氏名	藤田 悦郎			
学位（専攻分野）	博士（情報学）			
学位記番号	総研大甲第1602号			
学位授与の日付	平成25年3月22日			
学位授与の要件	複合科学研究科 情報学専攻 学位規則第6条第1項該当			
学位論文題目	Efficient Retrieval of Highly Ranked Documents for Informational Search on Large Scale Text Databases			
論文審査委員	主査	教授	大山 敬三	
		教授	高須 淳宏	
		准教授	定兼 邦彦	
		准教授	片山 紀生	
		教授	相澤 彰子	国立情報学研究所

論文内容の要旨

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 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 below:

a) The proposal of new top-k document retrieval algorithms that efficiently process long queries generated by automatic query expansion, by introducing 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.

d) The extension of the above algorithms for supporting proximity search.

This thesis consists of eight chapters which are summarized as below:

1) Introduction.
We identify the background of the thesis. Although current web search engines support navigational search well, such engines do not necessarily support informational search well. We discuss why such engines fail for informational search, mention that potential solutions to this problem is to adopt an IR model in combination with enhancement techniques such as automatic query expansion and proximity search, and identify issues from the efficiency point of view, which we should address in this thesis.

2) Related Work.
First, we give an overview of basic techniques that improve retrieval effectiveness of informational search, including automatic query expansion and proximity search, and discuss retrieval efficiency issues for implementing such techniques in large scale text databases such as the web. Secondly, after reviewing various approaches to the efficiency issues briefly, we investigate previous work on top-k document retrieval in detail that is the most promising solution to the issues. Lastly, we give the reason why we focus on top-k document retrieval in this thesis, and then describe the differences of our approach from existing approaches in the literature.

3) Existing Top-k Document Retrieval Algorithm.
In this chapter, we describe the classical top-k document retrieval algorithms called the threshold algorithms provided by Fagin et al., which have been shown to find the top-k documents exactly and have a desirable property on processing cost that is described in Chapter 5. The proposed algorithms in this thesis are based upon Fagin's algorithms as stated below.

4) Proposed Algorithm introducing Query-Term-by-Document Binary Matrix.
In this chapter, we present novel top-k document retrieval algorithms for long queries generated by automatic query expansion. We extend Fagin's algorithms to the case where simple additional data structure called query-term-by-document binary matrix, which indicates which document contains which query term, is available. We first describe the key points on why the extended algorithms are much more efficient than Fagin's original algorithms, and then present the extended algorithms in detail.

5) Theoretical Study.
In this chapter, we discuss theoretical analysis on the extended algorithms presented

in Chapter 4. It is required for a top-k document retrieval algorithm not only to find the top-k documents exactly but also to have a desirable property on processing cost called instance optimality, which guarantees that the processing cost of a top-k document retrieval algorithm for any query is always linearly bounded with the cost of the minimum-processing-cost algorithm for that query. We show that the extended algorithms have exactness and instance optimality.

6) Empirical Study.
In this chapter, we provide empirical evaluation of the extended algorithms presented in Chapter 4. We show using the TREC GOV2 collection and expanded versions of the evaluation queries attached to this data set that the extended 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.

7) Extension to Proximity Search.
In this chapter, we extend the above algorithms for supporting proximity search along an existing framework for incorporating proximity search into top-k document retrieval. While existing proximity-aware top-k document retrieval algorithms have not been analyzed theoretically, we perform theoretical analysis and show that the extended versions of the above algorithms still have exactness and instance optimality.

8) Conclusion.
We give our conclusion, and state some future work.

The algorithms proposed in this thesis efficiently process an IR model combined with automatic query expansion and/or proximity search, by introducing 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 methodology 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 that has been considered difficult because of their inefficiency.

博士論文の審査結果の要旨

本論文は，Web 検索エンジンなどの文書検索システムにおいて，ランキングの上位文書を正確かつ高速に取得するためのアルゴリズムに関するものである。特に，クエリ拡張などによって生成されたロングクエリに対してベクトル空間モデルなどの情報検索モデルを高速に処理可能とするための手法をテーマとしている。

本論文は８章と付録から構成されている。第１章はイントロダクションであり，本研究の背景，貢献，及び論文の構成が述べられている。第２章には関連研究について，内容の概略と本研究との関係が示されている。第３章は本研究と直接関連する既存の Top-k 文書検索手法について，まず基本的な概念の定義や説明を行い，次にそれらの概念に基づいて２種類の具体的なアルゴリズムを紹介している。第４章では提案手法について述べている。本研究の核となるデータ構造である Query-term-by-document binary matrix（以下 BM）を示した後，第３章で紹介した２種類の各アルゴリズムに BM を導入して Top-k 文書検索を高速化したアルゴリズムを提案し，これらの定性的な挙動と優位性について分析している。第５章では提案アルゴリズムの理論的な分析を行っている。まず Top-k 文書検索アルゴリズムを特徴付ける二つの性質，即ち正確性と Instance optimality の定義を述べ，次に第４章で提案した２種類のアルゴリズムのそれぞれについて，これらの性質を有することの証明を与えている。第６章は実験による評価である。まず提案アルゴリズムの実装方法について述べ，次に大規模文書検索のベンチマークとして広く用いられている文書データセットとクエリセットを用いた実験設定について述べた後，提案アルゴリズムに対する実験結果を示している。また，一般的な検索アルゴリズム及び第３章で紹介した既存アルゴリズムによる同一条件での実験結果との比較によって提案アルゴリズムの優位性が示されている。第７章には近接検索手法を導入した情報検索モデルにおいて，BM による高速化を行ったアルゴリズムが示されている。２種類の既存アルゴリズムを紹介した後，これらに基づく提案アルゴリズムを示し，それらの理論的な分析を行っている。第８章は結論であり，論文のまとめと将来の課題が述べられている。最後に，クエリ拡張によって生成されたロングクエリを用いることによる検索の質の向上について，予備実験の結果が付録に示されている。

以上の結果は，提案手法が，処理負荷の高い情報検索モデルにおいてロングクエリを高速に処理するための汎用性の高い手法であることを示している。既存の Web 検索エンジンでは，処理性能上の制約により，情報収集目的を中心とした情報要求に対して十分に応えることができないという問題があるが，本手法はこの問題の解決に道筋を付けるものであり，学術的にも実用的にも高い価値を有する。

本成果は，電子情報通信学会英文論文誌に原著論文として１編の採録が決定しており，また査読付き国際会議においても採択されている。

以上により本審査委員会は，本論文が博士論文に値するものであると判断した。