称号及び氏名	博士（工学）	山本　裕
学位授与の日付	平成 17 年 9 月 30 日	
論　　文　　名	「全文検索システムのインデクスと遠隔データアクセスに関する研究」	
論文審査委員	主査	辻　　洋
	副査	市橋　秀友
	副査	石渕　久生

論文要旨

企業・公共機関・大学などで文書情報や知識情報を扱い計算機に蓄積する情報管理システム環境や，全世界に計算機ネットワークでつながっているインターネット環境には，各業種の実務に必要なデータベース情報，知識情報，マルチメディア系の各種大容量情報など，大量の情報が蓄積されている．多様な業種での社内文書情報の検索，自治体での公文書や議事録の検索，大学・図書館・博物館などでの学術文書・資料・文献情報の検索など，大量に蓄積された文書情報や統計情報の中から，有効な所望の情報を効率良く利用することは，情報システム構築の要件の１つである．

本研究の目的は，前述のようなシステム構築のための要素技術である「大量文書全文検索技術」と「大容量データ転送技術」に関して，効率的な使用方法を提案および検証することである．

最初に，これらの要素技術の情報管理システムにおける位置付けを以下に示す．大量文書全文検索技術は，Web でのキーワード検索や組織内文書情報の検索，組織内ポータル環境での文書参照・情報分類・類似文書検索など，情報管理に幅広く利用される．また，大容量データ転送技術は，情報管理システムの限られた通信資源を使用して，多種多様な形式（文書・マルチメディア情報）の大容量データを複数の計算機の間で移動するために利用される．

次に，これらの要素技術に対するシステム要件を示す．大量文書全文検索技術に対する要

件は，蓄積された文書情報から有効な所望の情報を取得するプロセスの効率を改善することである．そのためには特に，インデクスの有効な活用方法を確立することが必要条件である．また，データ転送技術に対する要件は，転送速度に加えて通信資源の占有の抑止など，システム全体の業務スループットを改善させることにある．特に，システム環境条件に応じた大容量データの分割転送方式を確立することはスループット改善の必要条件である．これらの要件を満足した要素技術を有効に利用することが，情報管理システム全体のパフォーマンスを向上するための重要なキーポイントとなる．

本研究では，全文検索インデクスの最適化に着目した．具体的には，N-gram インデクス型方式におけるインデクスの作成と使用方法に関して改善提案し，実証評価を行った．

この方式を採用する際に最も一般的な 2-gram インデクス型システムに関して，検索効率を最適化する手法として，追加型の高次（N-gram(N＞2)）インデクスとの併用方式がある．既存の研究では，特定の検索語に関して，1～2gram 及び 1～3gram のインデクスを使用した検索の場合と N-gram 追加型高次インデクスを使用した場合の検索性能とインデクス容量に関する比較評価が実施されている．しかし，これは語長が長い特定の検索語に限って性能改善を実現するものであり，実際に導入が想定される全文検索システムの検索条件と，高次インデクス数の条件を加味した検索性能の傾向分析が望まれている．

一方，追加型高次インデクスの作成基準に関しては，マシン性能とディスク性能を考慮して目標検索性能をクリアするための基準インデクス容量を算出し，このインデクス容量を閾値とする考え方が存在する．しかし，本方式は，実際にインデクスを作成した後に容量を分析して追加インデクス対象語を決定するものであり，インデクス設計のプロセスに課題があった．つまり，インデクスを作成する前に，どの検索語に関して追加インデクスを作成すれば効果があるかを判断する指標については言及されていない．

本研究では，追加型高次インデクスの対象となる検索語の選択基準の提案を行った．また，提案方法に関する実証評価計画を立案し，実際のシステムを想定した検索条件とインデクス条件での実証評価を行った．

また，大容量データ転送技術の利用に関して，システム資源の有効活用と効率の良い送受信プロセスという２つの観点で，大容量データを遠隔アクセスする場合の分割送受信の方法を提案した．具体的には，遠隔データベースアクセス手法として基盤となる OSI-RDA（Open Systems Interconnection - Remote Database Access）方式を分析し，大容量データをアクセスする手法の１つとして，サーバ主導型データ分割送受信方法を提案・評価した．

