様式３

論　文　内　容　の　要　旨

氏　名　（　西村　章宏　）	
論文題名	潜在トピックを利用した協調フィルタリングにおける文書の内容が推薦性能に与える影響

論文内容の要旨

大量の情報の中からユーザの情報獲得を支援する方法の一つとして推薦システムがある．推薦システムを実現する方法としては，内容ベースフィルタリングと協調フィルタリングに大きく分けることができ，これらを組み合わせた手法も数多く存在する．協調フィルタリングの中では，Matrix Factorization (MF) が評価値の欠損が多い現実のデータセットに対して優れた推薦を行うことができる．ただし，MFは評価値の欠損に強いものの，評価値が極端に少ない場合には適用することが困難である．これは一般にコールドスタート問題と呼ばれており，評価値だけでは不足している情報を補うために，評価値以外の情報（サイド情報と呼ばれる）が利用される．評価値とサイド情報を合わせて活用するアプローチとしては，協調フィルタリングと内容ベースフィルタリングの両方を組み合わせたハイブリッド法が主流である．近年のハイブリッド法の多くは，結合させる手法に前後関係がなく同等のレベルでモデル化を行う完全結合である．MFをベースとした完全結合に注目すると，サイド情報としてアイテムに関する文書を利用する手法が多く提案されている．その中で Collaborative Topic Regression (CTR) は文書から抽出したトピックを利用するアプローチの先駆けとなったモデルであり，その組み合わせ方法がシンプルであるため様々な派生手法を考える上で基準となる手法である．本論文では，CTR において3つの課題を提起する．課題1は``ユーザに関する文書を用いる場合の推薦性能についての詳細な分析″であり，従来のアイテムに関する文書を用いる場合に比べて文書から抽出されるトピック内容にどのような差異があるか，そしてトピックの差異により推薦性能にどのような差異が生じるか分析する必要がある．課題2は``評価値行列と文書が持つ特徴およびハイパーパラメータ間の関係性が推薦性能に与える影響の分析″であり，評価値行列と文書における特徴とは何であるか具体的に定義し，そられの特徴とハイパーパラメータ間の関係についての仮説を立てて検証する必要がある．課題3は``文書中の語彙を制御することによる推薦性能に与える影響の分析″であり，文書中の単語を何らかの基準によって削除，または他の情報源を利用して追加することによる推薦性能への影響を分析する必要がある．この3つの課題への取り組みとして，著者は4つの研究を行った．課題1に対しては，ユーザに関する文書を用いた際のトピック内容と可視化結果についての分析を行った．その結果，ユーザに関する文書から抽出されたトピック内容はユーザが興味を持っている対象やユーザの属性であり，これらはユーザ層という観点を表していることが分かった．次に，CTRモデルにおいて，トピック情報源としてアイテムに関する文書情報を用いた場合と，ユーザに関する文書を用いた場合を実験により比較し，正確性と利便性という2つの観点から推薦性能がどれだけ異なるのかを分析した．実験の結果，アイテムに関する文書を用いた場合では利便性の高い推薦，ユーザに関する文書を用いた場合では正確性の高い推薦を行う傾向が分かった．課題2および課題3に対しては，スパース度と利用する単語の種類およびハイパーパラメータにおいて推薦性能に影響を与える要因について仮説を立て，複数のデータセットを用いた実験によりそれらの要因間の関係性についての分析を行った．実験の結果，スパース度と文書頻度の組み合わせに関しては，正確性と利便性のバランスが良い推薦を行うのに複数のデータセットで共通する傾向が存在することが分かった．課題3に対しては，アイテムに関する文書とは異なる情報源を用いて文書の内容に合った単語を追加することにより，推薦性能が向上するケースが存在するかの分析を行った．実験の結果，興味の幅が広いユーザに限定した場合には，語彙補完を行うことで推薦性能が向上することが分かった．最後に，本研究をふまえて筆者が考える今後の展望を述べる．

様式７

論文審査の結果の要旨及び担当者

氏　名　（　西村　章宏　）		
	（職）	氏　名
論文審査担当者	主査　教授	佐藤宏介
	副査　教授	飯國洋二
	副査　教授	長井隆行
	副査　教授	土方嘉徳（関西学院大学）

