様式 C-19

科学研究費補助金研究成果報告書

平成２１年3月３１日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18700084
研究課題名（和文）　持続型テキスト分類学習のためのフィードバックナビゲーション
研究課題名（英文）　Feedback Navigation for Persistent Text Classification

研究代表者
岡部　正幸（MASAYUKI OKABE）
豊橋技術科学大学・情報ﾒﾃﾞｨｱ基盤ｾﾝﾀｰ・助教
研究者番号：50362330

研究成果の概要：本研究では，テキスト分類学習における訓練データの準備に要するユーザの負担を軽減するための方法の開発を目的とし，トランスダクティブ学習を利用したクエリ拡張，制約付きクラスタリングにおける制約選択，アクティブ学習を利用したスパムフィルタリングなどの研究を主に行った．各研究とも従来手法に比べより良い手法を提案することができた．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	1,800,000	0	1,800,000
２００７年度	1,000,000	0	1,000,000
２００８年度	500,000	150,000	650,000
年度			
年度			
総　計	3,300,000	150,000	3,450,000

研究分野：総合領域
科研費の分科・細目：情報学，メディア情報学・データベース
キーワード：情報検索

１．研究開始当初の背景

スパムフィルタ，Web コンテンツフィルタなどテキスト分類学習の実用的な利用を考えた場合，訓練データのラベル付けに要するユーザの負担は大きな問題であった．例えば，スパムフィルタのように新手のスパムが次々と生成されているような環境では，分類器を継続的に更新する必要があり，ユーザはスパムを一つ一つ特定して，再学習のための新しい訓練データとしてシステムに提示するというフィードバック作業を余儀なくされていた．このため，システム側で学習効果の期待できるデータを推定することでユーザの負担を軽減できるフィードバックシステムが必要とされていた．

２．研究の目的

本研究では，テキスト分類学習を継続的に行う際に生じる訓練データのラベル付けに要するユーザの負担を軽減することを目的とする．提案するシステムは人工知能分野で研究されているトランスダクティブ学習，アクティブ学習などの基礎技術を用いて構築し，新しい訓練データをユーザから提示されるのを待つのではなく，システム自らが発見し，ユーザへ積極的に確認を求め，再学習を行うというシステム誘導型のフィードバック環境を実現する．これにより，ラベル付けに要するユーザの負担を減らし，学習サイクルを無理なく持続していくことが可能となることを目的とする．

３．研究の方法

本研究では，トランスダクティブ学習とアクティブ学習について新しいアルゴリズムを提案する．前者のトランスダクティブ学習については，既存のアルゴリズムはいずれも全データに占める正解判定データの割合をパラメータとして予め与えておく必要があり，判定済みデータ数が極端に少ない場合に問題となることが分かっている．このため，正解判定データの割合を複数変えた場合の学習結果を重ね合わせることで解決を行う方法について検討する．また，後者のアクティブ学習については，仮説空間を考慮した訓練データ候補の選び方について検討する．また異なる属性を用いた学習方法の利用についても検討を行う．

４．研究成果

本研究では，トランスダクティブ学習を利用したクエリ拡張，制約付きクラスタリングにおける制約選択，アクティブ学習を利用したスパムフィルタリングなどの研究を主に行い，従来手法に比べより良い手法を提案することができた．以下に，各研究の概要について記述する．

(1)トランスダクティブ学習を利用したクエリ拡張

情報検索を行う際，クエリ（検索キーワード）を一度入力しただけで満足のいく検索結果を得られることはあまりない．多くの場合，ユーザは新たな単語をクエリに追加して結果の改善を試みるが，この追加単語の選択を自動的に行ってユーザを支援する技術は，一般にクエリ拡張（query expansion）と呼ばれる．クエリ拡張は，シンプルではあるが検索支援における効果的なアプローチであるため，これまでに多くの手法が提案されている．中でも，ユーザに一定数の文書の適合性を判定してもらい，その判定情報を利用してクエリ拡張を行う手動フィードバック型の方法は，検索性能を向上させる良いクエリを選択できることが知られている．しかし，文書の適合／非適合の判定はユーザにとって多大なコストを要するため，通常の検索場面においてユーザから質量ともに十分なフィードバック情報を得ることは極めて難しい．よって，非常に少ない判定情報からでも有効なクエリ拡張が必要となる．そこで本研究では，手動フィードバック型クエリ拡張を行うために必要な最小のユーザフィードバック，つまり，ユーザが適合文書を１つ見つけるまでの文書判定情報のみを利用した場合に有効なクエリ拡張方法を提案した．フィードバック情報が最小の場合の問題点として，追加単語の選択肢が１文書内に現れる単語だけに限られてしまうこと，また各単語のスコア計算を行う際の統計情報が信頼性に欠けることが挙げられ，実際これまでに提案されている標準的なクエリ拡張方法では，検索性能を向上させるクエリ拡張を行うことができない．我々はこうしたフィードバック情報の不足から起こる問題点を解決するため，トランスダクティブ学習と呼ばれる機械学習の方法を用いて，適合文書である可能性の高い他の文書を探し出し，これらをクエリ拡張の際のスコア計算に利用することを試みた．

