Query-Session-Based Term Suggestion for Interactive Web Search

Chien-Kang Huang
Dept. of Computer Science and Information Engineering,
National Taiwan University, Taiwan
ckhuang@mars.csie.ntu.edu.tw

Lee-Feng Chien
Institute of Information Science,
Academia Sinica,
Taiwan
lfchien@iis.sinica.edu.tw

Yen-Jen Oyang
Dept. of Computer Science and Information Engineering,
National Taiwan University, Taiwan
yjoyang@csie.ntu.edu.tw

ABSTRACT

This paper presents a new effective log-based term suggestion approach for interactive Web search. Using the approach, it is able to suggest more precise search terms for user's query sessions rather than single queries, and the suggested search terms are automatically extracted with query session logs from proxy servers. The achieved performance shows that the proposed approach is very effective in recommending relevant search terms especially for high-frequency queries, and can improve real-word search services in several aspects.

Keywords

Interactive Search, Term Suggestion, Query Context.

1. INTRODUCTION

Web users' queries are often too short to contain sufficient and discriminated keyterms in the process of information retrieval [1]. For example, an analysis of web search engine logs reveals that the average query length for Web search is about 2.3 words [1,2]. To alleviate such a short query problem, high-performance Web search engines often incorporate interactive search and/or term suggestion techniques [3]. These search engines attempt to identify some of the users' intentions and suggest more precise search terms, in particular for high-frequency queries. Several researches on mining search engine logs are believed helpful in dealing with the problem [4].

In this paper we are going to present a new effective log-based term suggestion approach for interactive Web search. The most important feature of the proposed approach is the development of a query-session-based term suggestion method that exploits the contextual information in the query session. Using the method the suggested terms in each interactive step can be obtained by their relevance with the whole query session, rather than as conventional approaches with the most recent single query.

The significance of the contextual information is well illustrated in the example in Fig. 1. The query terms in these three query sessions were submitted by three different users with different information needs but all three sessions contain the term "Taiwan University Hospital". By looking at the entire sessions, one can easily figure out that the first user was looking for a hospital with a high-quality department of Obstetrics and Gynecology, and the second user wanted to find some medical journals, and the third user wanted to find a hospital in the southern part of Taipei. However, no one can figure out the real needs of the users if only given the term "Taiwan University Hospital".

Session 1	Obstetrics and Gynecology Department Women and Children Hospital Taiwan University Hospital
Session 2	Taiwan University Hospital Medical College of National Taiwan University National Taiwan University Medical Library Journal Medial Journal
Session 3	Cathay General Hospital Taipei Municipal WanFang Hospital Taiwan University Hospital Tri-Service General Hospital

Fig 1. Examples of contextual information in query sessions.

2. THE PROPOSED APPROACH

Fig. 2 depicts the basic operations of the proposed term suggestion mechanism. The kernel term suggestion module operates based on a term relevance analysis conducted in advance on a collected log of users queries from proxy servers. To reduce replicated web transmission, proxy servers cache most of local users' network transmission, especially for the retrieved Web pages. Through proxy logs, it is possible to extract local users' search requests instantly. A feasible solution for segmenting query sessions from proxy server logs is, therefore, developed.

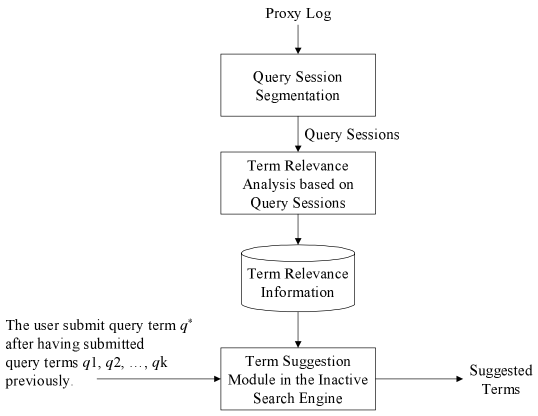
Proxy Log
Query Session Segmentation
Query Sessions
Term Relevance Analysis based on Query Sessions
Term Relevance Information
The user submit query term q^* after having submitted query terms $q1$, $q2$, ..., qk previously.
Term Suggestion Module in the Inactive Search Engine
Suggested Terms

Fig 2. The basic operations of the proposed term suggestion mechanism.

The collected proxy log is first partitioned into a number of query sessions and relevance among the query terms in the log is computed based on how these query terms are clustered in query sessions. The result from the term relevance analysis is then exploited by the term suggestion module on-the-fly. When a user submits a new query term $q*$ after having submitted a series of query terms $q1$, $q2,...,qk$ previously, the term suggestion module will suggest terms that are not only relevant to the currently submitted term *q** but also relevant to the previously submitted terms *q*1, *q*2,…,*qk*.

