氏　　名	村田　博士
学位（専攻分野）	博士（情報学）
学位記番号	総研大甲第1510号
学位授与の日付	平成２４年３月２３日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当
学位論文題目	サポートベクターマシンを用いた対話的文書検索
論文審査委員	主　査　教授　山田　誠二 教授　佐藤　健 客員教授　相澤　彰子 准教授　市瀬　龍太郎 連携教授　小野田　崇　東京工業大学

論文内容の要旨

近年の情報技術の発展に伴い,個人で扱えるテキストデータの量が急激に増加している．このような状況で，膨大なテキストデータから必要な情報を検索する機会も増え，情報検索，特に文書検索に対する期待が高まっている．文書検索では，一つの適合文書を見つければよいタスクもある一方，システムから提示された文書をユーザが評価する負荷をおってでも，できるだけ多くの適合文書を獲得したいタスクも多い．本研究で扱う検索タスクは後者であり，このようなユーザから対話的にフィードバックをかける情報検索システムは，その有効性の検証とともにさまざまな研究が展開している．

一般に，ユーザが検索意図を記述したクエリ（検索キーワード）による一回だけの検索により，多くの適合文書を獲得することは容易でない．そのため，ユーザからのフィードバックを利用して検索を繰り返すことで，できるだけ多くの適合文書を検索することが現実的である．このようなユーザからのフィードバックを利用する手法として，検索結果である文書をユーザに提示して，ユーザが適合，非適合の判定を行い，その判定結果をもとに，より精度の高い再検索を行うことを繰り返す適合フィードバック（relevance feedback）がある．

この適合フィードバックを対話的分類学習として捉え，現在最も性能の高い分類学習アルゴリズムの一つであるサポートベクターマシン(SVM : Support Vector Machines)を適用する方法が提案されている．先行研究として，分類学習としてサポートベクターマシンを適用し，比較実験により検索性能が向上することを示した研究，特に適合文書が文書データベース中に少ない場合に，サポートベクターマシンベースの対話的文書検索が有効であることを確認した研究などがあるが，これらの先行研究ではサポートベクターマシンを用いた適合フィードバックにおける提示文書の選択に対する詳細な研究は行われていない．

そこで本研究では，サポートベクターマシンを用いた適合フィードバックにおける対話的文書検索での検索性能と学習性能をともに向上させる，ユーザへ提示する文書の選択のためのヒューリスティクスを提案し，それにより単純な文書提示や従来の適合フィードバックを凌ぐ性能向上が達成されることを実験的に示す．この文書提示のヒューリスティクスは，大規模の文書データにおける正データと負データの極端な偏りに基づくもので，サポートベクターマシンのサポートベクターを効率よく集めることができる能動学習を実現している．

文書検索のテストベッドとして広く使われている国際会議TREC（Text REtrieval Conference)のデータセットを用いて,従来の適合フィードバックで用いられるRocchioの手法と，従来のサポートベクターマシンにおける能動学習手法による提示文書選択方法について，提案手法との比較実験を行った．その結果，提案手法は，従来手法に比べて検索性能，学習性能とも同等あるいはそれ以上のパフォーマンスであることが示された．

また，このときの文書提示における，提示文書の順位付けは，その文書が適合文書にどれだけ似通っているかを判定する適合度を計算することで行う．サポートベクターマシンを用いた適合フィードバックのシステムにおいて，この適合度として判別関数と文書ベクトル間の符号付距離が用いられるが，この適合度が従来手法であるRocchioの手法などで用いられているベクトル空間モデル上でどのような特性を持つのかは明らかになっていない．

そこで本研究では，サポートベクターマシンにおける距離を用いた適合度を定式化し，対話的文書検索における従来手法であるRocchioの手法との比較分析を行う．

比較分析を行った結果，Rocchioの手法におけるクエリベクトル更新式が，サポートベクターマシンに基づく適合フィードバックの重みベクトルの近似となっていること，そして，Rocchioの手法ではクエリベクトルとのコサイン類似度となっている適合度の計算式が，サポートベクターマシンに基づく適合フィードバックについては，重みベクトルとのコサイン類似度に評価対象文書ベクトルのノルムをかけたものになっていることがわかった．

この比較分析により得られた知見から，サポートベクターマシンに基づく対話的文書検索における類似度が文書ベクトルのノルムの影響を避けて文書ベクトルモデル元来のコサイン類似度に近づける効果のあるカーネルとして，コサイン類似度に対応したコサインカーネルを提案した．提案手法の有効性を検証するために国際会議TRECのデータセットを用いた検証実験を行い，Boolean，TF，TFIDFの文書ベクトル表現について比較した結果，すべてのベクトル表現で性能が向上し，特にTFベクトル表現において，性能が大きく向上することを示した．

