様式 C -19

科学研究費補助金研究成果報告書

平成２２年　５月３１日現在

研究種目：若手研究（Ｂ）
研究期間：２００７　～　２００９
課題番号：１９７００２３２
研究課題名（和文）：ウェブページを対象にした主題だけではない多様な観点からの分類手法の検討
研究課題名（英文）：Exploring Automatic Non-Topical Classification

研究代表者
石田　栄美（ISHITA EMI）
駿河台大学・文化情報学部・准教授
研究者番号：５０３６４８１５

研究成果の概要（和文）：本研究では、主題カテゴリに加えて、文書タイプ、対象者などウェブページが持つ様々な属性や非主題カテゴリなど多様な観点からの分類の可能性を検討することが目的である。この目的に従い、本研究期間に 2 種の自動分類について検討した。ひとつは、ウェブ上である学術論文を自動的に判別するサーチエンジンの開発である。これは、文書タイプによる分類といえる。ウェブ上から PDF ファイルを自動的に収集し、ファイルに出現する語彙などの属性を用いて自動分類を行った。もうひとつは、非主題カテゴリとして人の価値観を表すカテゴリセットを設定し、米国の公聴会での証言の自動分類を試み、非主題カテゴリに対する分類の可能性を検討した。まず、自動分類のためのテストコレクションを作成し、自動分類実験を行った。自動分類に用いた手法は基本的な手法のみであるが、結果は将来的な可能性を示唆するものであった。

研究成果の概要（英文）：The goal of this research project is to explore the potential for automatic Web page classification based on non-topical categories (in addition to topical categories). Two kinds of classification have been explored in this project. The first was to develop a search engine to automatically detect academic articles on the Web, this was classification by document type. PDF files were collected from the Web and classified using attributes such as terms in PDF files. Second was the development and use of a new test collection for automatic labeling of sentences with ten human values. Experiment results appear promising in this preliminary study, clearly pointing to productive directions for future work.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,500,000	0	1,500,000
２００８年度	900,000	270,000	1,170,000
２００９年度	700,000	210,000	910,000
総　計	3,100,000	480,000	3,580,000

研究分野：総合領域
科研費の分科・細目：情報学、図書館情報学・人文社会情報学
キーワード：テキスト自動分類、非主題カテゴリ、ウェブページ

１．研究開始当初の背景

テキストの自動分類とは、テキストを設定したカテゴリに分類することである。従来のテキストの自動分類では、主題を表すカテゴリに分類することが主流であった。テキストを主題カテゴリに分類するには、テキスト中に用いられる語の出現傾向などを利用して、その主題的特徴をもとに分類する。テキストの自動分類研究では、主題的特徴を抽出することや分類手法の開発が主な研究対象でった。主な分類手法には、情報検索分野で考案された統計的手法を用いたものや機械学習手法を用いたものなどがあり、現在では機械学習手法のひとつである Support Vector Machine (SVM)や統計的に手法を用いた k 近傍法(kNN)などが性能が高い手法として知られている。現在、自動分類に用いられる基本的な分類手法は確立されているといってよいが、様々なカテゴリに対する分類が実用可能なレベルになったとは言い難い。今日では、テキスト自体の特徴に応じた様々なカテゴリ、レベルでの分類など、よりきめの細かい分類が求められてきている。

これまで、分類対象テキストは書誌データや新聞記事であったが、最近はウェブページを対象とした研究が多くなってきている。ウェブページの特徴のひとつは、書誌データや新聞記事等とは異なり、行政の情報や企業情報、個人のブログなど様々なタイプのページが入り混じり、テキストの形式も統一されていないことである。これらタイプの異なるページを分類において同じように扱うことは、結果的に、分類精度を下げる可能性がある。そこで、本研究では、主題カテゴリに対する自動分類だけではなく、それぞれのウェブページに適した分類が必要であると考え、ページタイプや主題以外のカテゴリに対する分類について検討した。

