Abstract

With the successful adoption of link analysis techniques such as PageRank and web spam filtering, current web search engines support navigational search well, where a user is looking for a particular web resource that the user has in mind. However, such engines do not necessarily support informational search well, where a user is looking for information about a certain topic that might be on diverse web resources. This is because a user often forms an informational query by a few keywords that does not necessarily model the user information need well while such engines search web documents basically based on conjunctive Boolean searching using the submitted keywords. Informational search would be better handled by a web search engine based on an information retrieval (IR) model combined with automatic query expansion. Moreover, the realization of such an engine requires a method to process the IR model efficiently. So in this thesis, we propose new top-k document retrieval algorithms that efficiently process long queries generated by automatic query expansion, by introducing a simple additional data structure called "query-term-by-document binary matrix," which indicates which document contains which query term. We show on the basis of theoretical analysis that our algorithms not only find the top-k documents exactly but also have a desirable property on processing cost as described below. Furthermore, we show on the basis of empirical evaluation using the TREC GOV2 collection that our algorithms achieve considerable performance gains over existing algorithms especially when the number of query terms gets larger, yielding speedup of up to a factor of about 2 over existing algorithms for top-100 document retrieval for 64-term queries. Then, we extend our algorithms for supporting proximity search to take advantage of the structured nature of web documents, and show that the extended versions of our algorithms are still exact for finding the top-k documents and desirable on processing cost. The proposed algorithms presented in this thesis are applicable not only to web search but also to other areas such as enterprise search. The novel contribution of this thesis is summarized as follows: (a) The proposal of new top-k document retrieval algorithms that efficiently process long queries generated by automatic query expansion, by introducing a simple additional data structure called query-term-by-document binary matrix. (b) The theoretical analysis on the proposed algorithms. We show that our algorithms not only find the top-k documents exactly but also have the desirable property on processing cost. (c) The empirical evaluation of the proposed algorithms. We demonstrate that our algorithms achieve considerable performance gains over existing algorithms. And (d) The extension of the above algorithms for supporting proximity search. The algorithms proposed in this thesis efficiently process an IR model combined with automatic query expansion and/or proximity search, by introducing a simple additional data structure called query-term-by-document binary matrix. Due to the simplicity of our method using query-term-by-document binary matrix, our method is also applicable to other IR techniques. We believe that our method paves the way for practical use of various IR techniques, including an IR model combined with automatic query expansion and/or proximity search, in large-scale text databases such as the web, which has been considered difficult because of their inefficiency.