CITY UNIVERSITY OF HONG KONG
香港城市大學

Effectiveness of Phrase in Information Retrieval
片語在信息檢索的成效影響

Submitted to
Department of Computer Science
電腦科學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Chang Chor Ming Matthew
鄭礎明

August 2008
二零零八年八月

Abstract

With the advent of Internet and advancement in computer technology, there are two notable challenges in document retrieval: retrieving huge amount of data *efficiently* and identifying the most useful documents, out of many relevant pages, *correctly*.

Although various phrase-finding and indexing methods have been proposed in the past, conclusions on the usefulness of phrases as indexing units have been generally inconsistent. Nevertheless, a number of recent research groups, including the leading groups who have participated in TREC campaigns, have used phrases as indexing units and have been able to obtain some improvement. As phrases have traditionally been regarded as precision-enhancing tools, recent research continues to apply the concept of phrase in different IR problems. In this thesis, following the tradition, we are interested in the concept of phrase in information retrieval, especially for document retrieval.

To address the two challenges, we first propose a *common phrase index* as an efficient index structure to support phrase queries in a very large text database. Our structure is an extension of previous index structures for phrases and achieves better query efficiency with modest extra storage cost. Further

improvement in efficiency can be attained by implementing our index, according to our observation of the *dynamic nature* of common word sets. In experimental evaluation, a common phrase index using 255 common words yields an improvement of about 11% and 62%, in query time for all queries and large queries (queries of long phrases) respectively, over an *auxiliary nextword index*. Moreover, it needs only about 19% extra storage space. Compared with an *inverted index*, our improvement works out to about 72% and 87% for all and large queries respectively. We also propose to implement a common phrase index with dynamic update feature. Our experiments show that more improvement in time efficiency can be achieved.

For improving the quality of retrieval results, we devise a proximity-based ranking function that combines an "*ordered loose phrase*" scoring with the state-of-the-art Okapi probabilistic model (BM25). We say that a phrase occurs in a document in an *ordered loose phrase* form, when the words of the phrase appear sufficiently close to each other, and in the same order as in the query. The occurrence of an ordered loose phrase, constituted by words of a query phrase in a document, may indicate a high relevance of the document to the query. We design our experiments using the query sets in TREC-11, TREC-12 and TREC-13 and the .GOV document collection. The results show that our method compares favorably with the pure BM25 and three recent works based on term proximity and co-occurrence, in most of the performance

measures. For TREC-12 and TREC-13, our results successfully demonstrate that our method can improve the quality of search results significantly.

Contents

List of Figures

List of Tables

List of Algorithms