２．研究の目的

同じ主題を持つウェブページでも、様々な読者を対象にしたものが存在する。ウェブ上から情報を探す利用者からみれば、利用者の属性や求める情報のタイプに従って分類され、提示されるほうがよい。そのためには、主題だけではなく、各テキストがもつ特徴を考慮した様々な観点からの分類を行う必要がある。本研究では、主題カテゴリに加えて、対象者、文書タイプなどウェブページが持つ様々な属性も含めた多様な観点からの分類の可能性を検討することが目的である。主題カテゴリの他に、どのようなカテゴリを設定することが利用者にとって有効か、また設定したカテゴリに対する分類は、どの程度可能かを検討する。

本課題においては、主に 2 つの自動分類研究を行った。文書タイプを考慮したウェブ上に存在する学術論文の自動判定を行うサーチエンジンの開発と、主題以外のカテゴリである「人の価値観」を表すカテゴリを対象にしたテキストの自動分類である。

ウェブ上には様々なタイプのページが存在しており、研究者や学生がウェブ上で学術情報を検索することは一般的に行われるようになってきている。現在では Google Scholar や Scirus のように学術情報に特化したサーチエンジンが登場している。しかしながら、これらのサーチエンジンはページの収集範囲や検索アルゴリズムが公開されているわけではない。これまで、共同研究として、ウェブページの文体や構造に着目し学術情報の識別を行う手法について検討してきた。本研究では、これらの手法を用いて公開されたアルゴリズムに基づくサーチエンジン「アレセイヤ」の構築と評価を行った。このサーチエンジンは全文検索システム Lucene と組み合わせることで、検索語による結果に加えて、学術論文と判定された割合が高い順に結果を表示することができる。

2 つめの分類として非主題カテゴリを用いた自動分類を試みた。現在、非主題カテゴリを対象にした分類も行われるようになっている。この中でも意見分析、評判分析を呼ばれるものが盛んに行われている。これは、ある事柄についてポジティブかネガティブかなどの意見を判別するものである。本研究では、これまで対象とされなかった「人の価値観(Human Values)」を表すカテゴリを非主題カテゴリとして設定し、どの程度、分類が可能かを検討した。主題以外のカテゴリにテキストを分類する場合、そのカテゴリを主題カテゴリと同様に扱うことができるのか、主題以外のカテゴリ独自の問題があるのかなどが焦点になる。

以下では、それぞれの研究方法について述べる。

３．研究の方法

(1)学術論文サーチエンジンの構築

①サーチエンジンの構築

学術論文サーチエンジンでは、それぞれファイルの学術論文判定が必要になるため、まず、学術論文のテストコレクションを作成した。2005 年 5 月と 11 月に Yahoo! Japan を用いて、検索対象は PDF ファイルに限定し収集した。検索語には IPAdic の名詞辞書から

無作為抽出した 20,000 語を用いた。各検索語の上位 100 件をダウンロードし、暗号化されたファイル、破損したファイル、重複などを除いた結果、得られた PDF は 599,673 件であった。これらの集合の中から、20,000 件を無作為抽出し、6 人の判定者が各 PDF について学術論文、非論文であるかを判定した。学術論文の判定規準として、1)論文の体裁である、2)タイトル、著者名、所属機関が明記されている、3)1 論文 1 ファイルである、4)引用・参考文献がある、5)2 ページ以上である、を用いた。結果として、論文の割合は、20,000 件の集合の中で 1.63%と低いものであった。

学術論文の自動判定には、PDF 中の出現語を特徴素として用いた場合(出現語アプローチ)と、経験則による 19 のルールを用いた場合(ルールベースアプローチ)がある。分類器で用いる学習用データは、Xpdf3.01p12 を用いて PDF からテキストを抽出し、テキストを形態素解析システム MeCab と bigram を用いてトークン化した。分類器は、出現語アプローチでは SVM と AdaBoost、ルールベースのアプローチでは SVM, AdaBoost, Naïve Bayes, Decision tree, Vote を用いた。さらに、出現語アプローチでは、学習用データとして空白改行処理を行うか否かの 2 バージョン、切り分け処理が MeCab か bigram の 2 バージョン、AdaBoost の学習ラウンド数を変えたものを用いた。これらすべての手法を合わせると 16 通りとなる。検索エンジン部分には Lucene を用いた。アレセイヤでは「学術論文らしさ」により順位付けを行うため、検索結果の PDF ファイルを論文と判定した分類器数が多い順に並び変えた。

