Query Expansion Using Web Access Log Files[1]

Yun Zhu and Le Gruenwald

The University of Oklahoma, 200 Falgar Street, Norman, OK 73019, USA
{zhujulie, ggruenwald}@ou.edu
http//www.cs.ou.edu/~database/faculty.htm

Abstract. Query Expansion has long been recognized as one of the effective methods in solving short queries and improving ranking accuracy in traditional IR research. Many variations of this method have been introduced throughout the past decades; however, few of them have incorporated web log information into the query expansion process. In this paper, we propose an expansion technique that expands document content at the initial index stage using queries extracted from the web log files. Our experimental results show that even with a minimal amount of real world log information available and a professionally cataloged knowledge structure to aid the search, there is still a significant improvement in using our query expansion method compared to the conventional query expansion ones.

1 Introduction

With the explosive growth of the World Wide Web, many web sites are providing web interfaces for users to access their databases. Thus, it is becoming increasingly important to find an optimum retrieval system suitable for such applications. Compared to the conventional full text retrieval systems, the web-interfaced retrieval systems face additional obstacles and opportunities that need to be addressed.

One of the well known problems posed to the web search systems is that web users tend to enter very short query terms, generally two to three keywords per query [11]. With the paucity of query terms, it is much more likely to have the word mismatch scenario where query terms do not match any keywords in a relevant document. Although these web-based search systems encounter the severe word mismatch problem, they also have an additional piece of information available that can be exploited for search purposes – web access log files recording user activities when they access the web site.

For decades, many studies in IR have tried to address the short query problem. One of the most recent and effective approaches is query expansion [5]. This technique involves expanding and reweighing query terms and reforming query results based on the expanded query. The source for term expansion is typically derived from user feedback, or documents assumed to be relevant to the original query or knowledge structure such as thesaurus. Although this technique has been

[1] This work was partially supported by the National Library of Medicine, Grant No. 1 G08 LM007877-01 and 1 G08 LM008054-01.

K.V. Andersen, J. Debenham, and R. Wagner (Eds.): DEXA 2005, LNCS 3588, pp. 686 – 695, 2005.

shown to be effective [5, 18], few studies have so far explored its application in web search and incorporation of web log information. The only experiment that utilizes log information was limited to expansion term selection [8]. However, there are more areas in the retrieval process where log information can be applied.

In this paper, we propose a new query association and expansion method that utilizes web log information to the full extent. By applying log mining techniques, we are able to expand the document contents with query terms entered by the users. We then perform expansion search based on the expanded document contents. Our query expansion method achieves good improvement compared to conventional expansion techniques even when there is professionally cataloged knowledge structure to aid the search and only a minimal amount of log information available.

2 Related Works

The concept of query expansion was originated in the 1970's [12] where user feedback is applied not only to the reweighing of terms, but also to the expansion of search terms for further retrieval improvements. The basic idea of this technique is that each time the system retrieves a set of documents based on a user's query, a set of extra terms is then selected from the relevant documents identified by the user, then finally a new query is formed with the selected terms and a new search is performed. This process can be carried out iteratively, and it is expected that the more iterations it goes through, the more number of relevant documents will be retrieved. Although this technique has been shown to be effective [13, 16], the requirement for users' constant feedback on relevance is not appealing.

Throughout the years, many variations of query expansion techniques have been introduced. They can be categorized into three main groups: manual query expansion [4], query expansion based on the complete document collection [5, 6, 7, 9] and query expansion based on local analysis [2, 17, 18] (also referred to as *Relevance Feedback* or *Pseudo Relevance Feedback*). The manual query expansion involves users' judgment on which terms to select for expansion. This technique is rarely implemented because studies have shown that it does not improve search results effectively [4]. The basic idea for query expansion based on the complete document collection is to study the correlation between terms in the documents and identify the relationships between terms throughout the collection. This technique usually involves the manual or automatic building of a thesaurus type knowledge structure to aid the search. Unfortunately, building such a knowledge structure is extremely expensive, and has not achieved any significant improvements in experiments [5]. The third group, query expansion based on local analysis, assumes that the top n documents retrieved based on users' original queries are always relevant, thus the expansion terms are selected via studying the correlations between the terms within those n documents. This technique dismisses users from any form of input, and has shown improvements in many studies, especially in the experiments from TREC [5, 18].

Recently, there have been two interesting studies on applying alternative pseudo relevance feedback techniques to web search and both have achieved significant improvement. The first study by Billerbeck and others [3] utilizes a query association