In order to complete the above term suggestion mechanism, three functional modules are designed.

1. Query Session Segmentation – this module uses a time threshold as delimiter to segment query sessions.
2. Relevant Term Extraction – this module uses a synthesize method for relevance estimation. Relevant terms with different co-occurrence will be categorized into 3 type – highly co-occurred, co-occurred, and rarely co-occurred. Terms of each type will be extracted with a different relevance estimation function.
3. Query-session-based Term Suggestion – in this module we apply the following procedure.

```
function term suggestion(qc, q1,…,qk, Q, C, R, S)
{
  Input:
    qc: the query term currently submitted by the user
    q1,…,qk: the query terms previously submitted by the user
    Q: the set of query terms in the log
    C: the co-occurrence matrix,
      Cij = the number of sessions containing both qi and qj.
    R: extracted relevant term set of qc.
  Output:
    S – suggested term set

    $S = \varnothing$
    For every $q_i$ in R {
        Calculate scalar cluster measure of $q_c$ and $q_i$ :
            $\frac{\sum_{\forall qj \in Q}(Cc,j \cdot Ci,j)}{\sqrt{\sum_{\forall qj \in Q} Cc,j^2} \cdot \sqrt{\sum_{\forall qj \in Q} Ci,j^2}}$
        If scalar cluster measure of qc and qi > threshold ,
        then $S = S \cup \{qi\}$
    }
    return S;
}
```

3. EXPERIMENT RESULTS

The utilized query session log contains 160,180 sessions for the relevant term extraction experiments. Based on the log and the most proper thresholds, it is found that it can successfully extract relevant terms for 3,330 of 5,366 unique search terms whose occurrences are large than 10 in the log. On average there are 9.28 extracted relevant terms for high-frequency terms, 4.82 for medium-frequency terms, and 2.77 for low-frequency terms. The achieved relevance between the search terms and extracted relevant terms is highly out of our expectation. Though the test log size is small, the extracted terms especially for those high-frequency search terms, which are most ambiguous and need sub-requests to further clarify, are most relevant and hard to be obtained by manual analysis.

To realize the proposed approach in dealing with real-world search services, two query logs from real-world search engines, i.e., Dreamer and GAIS in Taiwan, were collected as the basis for analysis. The Dreamer's log contains 228,566 unique search terms in a period of over 3 months in 1998, and the GAIS's contains 114,182 unique search terms in a period of 2 weeks in 1999. It is noted that the top twenty thousands search terms in the Dreamer log occupy 81% of the search frequencies, and 9,709 of which still occur in the GAIS log, which are called core search terms in our analysis. Although the log tested is small, the proposed approach has been proven can successfully extract relevant terms for 2,856 of the core search terms, and on average there are 9.36 extracted relevant terms for high-frequency terms, 4.73 for medium-frequency terms, and 2.73 for low-frequency terms.

Another experiment was performed to realize the performance of query-session-based term suggestion. More than 10K query sessions in the session log were tested to see the reduction of irrelevant term suggestion. Some of the obtained results are illustrated in Table 1. It can be found the average number of final suggestions can be effectively reduced.

Number of Queries in Session	Number of Effective Sessions	Average Number of Largest Extracted Relevant Terms	Average Number of Relevant Terms for Last Query	Average Number of Final Term Suggestions
2	13,659	12.517	9.677	3.526
3	1,211	16.573	11.889	2.449
4	139	21.129	12.942	1.885
5	30	19.7	10.8	1.567

Table 1. Suggestion reduction in an experiment on query-based term suggestions.

4. FUTURE RESEARCH ISSUES

Though the experimental results look promising, further study is needed on the following three issues that concern implementation of the proposed term suggestion mechanism:

1. how to partition the query log into query sessions so that each query session really corresponds to the query terms submitted by a user in a single information need.
2. how to measure the relevance between each pair of query terms in the log.
3. how to exploit term relevance in making term suggestion on-the-fly.

More details of the research work can be found in [5].

5. REFERENCES

[1] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR FORUM*, 32(1), 1998.

[2] C. Silverstein, M. Henzinger, H. Marais, and M. Morics. Analysis of a very large AltaVista query log. Technical Report 1998-014, Digital Systems Research Center, 1998.

[3] N.J. Belkin. Helping people find what they don't know. *Communication of ACM* (CACM), Vo.43, No8, pg 58-61,Aug 2000.

[4] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. *Proceeding of International ACM SIGKDD Conference on Knowledge* (KDD-00), pages, 2000.

[5] C.K. Huang, L.F. Chien and Y.J. Oyang. Query-Session-Based Term Suggestion for Interactive Web Search, 2000. http://mars.csie.ntu.edu.tw/~ckhuang/pub/www2001full.pdf.