②クローリング手法を検討するための生存調査

サーチエンジンを開発する中で、テストコレクション作成は大きな問題である。テストコレクション作成のための効率的なクローリング手法を検討するために、ウェブページの生存調査を行った。約 2 年前に収集していた 584,973 件の PDF ファイル集合に対してクローラーによる生存調査を 2007 年 12 月から 2008 年 1 月にかけて行った。

(2)人の価値観を表すカテゴリを用いた分類

①テストコレクションの作成

人の価値観を表すカテゴリとして、社会科学の研究分野において広く用いられている Schwartz Values Inventory（SVI）を用いた。SVI は「Freedom」「Capable」「Equality」など 56 の基本カテゴリから構成されている。このカテゴリを用いて内容分析を行った結果、カテゴリ数が多いことや人の価値観を表すカテゴリ以外も含んでいることから、SVI をもとにして、より対象テキストに適した新しいカテゴリセットを構築した。これは、「 Effectiveness 」「 Human Welfare 」「Importance」「Independence」「Innovation」「Law and Order」「Nature」 「Personal Welfare」「Power」「Wealth」の 10 カテゴリからなる。SVI を用いた自動分類実験も行っているが、本報告では、この新しいカテゴリセットを用いた結果を示す。

対象としたテキストは、「ネットの中立」に関する米国の公聴会での 28 証言である。これらの証言に対して、新しいカテゴリセットを用いて内容分析を行い、テストコレクションを作成した。自動分類は、どのような分類器を用いるかによって、性能に大きな影響を与えるが、すでに数多くの性能の高い分類器が提案されており、本実験ではそれらの分類器を用いることにした。本研究で注目したのは、主題カテゴリで用いられている分類器を用いて非主題カテゴリに対する自動分類が、どの程度可能であるのかを検証することである。

テストコレクションは一つのデータに対して複数カテゴリが付与されていたため、分類器を効果的に学習させるための学習用データの表現方法と、分類器が出力したカテゴリの中からいくつのカテゴリを最終的な分類結果とするかという分類先の決定手法についても検討した。

②学習用データの表現方法

テストコレクションには複数のカテゴリが付与されているものがある。分類器は、カテゴリが付与されているデータを用いて学習することが必要であるが、データに複数のカテゴリが付与されている場合、学習用データをどのように表現すれば最も効果的であるかを検討しなければならない。学習用データの表現方法は、以下の 2 つの手法(Train1、Train2)を実験した。

Train1 は、複数カテゴリが付与されている場合でも、一つのカテゴリが付与されていた場合と同様に、各データを複製し、各々のカテゴリの学習用データとするという方法である。同じデータを複数のカテゴリに学習させてしまう可能性はあるが、分類性能が高い分類器を用いる場合は、効果的であるといえる。この手法は、すべてのデータを用いることができるため、データ量に制限がある場合には、有効であると言われている。Train1 は、偏りがある学習用データの場合には、出現回数が多いカテゴリに対して過学習してしまう恐れがある。

一方、複数カテゴリが付与されている場合、その中から適切な一つのカテゴリを選択し、そのカテゴリのみの学習用データとするという方法も提案されている。本実験では、複数のカテゴリが付与されている場合、付与さ

れているカテゴリの中でもっとも出現回数が少ないカテゴリを選択するという方法を用い、これを Train2 とした。この手法を用いると、学習用データが小さくなってしまうという問題があるが、その一方でデータの偏りの問題が解消されると予想できる。その他に、複数カテゴリの組み合わせを一つのカテゴリとみなす方法、一つのカテゴリだけが付与されているデータだけを学習用データとして用いる方法、カテゴリごとにデータを正と負のサンプルに分けて学習する方法が提案されている。最初の 2 手法については、プレ実験で Tran1, 2 によりも性能が低かったため、本実験は行わなかった。3 番目の手法は、他の分類器を用いて、今後、実験する予定である。

