報告番号	※ 甲 第 10256 号

主 論 文 の 要 旨

論文題目　数量化理論III類・コレスポンデンス分析におけるスコアベクトルの構成法とその応用に関する研究

氏　名　鈴木　誠

論 文 内 容 の 要 旨

ユーザの検索を支援する手法やユーザのニーズを的確に捉える手法としてデータマイニング技術の必要性が高まっている．データマイニングとは「大規模データから未知のパターンを発見する計算プロセス」のことであり，計算科学の学際分野の一つを構成する．

データマイニングにおいて，ユーザの年齢，居住地域，収入などのデモグラフィック（グループ）情報と，購買履歴，好みなどの行動（属性）情報の絞り込みを行うユーザプロファイリングおよび保有する大量の電子情報の中から，必要な文書を探し出す文献検索は，重要な課題である．

データマイニング技術の手法はテキストマイニング手法やベイズ推定による予測手法，マーケット・バスケット分析，ニューラルネットワーク，自己組織化マップなど多種である．また，回帰分析，主成分分析，数量化理論，コレスポンデンス分析などの多変量解析の手法も幅広く用いられている．この中で数量化理論III類・コレスポンデンス分析は，ユーザとデモグラフィック・行動情報を対応させながら，それぞれの関係性を明らかにするユーザプロファイリング手法として捉えることができ，外的基準のない解析手法に分類される．数量化理論III類・コレスポンデンス分析を用いた解析において，ユーザプロファイリングでは，グループ情報と属性情報の絞り込みを行い，グループ／属性間の関係を明確に表すことのでき，データの変動に影響されにくい可視化手法が求められている．また文献検索では，検索キーワードの含／不含に関わらず，ユーザの求める文書を検索・表示できる検索支援技術が求められている．

本研究では，ユーザプロファイリングおよび文献検索において，(1) データのゆらぎを受けにくく，カテゴリ（グループ／属性）間の関係を適切に表すことのできるユーザプロファイリング手法の開発，および，(2) 検索語キーワードの内容を反映して，検索語キーワードを含まなくても関連している文書を抽出でき，検索語キーワードを含んでいても関連性の低い文書を除外して，関連度に応じて検索結果を表示できる検索支援技術の開発を目的として，数量化理論III類・コレスポンデンス分析におけるスコアベクトルの構成法およびデータマイニングへの応用を採りあげる．

第1章では，本研究の背景と目的について示す．

データマイニングの種類と動向，データマイニングにおける重要な課題であるユーザプロファイリングおよび文献検索に関する現況について説明する．多変量解析の種類と特徴を概説，対象とするデータが量的データ／質的データ，また外的基準の有無によって多変量解析を分類し，ユーザプロファイリングおよび文献検索などの質的データを対象とした外的基準のない多変量解析手法として，数量化理論III類とコレスポンデンス分析に着眼する．数量化理論III類・コレスポンデンス分析開発の経緯を概説し，簡単な解析例を示すことによって，数量化理論III類・コレスポンデンス分析を用いたユーザプロファイリングおよび文献検索における問題点を示す．さらに本研究の目的と構成を示す．

第2章では，本研究の第3章および第4章で用いる相関の定義として，数量化理論III類・コレスポンデンス分析をもとにしたサンプルおよびカテゴリ間の相関を示す．

まず従来用いられている数量化理論III類・コレスポンデンス分析を説明し，次に全スコア軸にそれぞれ対応する固有値の二乗根を掛け，全スコア軸の情報を用いてカテゴリスコアベクトルおよびサンプルスコアベクトルを定義する．本定義で，内積値で相関を表示する場合，カテゴリ間の相関関係を直感的にも分かりやすいものとするため，内積値に1加算して，最小値が0とする．この処理は，最も大きい固有値とこれに対応する固有ベクトルもカテゴリスコアベクトルに用いることと対応していることを示す．サンプルスコアベクトル間の内積によりサンプル間の相関を，カテゴリスコアベクトル間の内積によりカテゴリ間の相関を定義する．従来手法との比較，先行研究の調査から，本研究で定義するすべてのスコア軸を用いてスコアベクトル間の相関を求めるデータマイニング手法が従来は用いられていないことを示す．

第3章では，前章で定義したカテゴリ間相関をもとに，グループ情報間の相関関係の可視化法と属性情報とグループ情報間の相関関係の可視化法を示す．この可視化法を用いて，アンケートデータのグループ情報と属性情報の関連付けを行い，ユーザプロファイリングを行う．