第２章以降の各章の内容は以下の通りである．

第２章では，N-gram 型全文検索システムにおける「追加型高次インデクス併用方式」を利用する場合の課題を整理し，高次インデクス作成対象の検索語の選択に関するアプローチを提

起した．まず，高次インデクス数が多い場合には，インデクス容量が増大し全文検索処理全体の性能が劣化してしまう．逆に少ない場合には，高次インデクスの効果が出なくなる．検索性能に関して以上の課題があるため，語の選択が必要であることを示した．

提起したアプローチの１つは，過去に検索された語（検索語＝サーチターム）に着目するサーチタームインテンシブアプローチ，もう１つは，蓄積された文書中に含まれる語に着目するデータベースインテンシブアプローチである．さらに，各アプローチに則った，高次インデクスの対象語を選択する具体的方法に関する実証評価計画を示した．

第３章では，サーチタームインテンシブアプローチに則った追加型高次インデクス対象語の選択方法を提案し，実証評価を行った．具体的には，検索語として出現頻度が高い上位 10～100 語を高次インデクス対象として選択する方式を提案し，検索性能を評価した．評価対象は，新聞記事 20 万件とし，追加で作成する高次インデクスは 3 および 4-gram とした．

評価は，高頻度語の出現比率を変えた検索パターンと高次インデクスの作成数を変えたインデクスパターンの組み合わせを条件とし，検索時の高次インデクス利用率とインデクス容量，検索性能の関係を分析した．

実験したところ，高次インデクスを使用した検索では，インデクス容量が要因と想定されるメモリ swap 現象の頻発により性能が劣化した．つまり，20 万件程度の環境では 2-gram 方式の適用が妥当である結果となった．しかし，swap を除いたデータ範囲では改善が認められたためインデクス容量を抑止できれば改善効果を得る確証を得た．

第４章では，第３章と同じくサーチタームインテンシブアプローチに則った追加型高次インデクス対象語の選択方法を調査した．具体的には，検索語としての出現頻度が高い上位 100～1000 語を高次インデクスとして実証評価を行った．本章の評価対象は，新聞記事 10 万件とし，追加で作成する高次インデクスは，5-gram 以上とした．

また，評価は第３章と同じく，高頻度語の出現パターンを変えた検索パターンと高次インデクス作成数を変えたインデクスパターンの組み合わせを条件とし，検索性能の傾向を分析した．

数値実験で，2-gram インデクス使用時と比較して，高次インデクスを使った性能改善効果が確認できた．この要因は，語長が長い頻出語をそのまま 1 語の高次インデクスとしたこと，置換した 3 文字が文書 DB 中唯一の文字列であったことから高次インデクス容量を抑制できたことであると考えられる．従って，高次インデクス対象語を，検索に利用される頻出語に絞り込んでインデクス容量を抑制できれば，性能改善できる見込みを得た．

第５章では，データベースインテンシブアプローチに則り，文書データベース中の高頻度語

を対象として追加型高次インデクスを作成する方式を提案し評価した．具体的には，高次インデクス作成対象の検索語を，文書データベースに多く含まれる語の中から 2-gram の後の文字列出現確率に関するエントロピー値を用いて選択する方式を提案した．

また，本提案方式で高次インデクスを作成し，新聞記事 10 万件を対象として実験を行い，検索性能の改善効果を確認した．この結果，本提案方式が有効である確証を得たので，システムに最適な高次インデクス選択のガイドラインの設定を目的として，サーチタームインテンシブアプローチ（３章の検証）での評価との比較も含めた将来の可能性についても言及した．

第６章は，遠隔データベースアクセス方式として実績がある Remote Database Access 機能を分析し，「大容量データ転送方法」として， データベースにアクセスするためのサーバ主導型データ分割送受信方式を提案・評価した．

従来の方式は１要求－1 応答型方式であったのに対し，提案方式は１要求－N 要求型（サーバ主導型）方式である．評価指標は，従来方式と提案方式の転送効率の比較，および提案方式を適用した場合の分割数と転送処理時間との関係の分析とした．