③分類手法

分類器として性能が良いと報告されている kNN(k=1,3,5,10,15,...,40,)を用いた。本実験では、プレ実験の結果から、データに対し、ポーターのアルゴリズムを用いて語幹処理を行い、4 回以下しか出現しない語は削除した。Weka は、サンプルの近さをはかる尺度として 1)同じ重み(vote)、2)逆距離加重法(w=1/distance, iw)、3)類似性加重法(w=1-distance, sw)の 3 つの重みづけ手法を提供している。本実験ではこのうち vote と iw を用いた。また、Weka が提供している kNN の分類器は、評価用データに各々のカテゴリに対する確率分布を付与して、分類結果を出力することができる。この確率分布を用いて最終的なカテゴリを決定する方法を実験した。

実世界では、各データにいくつのカテゴリを付与することが適当であるかはわからない。そこで、閾値を用いて、選択するカテゴリ数を推定する方法(Threshold)を実験した。kNN の確率分布の値が閾値を以上であれば、最終的な分類結果のカテゴリとするという方法である。この場合、適切な確率分布の値を設定することが必要である。

まず、テストコレクションを、学習用データ、閾値用データ、評価用データと 3 分割した。それぞれの割合は、80%、10%、10%である。次に以下の手順で閾値を設定した。

1)学習用データを用いて分類器を学習させる
2)分類器に、閾値用データを入力し、各カテゴリの確率分布を得る
3)閾値用のデータの正解と分類結果を比較し、閾値用データにおいて、もっとも F 値が高くなる閾値を選択する
4)評価用データに対して 3)で得られた閾値を用いて、カテゴリを選択する

この閾値を用いた手法(Threshold)と比較するため、評価用データに正解カテゴリとして付与されているカテゴリと同じ数のカテゴリを選択する方法(Oracle)も実験した。分類結果のカテゴリは以下のように選択した。あるデータ s に i 個のカテゴリが付与されている場合、kNN の確率分布をもとに上位 i 番目までのカテゴリを、s に対する最終的な分類結果とした。i 番目のカテゴリと同じ確率分布を持っているカテゴリがあった場合は、それらも選択した。

４．研究成果

(1)学術論文サーチエンジン

①サーチエンジンの評価

サーチエンジンの構築後、専門用語を用いた評価実験を行った。東京大学、京都大学、慶応義塾大学の博士論文データベースにある 2005 年度博士論文のタイトルを収集し、テクニカルターム 180 語を選択した。これらを検索語として人手で検索を行い、学術論文が含まれる割合を調べた。検索結果が得られたのは 180 語中 126 語であった。各検索語を用いて表示された検索結果の上位 20 以内のファイルを確認したところ、学術論文であった割合は 0.475 だった。

次に、一般的な利用を想定した評価を行うために、検索結果の上位 10 位に含まれる学術論文の割合を調べ、他の学術情報専門の検索エンジン Google scholar BETA と Scirus の検索結果と比較した。180 の検索語で検索したとき、検索結果が 0 件とならない割合は、アレセイヤが 140/180、Google scholar が 168/180、Scirus が 37/180 であった。ただし、Scirus はあらかじめ検索語を分割すれば検索結果が改善する。検索結果は、Google のほうが多い。検索結果に学術論文が含まれる割合は、それぞれ 0.406, 0.291, 0.194 であり、アレセイヤが最も高かった。全文へのアクセスは、アレセイヤはすべての論文にアクセスできたのに対し、Google scholar はアクセス不可、リンク切れ、リンクなしなどの理由で 636 件、Scirus は 36 件にアクセスできなかった。