グループ情報間の相関関係の可視化のために，対応するカテゴリスコアベクトル間の角度を求め，この角度関係を維持しながら2次元平面上にグループを布置する手法を示す．さらに，グループと属性からなる各カテゴリ間の関係を可視化するために，属性情報とグループ情報のカテゴリスコアベクトルの内積値を用いて，属性（行動）ごとに各グループの定量的プロファイルを表す方法を示す．この手法をテストデータおよびアンケートデータに適用して，グループのプロファイルを把握しやすい可視化ができることを示す．テストデータを用いて，内積値によって元のデータの特徴を抽出できること，およびカテゴリスコアベクトル間の角度および行動情報とグループ情報のカテゴリスコアベクトルの内積値を可視化する方法を示す．さらに，こだわりの領域に関するアンケートデータを用いて，本章で示した可視化手法による各対象に対する内積値の一覧情報および詳細情報を示す．アンケートデータは，情報機器に対する20代から60代までの男女の，こだわりの領域を回答するもので，本研究の提案手法によって従来の数量化理論III類のようなユーザプロファイルの俯瞰に留まらず，定量的で詳細の個別の属性の特性を可視化できることを示す．また，同じアンケートデータを用い，本研究で定義した相関と新たに導入した可視化手法によって，スコア軸に割り当てられる特徴の順位が入れ替わるような，データにゆらぎがある場

合でも，安定した結果が得られることを示す．本章で示した可視化法によって，世代ごとの変化や男女の特徴，さらに世代による嗜好性の男女逆転など，マーケティングを行う際に重要となる，デモグラフィック情報と行動情報の関連付けできることを示す．

第4章では，数量化理論III類・コレスポンデンス分析において，文書と検索語キーワードの相関を，サンプルスコアベクトル間の類似度によって定義し，検索語キーワードに近い順に検索文書を提示する手法を示す．

文献検索の手法として，シソーラス（類義語辞書）を使用する／しない，電子ファイルの属性情報を使用する／しないによって分類し，シソーラスおよび属性情報を使用しない文献検索手法の性能改善をめざし，数量化理論III類・コレスポンデンス分析を適用できることに着眼する．文書スコアベクトルに加えて新たに検索語スコアベクトルを導入する．文献検索時には，検索語スコアベクトルと個別の文書スコアベクトルとの相関として，類似度を求め，この結果をもとに検索結果の順位付けを行う．提案手法によって検索キーワードの含／不含に関わらず，いずれも安定した検索特性が得られることを示す．また，検索語スコアベクトルを用いることによって，文書蓄積時にあらかじめ文書スコアを計算することが可能になる．文書中に検索語キーワードは含まれていないが関連性の高い文書，逆に検索語キーワードは含まれているが内容はあまり関連のない文書に対する提案手法の検索性能を定量的に評価するため，検索語キーワードの削除および追加による実験を行う．名古屋大学から出願された公開特許1000件の検索に本手法を適用し，提案手法をキーワード一致およびLSI（潜在意味インデックス）による従来法と比較する．検索語キーワードは，異なった分野から「遺伝子」,「カーボンナノチューブ」および「びびり振動」を用いる．提案手法によって各検索語キーワードに対して抽出される文書，検索に寄与するキーワードを示す．評価には文書抽出における適合度と再現率の調和平均であるF尺度を用いる．LSIでは，各単語の出現頻度において単語の出現頻度と逆文書頻度をもとにしたtf-idfによって重みづけを行い，特異値分解を行う．このような条件でも，提案手法によって検索語キーワードの含／不含に関わらず，キーワード一致およびLSIの従来手法と比較して，いずれも安定した検索特性が得られることを示す．また，検索に使用するスコア軸数を変化させる場合の影響，およびスコア軸に掛ける重みの有無の影響を示す．検索語キーワードの追加・削除において，追加・削除を行う件数を増すに従って，キーワード一致検索と比較して提案手法の特性低下が少ないことを示す．さらに検索語キーワードの削除と追加が同時に起こる場合の特性も示す．提案手法は，構文解析を行う段階で辞書を使用するものの，検索では類義語辞書などのシソーラスを使用することなく，文書の検索を実現する．

本実験で使用した文書では，前処理は最低限に留めているため，構文解析によって抽出されるカテゴリ情報には対象の情報を適切に反映していないものも多数含まれる．このような条件であっても提案手法により効果的な抽出ができていることから，今後はより高度な構文解析手法を組み合わせる方法や，他の手法との連携などの応用も期待できることを示す．

第5章では，本研究で得られた成果をまとめ，今後の課題と展望について述べる．