ある商用 DB の分散機能を対象として性能値を算出したところ，従来の１要求－1 応答型の分割送受信方法と比較し転送効率を２倍程度改善することができた．また，通信資源占有率に関しては，転送データが大きくなるほど改善効果が顕著化することが確認できたので大容量データの分割方式として有効である確証を得た．さらに，全体の送受信時間が最小になる分割数 N を導出するガイドラインを示すことができた．

第７章は，結論とし，本研究によって得られた結果を総括し，今後の課題について述べた．

本論文の基礎となる発表論文

No.	論文題目	著者名	発表誌名	本論文との対応
1	Incremental Indexing and Its Evaluation for Full Text Search	H.Yamamoto S.Ohmi H.Tsuji	Proceedings of the 2003 Information Resources Management Association International Conference, pp.688-690 (Philadelphia,USA,2003).	第２章
2	Experimental Simulation on Incremental Three-gram Index for Two-gram Full-Text Search System	H.Yamamoto S.Ohmi H.Tsuji	Proceedings of IEEE International Conference on Systems, Man & Cybernetics 2003, pp.4846-4851 (Washington DC,USA, 2003).	第３章
3	N-gram型全文検索システムにおけるインデクス長の実験的検討	山本　裕 辻　　洋	電気学会 C 部門論文誌（投稿中）	第４章
4	Entropy-based Indexing Term Selection for N-gram Text Search System	H.Yamamoto S.Ohmi H.Tsuji	Proceedings of IEEE International Conference on Systems, Man & Cybernetics 2003, pp.4852-4857 (Washington DC,USA, 2003).	第５章
5	N-gram全文検索におけるエントロピーを適用した高次追加インデクス検索語選択方式	山本　裕 佐賀　亮介 森山　悟 辻　　洋	電気学会 C 部門論文誌 Vol.125, No. 5, pp.730-731 (2005).	第５章
6	遠隔データベースアクセス向け大容量データの分割送受信方式	山本　裕 辻　　洋	電気学会 C 部門論文誌 Vol.124, No. 5, pp.1076-1082 (2004).	第６章

審査結果の要旨

本論文は、全文検索システムにおいて必要となる、大量文書から効率よく検索を行うための索引作成方法と大容量データ転送方法を論じたもので、新しい方法を提案すると共に数値実験を通してそれらの有効性を検証しており、次のような成果を得ている。

(1) N-gram 型全文検索システムにおける「追加型高次インデクス併用方式」 を利用する場合の課題を整理し、2 種類のアプローチ（サーチ・ターム・インテンシブ・アプローチとデータベース・インテンシブ・アプローチ）があることを示した。

(2) サーチ・ターム・インテンシブ・アプローチとして，3 および 4-gram の高次インデクスを追加した場合の評価実験方法を示し、新聞記事 20 万件を対象とした評価を行った。その結果、swap 現象の多発による副作用を発見し、同規模のデータベースでは 2-gram のインデクスの適用が妥当との結果を得た。

(3) 同アプローチの副作用を避けるため、検索出現頻度の高い上位の語をそのまま 5-gram 以上の高次インデクスとする方法をとることを提案した。これによりインデクス容量を抑制することが可能となり、結果として、最大 70%の性能改善を実現した。

(4) データベース・インテンシブ・アプローチとして、単語のスペルを前方より走査したときのエントロピー値を算出することにより、高次インデクスとして追加する語を選択する方式を示した。10 万件の新聞記事を対象とした実験で最大 18.5%の性能改善を実現した。

(5) 遠隔地からデータベースをアクセスするときに生じる性能劣化の問題をとりあげ、1 要求 -N 応答型の方式を考案し、転送効率の比較、データ分割の数と転送処理時間の関係を示した。さらに全体の送受信時間が最小となる分割数を算出するガイドラインを示した。

以上の研究成果は、経営工学分野における経営情報システムの構築方法論の発展に貢献するところ大である。また、申請者が自立して研究活動を行うに必要な能力と学識を有することを証したものである。