②PDF ファイルの生存調査

PDF ファイルの生存率は 55.2%であり、一般的なウェブページとほぼ同様の結果であることがわかった。さらに、2 年前に収集した時点での URL で保存できなかったファイルに対して人手による追跡調査を行った。PDF ファイルを論文と非論文に分けて調査したが、保存できなかった論文ファイル全てがウェブ上から消滅しているわけではなく、半数以上の 55%について移動先 URL を再発見することができた。非論文では 24%であった。この論文の PDF ファイルは URL が変更されたとしても生存している割合が高いことがわかった。これらの結果から、ウェブ上の情報の中

でも、内容によってその生存状況が異なることといえる。今後は、これらの結果を考慮したクローリング手法を検討する必要がある。

③サーチエンジンの今後

本研究では、学術論文に特化したサーチエンジン「アレセイヤ」の開発と評価を行った。他のサーチエンジンと比較すると検索結果が少ないという問題はあるが、検索結果に含まれる学術情報の割合が高いこと、学術情報の本文を実際に入手できることが保証されている点で有用であることが明らかになった。今後は、判定の精度を向上させていくこと、収録対象の範囲を広げていくことが課題である。

(2)人の価値観を表すカテゴリを用いた分類

①学習用データの表現方法、分類手法の結果

ネットの中立に関する米国の公聴会での28証言(2,294文)に対して内容分析を行った結果、新しいカテゴリセット中のいずれかのカテゴリが付与されたのは2,005文であった。これらのデータを用いて10点交差検定を用いた実験結果を下表に示した。この実験のために、閾値は0.01から0.25まで0.01刻みで評価した。最も性能が高いのは、0.48(k=25、vote)である。Train1とTrain2を比較すると、Train1の手法がわずかながらもよい性能を示している。閾値を用いた手法では0.45(k=25、vote)であり、これは正解カテゴリ数が最初からわかっている手法(0racle)を用いた場合と比べても、遜色ない結果が得られている。よって、閾値を用いた手法でも、正解のカテゴリ数を推定することができるといえる。

	0racle		0racle		Threshold	
	Train 1		Train 2		Train 1	
k	vote	iw	vote	iw	vote	iw
15	0.46	0.46	0.43	0.43	0.43	0.44
20	0.48	0.48	0.45	0.44	0.44	0.44
25	**0.48**	0.47	**0.46**	0.45	**0.45**	0.45
30	0.48	0.47	0.45	0.45	0.45	0.45
35	0.47	0.47	0.45	0.45	0.45	0.45

②今後の課題

これらの結果から、従来の分類手法を用いても、ある程度の分類は可能であることは明らかになった。しかしながら、性能は実用化のレベルに達しておらず、更なる改良が必要である。今後は、テストコレクションを増やすこと、また、分類器に入力するにはどのようなテキストの特徴がよいか検討する予定である。文単位での分類を行っているため、一文に含まれる単語の量は少ない。そのため、文に含まれる単語だけでなく、文脈を考慮した手法を取り入れることなどを考える必要がある。また、文章の中では、同じカテゴリが近くに出現するなどの傾向が見られるため、カテゴリの出現位置の情報を取り入れることも考えられる。

主題カテゴリを対象とした分類と比べて、人の価値観を表すカテゴリは、語とカテゴリの関係がそれほど強くないといえる。今後は、語以外の情報を用いた非主題カテゴリへの分類も検討していく。

５．主な発表論文等

〔雑誌論文〕（計４件）

① 石田栄美、An-Shou Cheng、Douglas W. Oard、Kenneth R. Fleischmann、人の価値観を表すカテゴリを対象にした複数カテゴリへの自動分類の試み、文化情報学：駿河台大学文化情報学部紀要、査読有、Vol.16、No2、2009、p.53-68

② Emi Ishita、Non-topical Classification for Healthcare Information (http://www.ieee-tcdl.org/Bulletin/current/Ishita/ishita.html)、Bulletin of IEEE Technical Committee on Digital Libraries、査読無、Vol.5、No.3、2009

③ Kazuaki Kishida、Emi Ishita、Translation disambiguation for cross-language information retrieval using context-based translation probability、Journal of Information Science、査読有、Vol.35、No.4、2009、p.481-495

