称号及び氏名	博士（工学）柳本　豪一
学位授与の日付	平成１８年４月２０日
論　文　名	「ユーザの興味を反映した情報フィルタリングの構築に関する研究」
論文審査委員	主査　大松　繁 副査　辻　　洋 副査　黄瀬　浩一

ユーザの興味を反映した情報フィルタリングの構築に関する研究

論文要旨

パーソナルコンピュータの普及にともない電子化された情報が増え，利用者が必要とする情報がコンピュータ上で検索できる可能性が高くなった．しかし，利用者が膨大な情報の中から必要とする情報を取捨選択できにくい状況や，利用者を満足させる情報が多くの関連した情報に紛れてしまう状況に陥ることが多い．このような問題を解決するため，情報検索や情報フィルタリングが開発されている．情報検索も情報フィルタリングもともに利用者の要求を満足する情報を利用者に提供することを目的としている．このため，これらの手法は多くの要素技術を共有しているが，取り組んでいる問題は異なっている．例えば，情報検索は利用者の情報に対する要求の満足や変則的知識状態の解消を目的とし，情報フィルタリングは長期間に渡って利用者が関心をもつ情報を提供することを目的としている．

このため，情報検索は，利用者による検索要求から最終的に検出された情報に対する評価までの処理に対する利用者の満足によって評価される．したがって，情報検索では情報の表現方法や利用者の検索要求と情報の比較方法が研究対象となる．情報フィルタリングは，情報提供を繰り返すことによる利用者の満足によって評価される．したがって，利用者の長期的な興味の表現方法やその作成方法が研究対象となる．

情報フィルタリングは，情報選別に用いる特徴量から実システムを分類すると，(a)コンテンツベースのフィルタリングシステム，(b) 協調フィルタリングシステムとなる．コンテンツベースのフィルタリングシステムでは，対象となる情報の内容を用いるので，情報

の内容をどのように表現するかが重要である．協調フィルタリングシステムでは，情報に対する他の閲覧者からの評価を用いて情報の選別を行うので，利用者間の類似性をどのように評価するかが重要である．

次に，情報フィルタリングシステムを興味抽出方法から実システムを分類すると，(a) 利用者へのアンケートより得られたトピックやキーワードを興味とする方法，(b) 閲覧した情報に対する評価から興味を作成する方法，(c)利用者の閲覧履歴や行動パターンから閲覧した情報に対する評価を推測して興味を作成する方法がある．アンケートを用いる方法は，初期のフィルタリングシステムで用いられた方法であり，主要技術はキーワードマッチングである．閲覧情報に対する評価から興味を作成する方法では，利用者が評価した文書を学習データとして識別器を構築する問題と考え，利用者の興味を作成する．このため，どのような識別器を構築するかが問題であり，従来は線形の識別器が多く利用されていた．利用者の閲覧履歴や行動パターンから評価を推測し興味を作成する方法では，利用者が何に意識を向けているかを把握することが重要である．

本論文では，文書を対象とし，利用者が評価した文書から作成した興味を用いたコンテンツベースの情報フィルタリングシステムを提案する．とくに，非線形な利用者の興味モデルを提案し，利用者の評価を用いてその興味モデルを最適化し，情報の選別性能の向上を図る．

本論文の構成は，以下の通りである．

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，本研究の位置づけを明確にする．

第 2 章では，利用者の興味を適切に表わすユーザプロファイルを作成するため，ユーザプロファイルの評価方法を提案し，そのユーザプロファイルを用いて文書の選別性能の改善を図る．ユーザプロファイルは文書が利用者の興味に適合しているか否かを判定するスコアを求めるために利用され，そのスコアに応じて文書が分類される．スコアの計算には cosine 距離を用いるため，線形識別器を構築することに対応する．このようなユーザプロファイルを改良するため，新たな評価方法を導入することで従来手法に比べ選別精度の改善を図る．ユーザプロファイルの評価方法は，(1) 文書を興味に応じて選別できていているか，(2) 興味の無い文書群と興味のある文書群を明確に分離できているかの観点から行った．興味に応じた文書の選別性能を評価するため，誤選別を少なくするという基準を用いた．これにより，学習に用いる文書が線形分離可能であるときには，誤選別が 0 となるユーザプロファイルが作成できる．しかし，誤選別が小さいという条件を満たすユーザプロファイルは数多く作成できるため，別観点からユーザプロファイルを評価する必要がある．したがって，学習文書において興味のある文書群と興味の無い文書群のスコアの差に注目

し,この差が大きくなるユーザプロファイルを求め,文書に対する興味の有無を分離する.評価には，興味のある文書群中の最小のスコアと興味の無い文書群中の最大のスコアを用いることで，文書群間の差の評価を行う．上記の評価にもとづいたユーザプロファイルを作成するため,単峰性正規分布交叉を用いた実数値遺伝的アルゴリズムを利用する.また,文書の選別実験を行うことで，提案手法の有効性を確認する.