論文審査の結果の要旨

○課題設定の明確性：本論文は、検索サービスが利用者に提供する推薦において、今日最も主要な推薦アルゴリズムである協調フィルタリングの挙動やデータ依存性、コールドスタート問題が未解明でありそれらを明らかにするという明確な問題意識に基づき、研究の意義や必要性が的確に述べられているものと認める。
○先行研究・資料の取扱いの適切性：当該分野の先行研究を渉猟し理解したうえで、協調フィルタリングにおける各利用者と各推薦対象とを行列化しその行列分解に基づいて推薦を行うMatrix Factorization（MF）と、各利用者に関する文書情報または各推薦対象に関する文書情報から抽出した潜在トピックを組み合わせるアルゴリズムであるCollaborative Topic Regression（CTR）に対して、十分に検討されていなかった評価値の存在比率や、利用単語の頻度特徴、アルゴリズム中のパラメータが推薦性能に与える影響を調査しており、本論文は提案する手法を当該分野の研究動向の中に適切に位置づけているものと認める。
○研究方法の妥当性：上記目的に照らして、課題を4種に整理し、１）利用者に関する文書から抽出した潜在トピックが推薦性能に与える影響、２）CTRにかかるハイパーパラメータが推薦性能に与える影響、3）文書中の語彙抽出の制御が推薦性能に与える影響、4）アイテムに関する文書の語彙補完が推薦性能に与える影響を取り上げ、CTRにおいて推薦対象に関する文書を用いるiCTR、利用者に関する文書を用いるuCTRの各手法におよびベンチマークとしてMFを設定し、短文投稿SNSサービスTwitterで有名人に関して言及したtweet、言及した利用者のユーザプロフィール、有名人本人が発したtweet、ソーシャルブックマークサービスCiteULikeでブックマークを行った利用者、利用者が登録した全ての論文と科学技術文書、イラストコミュニケーションサービスpixivにおけるイラストを推薦する利用者とイラストのタイトル・紹介文、書籍情報や批評を閲覧するウェブサイトGoodreadsの5万人の利用者データと書籍のタイトル・説明文を、自らクローリングして構築した性質の異なるビッグデータ間で精密に比較する実験に基づく分析は、適切な研究方法と分析が用いられているものと認める。
○学術的・社会的な貢献：本論文は、潜在トピックを利用した協調フィルタリングにおいて、適切なハイパーパラメータを策定できただけでなく、評価値行列のスパース度に応じて文書頻度を基準に語彙を抽出することで性能低下を抑制し、興味の幅が広い利用者に対して異なる情報源を追加して性能向上が図れること、ビッグデータ実験の結果、スパース度と文書頻度の組み合わせに関して正確性と利便性のバランス良い推薦は複数のデータセットで共通する傾向が存在することを見出したことは特筆できる知見と言える。加えて、評価値行列の特徴を調べることにより文書利用の性能向上の可能性が判断できるように実用性が高められ、評価値行列のスパース度と要求する推薦性能に応じて、推薦対象に関する文書、利用者に関する文書のいずれが優先するべきかを判断できるようにし推薦の説明性を高めたことは、国際的な学術水準及び学際的観点から見て、十分な独創性や重要性があり、社会的要請にも応える可能性を持つものと認める。
○構成・表現・表記法の適切性：本論文は、利用者文書抽出トピックと文書種類、スパース度と単語種類、語彙補完と関心幅が与える影響を、学術論文として体系的に構成されており、適切な表現・表記法によって記述されているもの、○論旨の明確性・一貫性：本論文の著述並びに図版、データは、上記の研究目的や分析、結果、考察の全ての過程においてその論旨が明確かつ一貫しており、論理的に明確な結論が導かれているものと認める。

本審査委員会は、基礎工学研究科における学位論文に係る評価に当たっての基準に基づき、その全ての評価項目を満たしていることを踏まえて、本論文を博士（工学)の学位論文として価値のあるものと認める。