④ 三根慎二、汐﨑順子、國本千裕、石田栄美、倉田敬子、上田修一、眼球運動から見た子どもの絵本の読み方、Library and Information Science、査読有、No.58、2007、p.69-90

〔学会発表〕（計１０件）

① Emi Ishita、Shinji Mine、Chihiro Kunimoto、Junko Shiozaki、Keiko Kurata、Shuichi Ueda、Analyzing Viewing Patterns While Reading Picture Books、ACM/IEEE Joint Conference on Digital Libraries (JCDL10) and the annual International Conference on Asia-Pacific Digital Libraries (ICADL10)、poster presentation、査読有、2010.06、Gold Coast、Australia (to appear)

② Emi Ishita、Teru Agata、Atsushi Ikeuchi、Nozue Michiko、Miyata Yosuke、Shuichi Ueda、A Search Engine for Japanese Academic Papers、ACM/IEEE Joint

Conference on Digital Libraries (JCDL10) and the annual International Conference on Asia-Pacific Digital Libraries (ICADL10) 、 poster presentation、査読有、2010.06、Gold Coast、Australia (to appear)

③ An-Shou Cheng、Kenneth R. Fleischmann、Ping Wang、Emi Ishita、Douglas W. Oard、Values of Stakeholders in the Net Neutrality Debate: Applying Content Analysis to Telecommunications Policy、43rd Hawai'i International Conference on System Sciences、査読有、2010.01、Kauai, Hawaii, USA

④ Kenneth R. Fleischmann、Douglas W. Oard、An-Shou Cheng、Ping Wang、Emi Ishita、Automatic Classification of Human Values: Applying Computational Thinking to Information Ethics、Annual Conference of the Association for Information Science and Technology、poster presentation、査読有、2009.11、Vancouver, Canada

⑤ 石田栄美、An-Shou Cheng、Kenneth R. Fleischmann、Douglas W. Oard、人の価値観を表すカテゴリを対象にした自動分類、第57回日本図書館情報学会研究大会発表要綱、査読無、2009.10.31-11.01、p.33-36、於明治大学

⑥ Emi Ishita、Shinji Mine、Masanori Koizumi、Yosuke Miyata、Chihiro Kunimoto、Junko Shiozaki、Keiko Kurata、Shuichi Ueda、Analyzing OPAC use with screen views and eye tracking、ACM/IEEE Joint Conference on Digital Libraries (JCDL09)、poster presentation、査読有、2009.06、p.405、Texas、USA

⑦ 石田栄美、宮田洋輔、池内淳、安形輝、野末道子、上田修一、生存分析からみた学術論文 PDF ファイルのクローリング、2008年日本図書館情報学会春季研究集会発表要綱、査読無、2008.03、p. 67-70、於東京大学本郷キャンパス

⑧ 三根慎二、小泉公乃、宮田洋輔、國本千裕、汐﨑順子、石田栄美、倉田敬子、上田修一、画面遷移と利用者特性からみた大学生における OPAC の閲覧、2007 年度三田図書館・情報学会研究大会発表論文集、査読無、2007.11、p. 45-48、於慶應義塾大学三田キャンパス

⑨ 石田栄美、三根慎二、小泉公乃、宮田洋輔、國本千裕、汐﨑順子、倉田敬子、上田修一、大学生は OPAC をどのように見ているのか、第55回日本図書館情報学会研究大会発表要綱、査読無、2007.10、p.101-104、於鶴見大学

⑩ 安形輝、池内淳、石田栄美、野末道子、宮田洋輔、上田修一、学術論文 PDF 検索システムの開発と評価、第55回日本図書館情報学会研究大会発表要綱、査読無、2007.10、p.57-60、於鶴見大学

６．研究組織

(1) 研究代表者

石田　栄美（ISHITA　EMI）
駿河台大学・文化情報学部・准教授
研究者番号：５０３６４８１５

(2)研究分担者

なし

(3)連携研究者

なし