論 文 の 内 容 の 要 旨

論文題目　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
氏　　名　小川 泰嗣

膨大な情報の中から必要な情報を的確に見つけ出すことが情報検索であり、情報検索は情報化社会における基本的な技能の一つに位置づけられる。文書検索はテキスト文書を検索対象とする情報検索である。本研究がテキストを対象としたのは、テキストが知識を表現する上で最も主要なメディアという地位にあるからである。電子化文書の増大・検索ユーザの拡大・文書共有化の進展という最近のＩＴ革命の進展に伴い、文書検索に対するニーズも非常に高まっている。

文書検索の対象は自然言語で記述されたテキストである。文法・語彙等の特性は言語によって異なるため、言語ごとに相応しい文書検索モデル・実装方法も異なる。特に、検索高速化のために必要不可欠な索引において、索引の基本要素である索引単位として何を選択するかは性能および運用しやすさに大きく影響する。代表的な索引方式には単語を索引単位とする単語索引と n-gram（n 文字組）を索引単位とする n-gram 索引がある。単語を分かち書きせず、造語力の高い日本語においては n-gram 索引が広く利用されているが、n-gram 索引には検索速度が遅いという問題がある。その理由は、ユーザから与えられる検索文字列が索引単位である n-gram に一致するとは限らず、一致しない場合には複数の n-gram を用いて検索結果を求める必要があるからである。特に検索文字列が n-gram よりも長い場合には、検索文字列中の n-gram が全て出現する文書を特定（候補文書特定処理と呼ぶ）した上で、そのような文書において n-gram が適切な相対位置にあって検索文字列を構成していることの確認（位置検査処理と呼ぶ）が必要であり、高速化の余地が大きい。

本論文では、日本語文書を対象とした n-gram 索引の検索処理の高速化について研究する。高速化の基本原理は、n-gram 索引における検索コスト増大の主要因である位置検査処理を可能な限り省略・低減するというものである。n-gram 索引における検索高速化の従来研究は、複数の n を組み合わせる、文字種に応じて n を調整する等の n-gram 抽出法に集中していた。これに対し、本研究は検索処理法および索引の物理編成法の改良による高速化である点に特徴がある。

検索処理法の改良では位置検査の削減・省略に着目した。日本語では造語力の高さから長い複合語がいくらでも生成されるため、n-gram 抽出法の工夫だけでは n-gram 索引における検索速度低下要因の位置検査を無くすことが不可能だからである。位置検査省略による検索高速化という考え方は従来研究にも存在している。しかし、従来研究が単一検索文字列の場合のみを極めて限定的に扱っていたのに対し、本研究では以下の３点について拡張する。

①単一検索文字列処理において、省略する n-gram を動的に選択するように拡張

従来は候補文書特定と位置検査の両フェーズに同一の n-gram 群を静的に選択していた。これに対し、本研究では両フェーズの処理特性の違いに着目して異なる n-gram 群を索引内容に応じて動的に選択する。実際には、位置検査用には検索文字列を被覆する n-gram 群のなかから n-gram の文書頻度の合計が最小となる組み合わせ（最小頻度パスと呼ぶ）を動的に選択し、候補文書特定には最小頻度パスにそれらの前後にある文書頻度の少ないものを加えた n-gram 群を用いる選択 n-gram 法を提案した。選択 n-gram 法により、位置検査を行うべき文書数を削減するとともに位置検査自体のコストも低減し、単一検索文字列の検索処理を高速化できる。

②AND・OR・ANDNOT の論理演算子に拡張

従来は単一検索文字列の処理のみを高速化の対象にしており、論理演算子処理を対象とする研究はなかった。これに対し本研究では、論理演算子の処理アルゴリズムを複数の検索文字列の論理関係を考慮することで不要な位置検査を行なわないように改良し、検索を高速化する。AND, OR, ANDNOT の３種類の論理演算子について、この考えに基づく拡張省略法を提示した。さらに、演算子が入れ子になった複合条件に対しても拡張省略法が適用できることを示した。

③ランキング検索に拡張

ランキング検索では検索条件に対する文書スコアに応じて検索された文書を順序付けする。文書スコア計算には文書頻度・文書内頻度の２種類の頻度情報が必要であるが、両頻度を求めるたびごとに位置検査が発生するため検索コストが増大する。従来研究として n-gram の頻度情報に基づいてスコア計算するスコア合成法が提案されているが、検索精度の低下という問題があった。これに対し、検索精度を低下させることがないように検索文字列の頻度情報に基づいてスコア計算しつつ、頻度情報を求める際に必要な位置検査を可能な限り削減する２つの高速化手法を提案した。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すると同時に文書内頻度も求めることで、従来必要であった文書頻度を単独で求める処理を省略する。もう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で位置検査を省略する。これら２つの高速化手法は組み合わせ可能であり、相乗効果が期待できる。なお、推定頻度は必ずしも正しい値にならないため、頻度推定法によるランキング結果は基本方式のものと同一とは限らないが、検索精度への影響は極めて小さいと考えられる。　□

