独立成分分析を用いた情報フィルタリングに関する研究

論文要旨

近年、急速な情報化が進み、一般ユーザが数多くの電子化された情報を容易に入手できるようになっている。それらの情報を効率よく利用するために、ユーザが求める情報を探し出す情報検索システムが数多く開発され、とくに、キーワードを用いた検索システムが普及している。しかし、キーワード検索システムでは、多くのユーザにとって必要な情報を得るための的確なキーワードを選別することは難しく、検索結果に不要な情報が含まれていることが多い。このような状況を解決するために、検索結果の表示順序をソーティングするランキング手法や、ユーザに必要な情報のみを選び出す情報フィルタリング手法が研究されている。

ユーザにとって必要な情報を選択する基準として、それらの情報に対する興味の有無が用いられ、その興味情報はユーザプロファイルと呼ばれている。ユーザプロファイルとしては、ユーザが興味のある単語で表現する方法や、過去にユーザが評価した情報から関連フィードバックや遺伝的アルゴリズムを用いて、ユーザの興味を抽出する手法が提案されている。ドキュメント検索システムにおいては、ドキュメントやユーザプロファイルはベクトル空間法により、索引語と呼ばれるドキュメントの特徴を表わす単語に対する重みを要素とするベクトルで表現される。このとき、扱うドキュメントが増えるにともない、索引語数が増え、ベクトルの次元数が増大するので、ユーザの興味抽出の処理効率に問題が生じている。

上述の問題を解決するため、ドキュメント中に含まれる潜在的意味を用いて、ドキュメントベクトルを変換する手法が提案されている。潜在的意味とは、索引語に対する重みのような表面的な特徴ではなく、統計的な解析などを施すことで浮かび上がってくる特徴である。ドキュメント中の潜在的意味を解析する手法として、LSA（Latent Semantic Analysis：潜在的意味解析）と呼ばれる手法が代表的である。LSAでは、各索引語に付与された重みのドキュメント間の分散に着目して、ドキュメント間の相互関係を解析し、その潜在的意味を用いて、ドキュメントベクトルの次元削減

を行なっている。

一方、音声処理や画像処理の分野で注目を集めている ICA（Independent Component Analysis：独立成分分析）を用いて、ドキュメント中の潜在的意味を抽出する試みもすでに報告されている。そこでは、ICA が独立成分を抽出できるという性質に着目し、ドキュメント中の潜在的意味を解析している。とくに、ICA により求められる独立成分は、特定のテーマに関連した索引語の集合（以後、トピックと呼ぶ）を表現できることが報告されている。それらのトピックは LSA によって得られる潜在的意味に比べ、独立性という基準を導入することによって、ドキュメントに含まれる特徴をより明確に表現した潜在的意味を捉えている。したがって、そのトピックを情報フィルタリングに用いれば、精度の良い情報フィルタリングが実現できると思われるが、現在までに、そのような手法に関する研究は報告されていない。

本研究では、ICA によって得られるトピックを用いてドキュメントを表現し、そのトピックによって表現されたドキュメントを利用してユーザプロファイルを作成し、情報フィルタリングの精度を改善する手法を提案した。また、ICA を用いて情報フィルタリングの精度を改善するための手法として、ICA によって得られるトピックに基づいて不必要な索引語を除去する手法も提案した。本論文の構成は以下のとおりである。

第 1 章では、本研究の背景ならびに目的を述べるとともに、研究内容の概要について述べた。

第 2 章では、ICA によって得られるトピックが構成する空間にドキュメントベクトルを写像し、その空間上でユーザプロファイルの作成を行なう手法について述べた。従来手法の LSA では、潜在的意味空間上の基底ベクトルは直交性という性質に着目しているが、本手法では、独立性に着目し、得られた独立成分にそのドキュメントに含まれるトピックという意味付けを与えることができた。

さらに、ユーザプロファイルを作成する方法として、学習ドキュメントの重心を用いる方法と遺伝的アルゴリズムを用いる方法を採用した。その結果、ユーザプロファイルの作成方法に関わらず、もとのドキュメントベクトルを直接用いる従来方法と比較して、ICA によって得られるトピックを用いてドキュメントを表現する提案手法

