follows:

$$Sim_{comp} = \alpha * Sim_{content} + \beta * Sim_{feedback}$$

There is an issue concerning the setting of parameters α and β. In our current implementation, these parameters are to be set manually by the users in order to obtain different behaviors.

3. EVALUATION

We collected one month user logs (about 22 GB) from the Encarta website. From these logs we extracted 2,772,615 user query sessions. Because the query number is too big to conduct detailed evaluation, we chose 20,000 query sessions from them randomly. We tested the following four similarity functions on the 20,000 query sessions:

– keyword similarity (K-Sim),

– similarity using single documents clicks (S-Sim),

– similarity using both keyword and single document clicks (K+S-Sim), and

– similarity using both keyword and document clicks in Encarta hierarchy (K+H-Sim).

In the last two functions, both α and β are set to 0.5.

We use precision and recall to measure the quality of clustering results. Precision is the ratio of the number of similar queries to the total number of queries in a cluster. Recall is the ratio of the number of similar queries to the total number of all similar queries for these queries (both in this cluster and not in). Because no standard clusters or classes are available, we manually checked the queries in every cluster to calculate precision and recall. Figure 1 is the F-measure [4] for the four functions, which illustrates a comparison of their global quality.

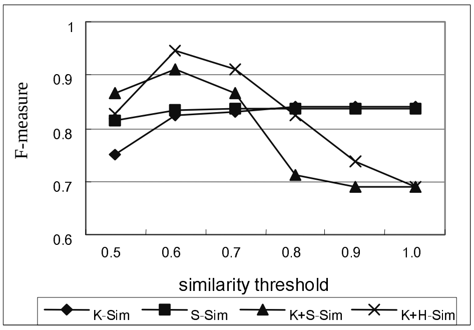

Figure 1. F-measures for four kinds of similarity functions

We can see that between the threshold range of [0.5, 0.7], K+H-Sim and K+S-Sim are better than the single-criterion functions. Especially, when similarity threshold is equal to 0.6, K+H-Sim reaches the highest F-measure value (0.94). This confirms that a proper combination of both query content words and the corresponding document clicks is a better approach to query clustering than using them separately.

4. CONCLUSION

The new generation of search engines for precise question answering requires the identification of FAQs. But the identification of FAQs is not an easy task. It requires a proper estimation of similarity between queries. Given the short length of queries, this similarity cannot be accurately estimated through an analysis of their content words alone. This study demonstrates the usefulness of user logs for query clustering, and the feasibility of building a tool to detect FAQs automatically.

REFERENCES

[1] Beeferman, D. and Berger. A., Agglomerative clustering of a search engine query log, Proc. ACM-SIGKDD, 2000, pp. 407-416.

[2] Ester, M., Kriegel, H., Sander, J. and Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, 1996, pp. 226-231.

[3] Porter, M., An algorithm for suffix stripping, Program, Vol. 14(3), 1980, pp. 130-137.

[4] van Rijsbergen, C.J., Information Retrieval, Butterworths, 1979.