審査の結果の要旨

論文提出者氏名　小川 泰嗣

本論文は「日本語文書を対象としたn-gram 索引による高速検索手法の研究」と題し、ＩＴ社会の進展に伴う文書検索に対する高速化要求の高まりを受け、日本語文書に対する代表的な索引の一つであるn-gram（n文字組）を索引要素とするn-gram索引の検索高速化に関し纏めたものであり、検索においてコスト増大の主要因である位置検査処理を可能な限り削減することにより処理の効率化を図る高速検索手法および当該手法に適したn-gram索引の物理編成法を提案すると共に、評価実験によりその有効性を検証しており、９章から構成される。

第１章は「序論」であり、本研究の背景および目的について概観し、本論文の構成を述べている。

第２章は「n-gram 索引の基本概念」と題し、３章以降の議論の前提となる n-gram 索引の基本概念を説明している。まず、登録処理、検索条件を満足する文書を特定するブーリアン検索処理、検索文書を検索条件に対する適切さに応じて順序付けるランキング検索処理などn-gram 索引の基本動作について述べ、つぎにn-gram索引のファイル形式である転置ファイルの基本構成と圧縮方法を紹介している。

第３章は「ブーリアン検索処理の高速化」と題し、検索文字列中のn-gram群が文書中で文字列を構成することを確認する位置検査を削減・省略することにより検索を高速化する手法を提案している。

第４章は「ブーリアン検索高速化手法の評価」と題し、新聞記事８年分、約85万件、約750MBの検索対象文書と、280個の検索条件を用いて、ブーリアン検索の高速化手法を評価している。単一検索文字列における速度向上は二次記憶装置へのアクセスを伴うCOLDでは8.3%、主記憶上の処理のみであるHOTでは73%であり、論理演算子における速度向上はCOLDでは10.7%、HOTでは110%に達し、提案手法の有効性が確認されている。

第５章は「ランキング検索処理の高速化」と題し、ランキング検索では検索条件に対する文書の適切さを表す文書スコア計算に文書頻度、文書内頻度の２種類の頻度情報が必要となり、両頻度を求めるごとに位置検査が発生するため検索コストが大きいことを指摘した後、位置検査を削減する２つの高速化手法を提案している。１つ目の順序入れ替え法は、検索文字列が出現する文書を特定した結果から文書内頻度を算出することにより、従来必要であった文書頻度のみを求める処理を省略し、検索を高速化するものであり、他の１つの頻度推定法は、検索文字列を構成するn-gramの頻度情報から文書頻度および文書内頻度を近似的に求めることにより位置検査を省略し、検索を高速化するものである。さらに、両手法の組み合わせについても考察している。

第６章は「ランキング検索高速化手法の評価」と題し、ランキング検索用の標準的な評価用データセットであるNTCIR-1予備版（論文要旨約33万件、約270MBの検索対象と、30件の自然言語で記述された検索要求から構成される）を用いて、提案した高速化手法を評価している。その結果、順序入れ替え法と頻度推定法の組み合わせにより、36.2%から91.7%の速度向上が得られる一方で検索精度には統計的に有意な差が生じないことが確認され、提案手法の有効性が検証されている。

第７章は「高速検索手法向きの転置ファイルの物理編成法」と題し、n-gram 索引を構成する転置ファイルの物理編成法を高速検索手法の視点から検討し、文書IDと文書内頻度・出現位置の分離、固定長ブロック単位の圧縮などを組み合わせた物理編成法を提案している。単純な物理編成法と比較して、最大で 9.6 倍の高速化が達成されることが確認されている。

第８章は「実システムにおける評価」と題し、提案手法を組み込んだ FTS という文書検索システムについて説明している。さらに特許システムの事例データを用いて FTS と他の検索システムの性能比較を行い、FTSの優位性が確認されている。

第９章は「結論」であり、本研究の成果と今後の課題について総括されている。

以上、これを要するに本論文は、n-gram 索引による文書検索に関して、位置検査の低減を図ったブーリアン検索およびランキング検索に対する新しい高速検索手法を提案するとともに、当該手法に適したn-gram索引の物理編成法を提案し、評価実験によりその有効性を示したものであり、電子情報工学上貢献するところが少なくない。

よって、本論文は、博士(工学)の学位請求論文として合格と認められる。