氏名	川原（かわはら） 稔（みのる）
学位(専攻分野)	博士（情報学）
学位記番号	論情博第42号
学位授与の日付	平成15年3月24日
学位授与の要件	学位規則第4条第2項該当
学位論文題目	データマイニング技術を用いた情報検索支援に関する研究
論文調査委員	（主査）教授 茨木俊秀　教授 金澤正憲　教授 髙橋豊

論文内容の要旨

本論文は，大規模な文書データの情報検索において，検索要求への明確な認識をもたないユーザをデータマイニング技術を用いて支援する手法に関して，知識発見アルゴリズムと情報検索システムの実行速度に関する研究成果を取りまとめたものであり，6章からなっている。

第1章は序論で，まず，計算機とネットワークの発達によるデジタルデータの大量蓄積と，テキストデータやマルチメディアデータなどへの蓄積形態の変化について述べている。大量の蓄積データからの知識獲得に関しては，データマイニング技術によって導出された知識を用いてユーザの検索要求を明確化することを基本方針として掲げている。そのため，データベース，人工知能，統計学の各分野の知識獲得アルゴリズムについてのサーベイを行い，それらの中から高速な相関ルール導出アルゴリズムを基本とした研究の概要を述べている。

第2章では，相関ルール導出アルゴリズムを改良するため，索引語の重み，構造化文書の付加情報，さらに対象データと独立した特定の領域に関するデータを用いて，支持度および確信度を基準にルール導出を制御する手法を提案し，大規模な文献二次情報データに対して情報検索支援用の実験システムを構築することによって提案手法の定量的な評価を行っている。情報検索では，データ全体から得られる知識よりも検索質問に関係する知識を導出することが求められることから，支持度を索引語の出現頻度と検索質問の包含するデータ領域を考慮したものに置き換えることにより，安定的なルール導出手法を提案している。この改良を基に，構造化文書から得られる付加情報間の概念的な階層を用いて，ルール導出対象となるデータ空間を拡張して，索引語に関連する相関ルール導出手法を提案している。さらに，対象となるデータとは独立した領域から導出されたルールを参照することで，対象のデータ空間にバイアスをかける手法を提案し評価を行っている。

第3章では，情報検索支援システムの応答速度に関して，時間コストモデルを定義し，前処理にかかるコストの評価基準の提案を行っている。すなわち，データマイニングにおける知識獲得アルゴリズムは，対象データが大規模化すると時間コストも増大するため，ユーザの待ち時間増加という応答速度の問題が発生する。そこで，ルール導出を事前に行っておき，検索時にその結果を用いる方法を提案し，計算機の主記憶・補助記憶アクセス，データの更新を考慮したコストモデルを定義している。それを基に，大規模な文献二次情報データを用いて提案方法の定量的な性能評価を行っている。

第4章では，相関ルールの導出基準となるパラメータの閾値の決定手法について述べている。相関ルールアルゴリズムでは，ルール導出の閾値として最小支持度と最小確信度が用いられるが，それらの妥当性の判定に制御分野などで用いられるROC解析手法を取り入れることを提案している。すなわち，検索質問が被覆する文書集合およびその補集合に対して，導出ルールである索引語集合が被覆する文書集合の比率を，それぞれ，TP（True Positive rate）およびFP（False Positive rate）と定義して，ROC解析におけるROCグラフを描き，その凸包を用いて最適な閾値を決定する手法を提案している。このとき，検索質問が被覆する文書数に応じて閾値を動的に変動させること，さらに，導出ルールの性質の優劣を測る基準としてROCグラフ上の距離を用いることなどを提案している。また，提案手法に対して，実験に基づく性能評価を行っている。

第5章では，導出された個々の相関ルールにおいて，TPは確信度に相当するが，FPはアルゴリズムに用いられていないことを指摘し，FPをパラメータとして導入することにより相関ルール導出アルゴリズムの改良を提案している。新規パラメータの導入によって，ベクトル空間モデル型情報検索システムSMARTで定義されているストップワードの導出を抑制する効果があることを示している。導入パラメータとして，FP，および，TPとFPを用いたROC距離を用いる手法を提案しており，いずれの場合においてもストップワードの導出が効果的に抑制されることを，大規模な文献二次情報データを用いて定量的に示している。

第6章は結論で，本論文で得られた結果を総括的にまとめ，今後の課題について述べている。

論文審査の結果の要旨

本論文は，情報検索支援における要素技術として，相関ルール導出アルゴリズムの改良および応答速度の最適化を目的として行った研究をまとめたものであって，得られた主な結果は次の通りである。

1．構造化文書から得られる付加情報間の概念的な階層を用いて，ルール導出対象となるデータ空間を拡張して，索引語に関連する相関ルールが導出可能であることを示した。また，複数の分野が混在したデータに対しても，対象データと独立した特定の領域に関するデータをルール導出に援用することで，対象データ空間にバイアスをかけることが可能であることを示した。

2．情報検索支援システムにおける応答時間に関するコストモデルを導入することで，システム管理者が経験的に与えていたシステムの閾値を客観的に決定可能であることを示し，情報検索支援システムの効果的な運用を実現するシステム設計の指針を与えた。

3．情報検索支援の根幹である相関ルール導出アルゴリズムに，新しいパラメータを導入することにより，ストップワードなどの知識ベースを用いることなしに，不適切なルールの導出を効果的に抑制できることを示した。このことは，情報検索だけでなく一般のデータマイニングにも適用可能であり，様々な分野に影響を与えるものと期待できる。

以上，本論文は，大規模な文書データベースを対象とする情報検索支援システムの構築についての研究結果をまとめたもので，学術上および実際上寄与するところが少なくない。よって，本論文は博士（情報学）の学位論文として価値あるものと認める。また，平成15年2月24日実施した論文内容とそれに関連した試問の結果，合格と認めた。