が、ユーザプロファイルの推薦精度と作成時間の点で有効であることを明らかにした。

第 3 章では、ICA によって得られるトピックに対して、トピックの選択を行なうことによって、ユーザプロファイルの推薦精度を改善する手法を提案した。ICA によって得られたトピックを用いて、ユーザプロファイルの精度を改善しようとする際、不必要なトピックやノイズ成分が含まれている。そこで、トピックには類似性が存在すると考え、トピックをクラスタリングし、必要なトピックを選択した。クラスタリングには、MDA（Maximum Distance Algorithm：最大距離アルゴリズム）を用いた。一般的に MDA の距離関数としては、ユークリッド距離が利用されている。しかし、ICA によって得られるトピックにはスケールと順序に任意性があるため、ユークリッド距離では各トピック間の類似性を評価できない。そこで、トピックに含まれる索引語の重みの分布の類似性に着目し、それを測る尺度として、昨今、情報幾何の分野において注目されている統計的情報量を用いた。本手法では、Jensen-Shannon 情報量を MDA の距離関数として導入し、クラスタリングを行なう手法を提案した。本手法でトピックの選択を行なうことにより、推薦精度が改善できることを明らかにした。

第 4 章では、SVD（Singular Value Decomposition：特異値分解）と ICA を組み合わせた潜在的意味の解析手法と、それらの潜在的意味を情報フィルタリングへ応用する手法を提案した。第 3 章の手法では、ICA によってトピックを求めた後、必要と思われるトピックを選別したが、ドキュメント数が増えるにしたがい、ICA の処理時間が増大する。そこで、ICA を適用する前に、累積寄与率に基づいて SVD の基底ベクトルを選択し、ノイズ削減と低次元化を行なった。また、それらの基底ベクトルが張る空間上で ICA を適用することにより、ICA の処理時間の短縮や推薦精度の改善を試みた。その結果、ノイズ削減を行なうことで推薦精度の改善を行なうとともに、低次元化により処理時間の大幅な短縮が実現できることを明らかにした。

第 5 章では、ICA を行列分解の一手法と考え、ドキュメントを構成する潜在的意味として ICA の混合行列を採用し、その潜在的意味を情報フィルタリングに応用する手法を提案した。従来、ICA の混合行列を潜在的意味とみなす研究が報告されており、その潜在的意味は、独立性や直交などの性質は仮定できないが、第 2 章で述べた手法に比べて、人間の考えるトピックの感覚に近いことが知られている。そこで、それ

らの潜在的意味を用いて情報フィルタリングを行なうことで、先に述べた手法とは異なる結果が得られると考え、その潜在的意味空間上でユーザプロファイルを作成し、ユーザプロファイルの推薦精度が向上できることを明らかにした。

第 6 章では、ICA によって得られるトピックに基づいて索引語の選別を行ない、情報フィルタリングの精度を改善する手法を提案した。従来、ドキュメント中から重要語を抽出するキーワード抽出という研究が行なわれている。それらの研究では、主にドキュメント中の高頻度語に焦点が当てられている。情報フィルタリングに用いるドキュメントベクトルの索引語を作成する際には、ドキュメント集合中に含まれるすべてのドキュメントに対して特徴付けを行なう必要がある。したがって、高頻度語が表わすドキュメントの大まかなトピックだけではなく、様々なトピックに関する単語も抽出しなければ、精度の良い情報フィルタリングを実現することが困難である。

ここでは、ICA によってドキュメントのトピックを解析し、そのトピックを表わす代表的な語と共起する語を索引語として用いることで、様々なトピックに関連する特徴的な語を索引語として選別した。さらに、それらの索引語を用いてドキュメントベクトルを再構築し、学習ドキュメントの重心を用いて、ユーザプロファイルの作成を行なった。そのユーザプロファイルによる推薦精度を検討した結果、興味のあるドキュメントが少ない状況では、推薦精度改善に効果があることが明らかとなった。

最後に、第 7 章では以上の本研究について総括を行ない、今後の検討課題について述べた。