博士論文の審査結果の要旨

出願者，村田博士氏は，「サポートベクターマシンを用いた対話的文書検索」と題する論文を提出し，この論文およびその内容に基づく研究発表に基づき博士論文の審査が行われた．本論文は，サポートベクターマシンSVMによる対話的文書検索の枠組みにおいて，能動的文書提示を効果的に行うヒューリスティクスの提案とその実験的評価を行い，さらに従来の適合フィードバックであるRocchio法とSVMの比較分析によりコサインカーネルを提案し，その有効性を実験的に評価している．

本論文は，序論から結論までの全5章からなる．第1章「序論」では，アクセス可能な膨大な情報に囲まれた現在の情報社会において，ユーザが欲しい情報をいかに効率的に検索するかが課題となっていることが本研究の背景として述べられている．また，そのような環境において，ユーザによる提示文書の適合判定であるユーザフィードバックを受けて対話的に文書検索を進める対話的文書検索をより効率的で精度の高いものにすることが本研究の目的であると述べている．

第2章「関連研究」では，文書検索で従来用いられているモデルを概観し，本研究で用いられる機械学習アルゴリズムであるSVMの基本的概念を説明している．さらに，能動学習，適合フィードバック，そして関連する機械学習であるトランスダクティブ学習，ランキング学習について解説している．また，これらの関連研究に対する本研究のオリジナリティについて議論し，研究の位置づけを行っている．

第3章「サポートベクターマシンに基づく能動学習によるインタラクティブ文書検索」では，対話的文書検索において能動学習に対応する能動的文書提示について述べ，その文書提示に有効である“SVMにより判定が曖昧なマージン中で適合文書（正データ）の領域に最も近い文書を提示文書とする”という新しいヒューリスティクスを提案している．そして，大規模文書データにおいて，従来の能動学習法とパフォーマンスの比較実験を行い，提案ヒューリスティクスが有効であることを示している．

第4章「サポートベクターマシンに基づく対話的文書検索の比較分析」では，文書間の類似度関数を中心として，SVMと従来の適合フィードバック手法であるRocchioアルゴリズムとの学習手続きの比較分析を行っている．その結果，SVMによる対話的文書検索における類似度である判別超平面からの符号付き距離がベクトル空間モデルのコサインに対応し，文書データのノルムのばらつきが類似度計算に大きな誤差となることを示している．この問題を解決するために，文書ベクトルを正規化するコサインカーネルを提案し，TF (Term Frequency)，TFIDF (Term Frequency Inverse Document Frequency)などの文書ベクトル表現と様々なユーザフィードバック回数で大規模文書データを対象に評価実験を行っている．その結果，特にWebなどの大規模かつ動的な環境に適しているTFにおいて，提案方法が有効であることを示している．

第5章「結論」では，本研究全体を総括し，対話的文書検索における本研究の意義をまとめている．

なお，本研究の成果として，出願者は，技術論文1編，査読付き国際会議論文2編を発表している．

論文発表後の質疑応答に引き続き，審査委員全員で博士論文の審議を行った．その結果，

本研究は対話的文書検索における文書提示に対し能動学習の新しいヒューリスティックスを提案し，またSVMと従来の適合フィードバック法との比較分析により，正規化された文書表現の妥当性に理論的基盤を与えている点が評価され，本論文は博士論文として十分な水準であると審査委員全員一致で認められた．

審査委員全員の出席のもと，公開の論文発表会を行った．まず，出願者による論文内容の発表を45分間行い，引き続き審査委員と一般傍聴者による質疑を15分間行った．出席者は，主査と4名の審査員，一般聴衆4名であった．その後，出願者と審査委員のみで，非公開の質疑応答を15分行った．

論文発表会において，出願者は，博士論文の研究についてその研究背景，研究目的，新たに提案した方法論，研究の位置づけと意義を明確に発表し，質疑応答ではその発表内容と博士論文に関する質問に対し，的確かつ簡潔に答えた．

以上の研究発表，質疑応答を鑑みて，出願者は情報学および関連する学問分野において十分な学識を有するものと審査員全員一致で判定した．また，英語による国際会議の論文執筆と口頭発表を2件行っており，

このことから出願者は十分な英語の語学力を有するものと審査員全員一致で判断した．

以上を総括し，最終試験は合格と判定された．