第 3 章では、組み合わせ学習アルゴリズムである AdaBoost を用いて，非線形な識別関数として表現されるユーザプロファイルを作成し，文書の選別性能の改善を図る．非線形な識別関数として表わされる利用者の興味モデルは文書の選別性能に影響を及ぼすため，適切な非線形識別関数をあらかじめ設計することは重要である.

このため，利用者が評価した学習文書に応じて非線形識別関数を構築することが望ましい.これを実現するため,組み合わせ学習アルゴリズムを導入する.組み合わせ学習では,複数の単純な学習機械を組み合わせることにより，学習文書に応じた複雑な識別関数を構築することができる．本手法では，AdaBoost を用い，弱学習機械は遺伝子的アルゴリズムを用いて構成する．このとき，計算時間を考慮して，遺伝的アルゴリズムのパラメータを調整する．さらに、文書の選別実験を行うことにより、提案手法の有効性を確認する.

第 4 章では，カーネル法を用いて従来の線形識別器として構成されるユーザプロファイルを非線形化し，文書の選別性能の改善を図る．とくに，未評価の文書をランキングするためのユーザプロファイル作成手法を提案する．カーネル法は，入力データを高次元の特徴空間へ写像し特徴空間で線形識別器を構築するとき，特徴空間での内積の計算量の増加を抑える写像を提供する手法である．特徴空間上で構築された線形識別器は，入力空間では非線形識別器とみなされるので，複雑な境界面を記述することが可能である．本手法では，文書のランキングが可能である関連フィードバック法をもとに，カーネル法による特徴空間上でのユーザプロファイルを作成する．また，文書の選別実験を行うことにより，提案手法の有効性を確認する．実験では，カーネル関数として多項式カーネル，ガウスカーネル，シグモイドカーネルを用い，各カーネル関数のパラメータを変更し，その影響についても考察する.

第 5 章では，確率空間での文書のフィルタリングを考え，Kullback Leibler 情報量をユーザプロファイルと文書の距離として利用し,文書の選別性能の改善を図る.本手法では,確率空間で処理を行うため，文書およびユーザプロファイルを離散確率分布として表現する．確率空間で文書を表現することは従来のベクトル空間法では捕らえきれない特徴を扱うことができる．このためには，確率空間で文書とユーザプロファイルの類似性を求める必要がある．本手法では，情報幾何の分野で偽距離として利用される Kullback Leibler 情報量を文書とユーザプロファイルの類似性を評価する関数として用いる．これにより，

確率空間上でユーザプロファイルを作成することを実現する．ユーザプロファイルを作成するため，Kullback Leibler 情報量を適合度関数として遺伝的アルゴリズムを用いることで，適切なユーザプロファイルを作成する．さらに，文書の選別実験を行うことにより，提案手法の有効性を確認する．

最後に，第 6 章では、本研究で得られた結果について総括し、今後の課題について述べる。

審査結果の要旨

本論文は、利用者の評価からその利用者の興味を抽出することを目標とし、遺伝的アルゴリズム、AdaBoost、カーネル法、確率的言語モデルを用いて、ユーザプロファイルによる判別性能の改善を行い、テストコレクションを用いて性能の検討を行ったものであり、以下のような成果を得ている。

（１） ユーザプロファイルが有すべき特性を検討し、非線形なユーザプロファイルの評価関数を提案した。非線形関数を用いて最適なユーザプロファイルを作成するため、遺伝的アルゴリズムを用いた探索手法を提案した。その結果、従来のユーザプロファイルの判別性能を改善できることが明らかになった。

（２） 非線形なユーザプロファイルを作成するため、組み合わせ学習手法である AdaBoost を導入した。AdaBoost における弱学習機械の作成に遺伝的アルゴリズムを用いる方法を提案した。その結果、利用者の興味を適切に表現するユーザプロファイルが作成できることが明らかになった。

（３） 高次元特徴空間で線形のユーザプロファイルを作成することで、入力空間における非線形なユーザプロファイルの作成を実現した。文書データを高次元特徴空間に写像するため、カーネル関数を導入した。この結果、文書の判別性能を改善できることが明らかになった。

（４） 文書を表現するモデルとして確率分布を採用し、Kullback Leibler 情報量を用いることで、確率空間上で文書の近さを定義できることを示した。Kullback Leibler 情報量を用いた適合度関数を提案し、遺伝的アルゴリズムにより最適化を行った。その結果、文書の選別性能を改善できることが明らかになった。

以上の諸成果は、非線形なユーザプロファイルの作成手法を提案するものであり、情報検索分野に貢献すること大である。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明したものである。