【18】

氏　名（本籍）	高木（たかき）　徹（とおる）（新潟県）
学位の種類	博　士（情報学）
学位記番号	博甲第 3850 号
学位授与年月日	平成 17 年 7 月 25 日
学位授与の要件	学位規則第 4 条第 1 項該当
審査研究科	図書館情報メディア研究科
学位論文題目	**主題解析に基づく情報検索システムの研究**

主　査	筑波大学教授	石川　徹也
副　査	筑波大学教授	田中　和世
副　査	筑波大学教授	杉本　重雄
副　査	筑波大学教授	石塚　英弘
副　査	国立情報学研究所教授	大山　敬三

論文の内容の要旨

本論文は，「主題解析に基づく情報検索システムの研究」と題し，それぞれの目的の下に 3 つの異なる情報検索システムの高精度化手法の提案と，情報検索テストコレクションによる評価実験による手法の有効性を論じたものであり，7 章から構成されている。

第 1 章の「序論」では，本研究の背景，研究の目的と範囲と論文の構成について示している。最近の電子化情報の増加により必要な情報を効果的に検索することの必要性や，利用者の様々なニーズに対応するために多様な情報検索システムが提案されている背景について論じ，高精度検索手法の実現の意義について述べている。また，検索質問や，検索結果の文書に内在している複数の主題を利用する検索手法の必要性について述べ，主題情報を考慮した検索手法による情報検索システムの高精度検索の実現の可能性を論じている。

第 2 章では「関連研究」として，本研究の前提となる情報検索モデルについて解説し，提案手法である文書の主題解析に基づく情報検索の関連研究について述べ，従来手法の問題点を示している。また，情報検索システムの有効性評価に関する指標および情報検索テストコレクションについて解説している。

以下，第 3 章から第 5 章において，本研究で提案している 3 つのシステムの手法について，それぞれの評価結果と共に示している。

第 3 章の「共起単語間の関連性を考慮した文書検索」では，既存の文書検索システムに対する高精度化検索手法として，共起単語間の関連性を考慮した文書検索手法を提案している。まず，単語出現頻度を基とする従来の文書検索モデルでは，検索質問に含まれる検索語同士の関連の考慮が行われておらず，不適合文書が誤検索されるという問題があることを示している。この問題に対して，関連する検索語が共起出現している部分を主題記述部分とする概念を提案し，検索語の共起出現関係の特徴を共起重要度とした検索モデルを提案している。共起重要度として，(1) 検索語間の出現距離の近さの指標である近接出現距離，(2) 検索語の共起出現の確率を用いた共起検索語間の関連度，および，(3) 共起検索語の重要度の 3 つの特徴量を用いた算出手法を提案し，さらに，既存の検索モデルと共起重要度を組み合わせた手法を提案している。日本語の情報検索評価用テストコレクション BMIR-J2（BenchMark for Japanese IR Systems Ver .2.0）を使用し，

従来手法と提案手法の検索精度を比較する評価実験を行うことにより，提案手法は，従来の方式に対して検索精度が向上し，有効であることの結果を示している。

第4章の「検索質問の主題分析に基づく類似文書検索」では，類似文書検索システムにおける高精度検索手法として，検索質問の主題分析に基づく類似文書検索手法を提案している。まず，文書には一般的に複数の主題の記述があり，類似文書検索の検索質問文書と検索対象文書について，主題ごとの対応を考慮しないと誤検索が発生するという問題が生じることを示し，この問題に対して，検索質問文書に含まれる複数の主題を抽出し，各主題に対して主題重要度を付与することにより類似文書の適合度を決定する手法を提案している。主題重要度判定のために，検索語の特定性をエントロピーを用いて算出する手法を提案している。この提案手法を特許文書の請求項を検索質問とする類似文書検索に応用し，NTCIR-4（NII-NACSIS Test Collection for IR Systems）の特許検索テストコレクションを使用し，従来手法に対して検索精度が向上し，有効であることの結果を示している。さらに，提案手法により算出した主題重要度と平均精度の相関の分析を行い，主題重要度が妥当であることを示している。

第5章の「質問応答システムにおけるパッセージ検索」では，質問応答システムにおける高精度検索手法としてパッセージ検索の適用を提案している。まず，一般的な質問応答の処理手順を示し，処理の中には自然言語処理や情報抽出処理による高コストな回答抽出処理が含まれる問題を指摘し，処理の効率性を考慮した場合，処理の初期段階での文書の絞り込み検索の検索精度および絞り込み文章量の少なさが重要であることを示し，この問題に対して，質問応答における絞り込み検索において，処理コストと検索精度の両者を満足する検索方式として，パッセージ検索の適用を提案している。質問応答におけるパッセージ検索の妥当性を評価するために，TREC-9（Text REtreival Conference）の質問応答テストコレクションを用いた評価実験を行い，出力された上位文書中に正解フレーズを含んでいるか否かを示す正解含有率の評価指標を新たに設定し，従来一般的に用いられている精度と合わせて，両指標の評価・分析を行っている。この結果，質問応答ではパッセージ検索が精度，処理コストの面で有効であることを示している。さらに，質問種別により，適したパッセージの長さを変更することが有効であることを示している。