トランスダクティブ学習は，transductionと呼ばれる推論方法に基づき，訓練データからラベル付け関数を生成することなしに直接テストデータのラベル予測を行う．この方法では，データ間の類似性を利用してラベル予測を行うため，テキスト文書などの類似計算ができるようなデータ集合では，訓練データ数が少なくても有効に働く．トランスダクティブ学習を実現するアルゴリズムは，これまでにいくつか提案されており様々なタスクにおいて訓練データ数が少ない場合での優位性を示している．本研究では，これらの中から高い性能を持つとされる Spectral Graph Transducer(SGT)アルゴリズムを用いた．SGT はラベル割り当てを制約付き ratiocut 問題として定式化し，その緩和問題を解くことによって近似解を導き出す．SGT によるトランスダクティブ学習では，まず各データについて，高い類似度を持つ k 個のデータを選び出し，それらのデータ間に辺を張った無向グラフを作成する．そして，このグラフを，与えられたすべての正例ラベルを含む点集合とすべての負例ラベルを含む点集合の 2 つに分割し，それぞれの集合に属するラベル無しデータに集合内のデータと同じラベルを割り振ることで学習が終了する．よって，グラフの分割方法が学習性能を左右する．SGT ではパラメータ fp を用いてグラフの分割度合いを調節することができるが，本研究ではクエリ拡張に適したパラメータ値を求める計算式を提案した．

実験を行った結果，提案手法は擬似フィードバックが苦手とする初期検索結果の悪いトピックに対する効果を発揮することが分かった．初期結果の改善手法自体はこれまでにも提案されている．例えば Mitra らは初期検索の後，上位にランクした一定数の文書を，クエリ単語の df(document frequency)値と共起度を用いた別尺度を用いて再度ランキングを行い，上位により多くの適合文書を集める方法を提案しており，この方法により，クエリ拡張によって初期検索結果を悪化させる``query drift"と呼ばれる現象を抑えることができたと報告している．方法は異なるが，一部の上位文書を再ランキングするという点で我々のアプローチと似ている．一方，Sakai らは，フィードバックに用いる文書に

ついて同じクエリ単語を含む同類の文書を過度にサンプリングせず，多様性を維持することにより，特定の単語を偏って重み付けしないようにするアプローチを提案している．このアプローチにより，従来手法では困難なトピックの性能を向上させることに成功している．我々のアプローチにおいても SGT を複数回実行することにより単語のスコア計算が特定単語に偏重しないようにしており，この点が似ている．以上のように，我々のアプローチは人手による文書判定というコストが必要ではあるが，初期検索結果が悪い場合に性能を向上させる 1 つのアプローチといえることが分かった．

(2) 制約付きクラスタリングにおける制約選択

データ間に存在する制約を利用することでクラスタリングの精度を向上させる方法は，制約付きクラスタリング，半教師ありクラスタリングなどと呼ばれ，近年精力的に研究が行われている．制約付きクラスタリングには主に 2 つのタイプが存在する一つは制約ベースのタイプで制約を満たしながら目的関数を最適化していくアプローチである．もう一つは距離ベースのタイプで制約を満たすクラスタリングを実現するために適応的にデータ間の距離関数を変化させるアプローチである．本研究では，Connectivity Kernel を制約付き k-means に組み込むことで 2 つのアプローチを同時に実現する方法を提案した．一方，制約はクラスタリングを行う前に予め分かっている場合もあるが，タスクによってはユーザとのインタラクションの過程において追加される場合もある．後者の場合，制約を与えることは一般にユーザにとって負担となる作業であるため，できるだけ精度向上の見込まれる制約を与えた方がよい．このようなアプローチは，実験計画法や能動学習において理論的に研究されているが，制約付きクラスタリングにおける制約選択方法については十分に研究されているとはいえない．本研究では，この制約選択の方法についても提案を行い，その有効性について検証した．

制約付きクラスタリングでは，一般的に must-link, cannot-link と呼ばれる 2 種類の制約を利用する．前者は必ず同じクラスタに属さなければならないデータペアとして与えられ，後者は必ず異なるクラスタに属さなければならないデータペアとして与えられる．本研究では，Wagstaff らが提案した制約付き k-means アルゴリズムに Connectivity Kernel を組み込んだ方法を提案した. ただし，オリジナルの方法はクラスタ中心をクラスタ内の重心としていたが, Kernel 計算の都合上，クラスタ中心は代表点を選択する k-medoid 型の方法を用いた．