n-gram 索引のファイル形式としては、n-gram の各文書における出現位置を格納可能である転置ファイルが一般的である。本研究では、位置検査省略による検索高速化手法向けの転置ファイルの物理編成法も提案する。単語索引用転置ファイルの物理編成法は従来から研究されており、その一部は n-gram 索引に対しても有効である。本研究では、n-gram 索引では検索処理が候補文書特定と位置検査の２つのフェーズから構成され、文書内出現位置は位置検査でしか用いられないという処理特性に着目した物理編成法を提案する。具体的には、n-gram の出現情報を、文書ＩＤとそれ以外の文書内頻度・出現位置情報に分離

して配置するとともに、出現情報を圧縮する際にブロック化する。以上の工夫により、ページアクセスおよび検索時の伸長処理を大幅に削減することが可能となる。

提案高速化手法の評価実験を行った。単一文字列と論理演算子については新聞記事８年分、８５万文書（テキスト量７４８ＭＢ）を用いて評価した。ランキング検索については論文要旨３３万文書（テキスト量２６７ＭＢ）から成るテストコレクションＮＴＣＩＲ－１を用いて評価した。以下の表は評価結果のまとめである。

	COLD	WARM	HOT
単一文字列	8.3%	20.8%	73.0%
論理演算子	***	***	***
AND	21.1%	40.1%	252.0%
OR	4.0%	11.0%	41.3%
ANDNOT	5.7%	14.5%	35.1%
ランキング	36.2%	40.7%	91.7%

この表は、全データの読み込みを伴う COLD、条件評価フェーズにおけるデータ読み込みのみの WARM、データ読み込みを伴わない HOT の３つ場合について、従来手法に対する提案手法の検索速度の向上率を示している。ランキング検索については、高速化効果が最大であった順序入れ替え法と頻度推定法を組み合わせの結果を用いている。この結果から提案手法により検索が大幅に高速化され、特に AND 演算子、ランキング検索に対する効果が大きいことが確認できる。また、COLD, WARM, HOT となるにしたがって高速化効果が大きく、HOT では 73～250%に達していることから、提案手法はディスクアクセスよりも CPU 処理の削減に効果が大きいことがわかる。

つぎにランキング検索における単語索引・n-gram 索引の性能比較結果を示す。

	単語索引	n-gram索引	改良n-gram
索引サイズ	197MB	490MB	←
登録時間	15246sec	13668sec	←
検索時間	***	***	***
COLD	0.698sec	1.291sec	0.948sec
WARM	0.410sec	1.023sec	0.696sec
HOT	0.319sec	0.663sec	0.346sec
検索精度	0.3699	0.3827	0.3791

この表で、n-gram 索引は従来方式、改良 n-gram 索引は提案方式の結果である。改良 n-gram 索引では検索精度をほとんど低下させることなく検索を高速化した結果、単語索引に対する検索時間の差異が小さくなることが確認できた。特に HOT においては検索時間の差は約 10%と小さい。本実験は bi-gram 索引を使用するという n-gram 索引にとって不利な状況での測定結果ということを考慮すると、検索が遅いという n-gram 索引の問題点は提案手法によりかなり克服できたと言える。

また、物理編成法の改良に関する性能評価も行った。本研究で提案した出現位置情報の分離とブロック化自体の組み合わせにより、基本的な転置ファイルの物理編成法に対して81%(COLD), 128%(WARM), 860%(HOT)の高速化を達成できることを確認した。

本研究の成果は全文検索エンジンである FTS サーバとして製品化され、当社の図書管理システムやオフィス向け文書管理システムなどで広く採用されている。FTS 最初のバージョンの開発時に行った性能比較によれば、当時市場で最高速であるという評判の商用文書検索システムに対して 2.5 倍（HOT）から 4.0 倍（COLD）高速であり、提案高速化手法の有効性が製品レベルでも確認できた。

今後の課題としては、日本語以外の言語に対する提案手法の有効性の検証、更新性能を考慮した複数索引への対応、質問拡張の高速化検討などがあげられる。日本語以外の言語の対応では中国・韓国語等の東アジアの言語が主な対象となるが、言語特性から提案手法が有効であることが期待できる。複数索引への対応についてはランキング検索の高速化が問題となるが、順序入れ替え法を拡張することで対応可能と考えられる。質問拡張については、選択する関連語の選択時の頻度取得が速度上の問題となるが、頻度推定法の適用により対応可能と考えられる。