第6章は，「考察」であり，提案手法について評価実験結果から，各提案手法の問題点を考察し，残された課題について整理し，今後の研究課題を示している。

最後に，第7章の「結論」で，本論文の総括を行っている。

審 査 の 結 果 の 要 旨

電子機器を用いた文書の作成が一般的になっている現代社会において，氾濫する電子化文書の情報を検索し，検索した文書の分析による意思決定や情報の再利用を行うことが必須となっており，この情報検索作業を支援する情報検索システムの役割は非常に大きくなっている。本論文は，情報検索システムに関して検索精度向上の観点から研究に取り組んだものである。語の頻度情報を基とする従来の文書検索手法の誤検索に関する問題点を解決するために，文書あるいは検索質問の解析を行い，内在する複数の主題を区別することに着目し，文書あるいは検索質問の本質的な情報を捉える検索手法を提案した。また，多様になっている現在の情報検索システムの形態に対応させ，3つの検索システム（文書検索システム，類似文書検索システム，質問応答システム）に対する検索手法に関する研究を行い，従来手法を上回る検索精度を達成する成果を得ていることは高く評価できる。特に，個々の提案手法について，以下の点に特徴があると認められる。

第1に，文書検索システムに対する情報検索手法として，検索語の共起出現に着目した新たな文書適合度付与方式のための3つの共起特徴量の組み合わせ手法を提案している。単独の共起特徴量のうち，近接出現距離や共起検索語間の関連度については，従来から提案のある特徴量であるが，特徴量を組み合わせて文書

検索に応用することにより，単独で利用する場合と比較して，さらに高い検索精度を実現できることを示している点で，有用である。また，本手法は，従来の文書適合度の算出要素である，語の出現頻度値を補正する方式で実現しているため，多くの文書適合度モデルへの適用が可能であり，汎用性も高い。

第 2 に，類似文書検索システムに対する情報検索手法として，検索質問文書の主題抽出を行い，この主題に対し主題重要度を付与することにより，検索質問中の重要な主題を考慮した手法の提案を行った。主題重要度付与では，各主題における検索語出現の確率分布を用いた手法を提案した。本手法は，検索質問の主題抽出に着目した従来にない独創的な手法であり，従来手法に比べて高い検索精度を実現している点で，有用である。

第 3 に，質問応答システムに対する情報検索手法として，文書の局所的検索が可能なパッセージ検索手法の提案を行った。質問応答システムに特有な高コスト処理の低減に着目した検索結果の文書量に対する評価観点の提案や，従来の文書検索システムとは異なる新たな評価指標として正解含有率の評価手法の提案を行っている。この結果，従来手法に対し高い検索精度を示している点で，有用である。

第 4 に，3 つの検索システムに対する有効性評価においては，当該分野において，認知されている情報検索テストコレクション（BMIR-J2, NTCIR-4 特許検索テストコレクション, TREC-9 質問応答テストコレクション）をそれぞれ利用し，客観的な評価基準により従来手法との比較実験を行い，提案手法の有効性について示していることから信頼性が高い。また，比較実験においては，検索課題特性による分析を行い，提案手法の効果を最大限有効とする結果を示している点も評価できる。

第 5 に，文書あるいは検索質問の主題を特定する手法は，検索結果の重要部分を提示することが可能となる効果もある。本研究の目的である検索精度向上以外にも，情報検索システムの機能として重要な，検索結果の理解支援や納得性向上の効果がある点は，有用である。

なお，情報検索システムの評価軸は，検索精度のみではなく，検索性能を考慮する必要がある。本提案手法は，従来手法と比較して，主題解析処理に関する計算量が追加されるため，高い検索性能が必要とされる実用的な情報検索システム構築の観点では欠点がある。しかし，この点については，本論文において並列処理の適用や，従来手法と提案手法を検索課題特性によって選択的に利用する方式といった一定の対応の方向を示しており，先に述べた本論文の本質的な価値を損なうものではないと判断できる。

また，本研究で実現している検索精度は向上の余地が残されている。主題解析において，高度な自然言語解析の利用によるさらなる高精度検索手法に向けた課題であることを示しているが，著者の今後の研究に期待したい。

以上のように，種々の利用形態に対応する情報検索システムを対象とし，文書検索システム，類似文書検索システム，質問応答システムに対する高精度情報検索手法をそれぞれ確立した本研究は，主題解析に基づく検索手法として独創性も高く，従来の問題点を改善した情報検索手法の提案として，情報学の分野において大きな貢献があると評価できる。

よって，著者は博士（情報学）の学位を受けるに十分な資格を有するものと認める。