制約を加えることでクラスタリング精度の向上が見込まれるが，より早く精度を向上させるには大きな効果の見込まれる制約を選択することが重要となる．本節では制約候補となるデータペアの選択方法について考えた. 制約候補の選択は k-近傍グラフの頂点ペアをそのユークリッド距離でソートしたリストを用いた．Connectivity Kernel はウォード法(Ward method)を用いて計算するがその際にこのリストを用いる．リストの最上位と最下位には既知の must-link と cannot-link がそれぞれリストされており，制約候補はそれらをのぞく頂点ペアから選択する．制約として効果が期待されるのは，頂点ペアの距離が短いにもかかわらず実は cannot-link である場合と頂点ペアの距離が長いにもかかわらず実は must-link である場合と考え，その可能性を構築されたクラスタ群において同一クラスタに属しているか，属していないかで判断する．つまり，上位の頂点ペアで頂点同士が同一クラスタに属していないもの，下位の頂点ペアで頂点同士が同一クラスタに属しているものを制約候補として選び実際に must-link か cannot-link なのかを判定する．上位と下位のペアは 2 つずつ同時に調べていき，該当するペアが先に見つかったものを制約候補とする．該当するペアが同時に見つかった場合は 2 つのうち一つをランダムに選択する．この方法の効果を実験で調べた結果，カーネルを用いた場合，制約選択を行った場合それぞれの効果を確認することができた．

(3) アクティブ学習を利用したスパムフィルタリング

インターネットを流れる 96.5%のメールがスパムであるというレポートがあるように，スパムメールの除去は大多数のユーザにとって面倒なルーティンワークになっている．最近のメール・クライアントの大部分はスパムフィルタ機能を持っているが，使い始めにおいては，ユーザがスパムメールをシステムに覚え込ませるという作業が必要となる．スパムメールを正確にフィルタリングするには，ユーザは十分な量のメールを判定しなければならず，判定量が不足するとそれだけスパムメールを見過ごす確率が高まってしまう．特に大量のスパムメールを受信するユーザにとってすべてのスパムメールをチェックすることは非常にコストのかかる作業であり，例えば似たスパムメールは代表的なものを知らせるだけでよいなど，必要最小限の判定のみに抑える工夫が必要である．本研究では，アクティブ学習と呼ばれる機械学習の

手法を使って，この判定すべきスパムまたはハムメールを選択することを試みた．アクティブ学習の方法にはいくつか代表的な手法がある．一つは，uncertainty sampling と呼ばれる判別関数が最も不確実な判定をするデータを判定すべきデータとして選択する方法である．これは仮説空間を減少させるタイプの手法である．もう一つは，判別の推定誤差を減少させようとする手法で，naive bayes を使った実験では uncertainty sampling よりも性能が良いとされている．

一方，フィルタリングの性能は訓練データ数だけでなく，データを表現する特徴集合にも大きく影響を受ける．一般にクラス情報を用いた特徴選択は性能が良いとされているが，この手法をアクティブ学習に適用する場合，学習の初期段階においてデータ数が不足するという問題が生じる．我々はこの問題を解決するため，推定による擬似的なクラスラベルを用いる方法を提案した．実験を行った結果，我々の提案手法が特徴選択を行いながらアクティブ学習によりフィルタリング精度を向上させていけることを確認できた．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2　件）

1. M.Okabe and S.Yamada,
"Semi-supervised Query Expansion with Minimal Feedback", IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 11, pp.1585-1589, (2007)

2. 岡部 正幸, 山田 誠二,
"トランスダクティブ学習による最小文書判定からのクエリ拡張", 人工知能学会論文誌, Vol.21, No.4, pp.398-405 (2006)

〔学会発表〕（計　4　件）

1. M.Okabe and S. Yamada,
“Interactive Spam Filtering with Active Learning and Feature Selection”, International Workshop on Intelligent Web Interaction(IWI2008)

2. M. Okabe and S. Yamada, “Spam filtering with Active Feature Identification”, in Proc. 4th International Conference on Soft Computing and Intelligent Systems (SCIS&ISIS2008), pp. 1218-1223

3. 岡部正幸, 山田誠二,“Connectivity Kernel を利用した制約付きクラスタリング”, 第 22 回人工知能学会全国大会, 2008

4. 岡部正幸, 三輪多恵子, 梅村恭二, "文字列解析に基づくネットワークトラフィックデータからの異常発見", インターネットカンファレンス, pp.67-74 (2006)

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計 0 件）

〔その他〕

６．研究組織
(1)研究代表者
岡部 正幸（豊橋技術科学大学・情報メディア基盤センター・助教）
研究者番号：50362330

(2)研究分担者

(3)連携研究者