社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ＦＩ－86 (5)
2007－ＤＤ－60 (5)
2007／3／27

プーリング手法を用いた学術論文の自動判別実験

池内 淳
大東文化大学

安形 輝
亜細亜大学

石田栄美
駿河台大学

野末道子
鉄道総合技術研究所

宮田洋輔
慶應義塾大学大学院

上田修一
慶應義塾大学

機械学習におけるテキスト分類実験を行うためには，予め判定されたテストコレクションを必要とする．しかしながら，ラベル付きデータの作成については，その多大なコストの問題がしばしば指摘されてきた．本研究では，ウェブから日本語学術論文 PDF ファイルを自動的に判別・収集することを目的として，20,000 件のラベル付きデータを学習集合とし，およそ 52 万件のラベルなしデータを実験集合とした自動判別実験を行った．また，複数の分類アルゴリズムによって学術論文であると判定されたファイルをプーリングすることによって，各々の手法の性能比較を行った．その結果，本実験環境におけるプーリング手法の有効性が示された．

Automatic Detection for Academic Articles Using Pooling Method

Atsushi IKEUCHI
Daito Bunka University

Teru AGATA
Asia University

Emi ISHIDA
Surugadai University

Michiko NOZUE
Railway Technical Research Institute

Yosuke MIYATA
Keio University

Shuichi UEDA
Keio University

In machine learning study, we need to prepare test collections for conducting text categorization experiments. However, it has been frequently pointed out that constructing labeled data set is expensive and / or time-consuming. The purpose of this study is automatically identifying and collecting academic articles in Japanese PDF files on the Web. Then, we conducted the automatic detecting experiment using pooling method and compared the performance of various classifiers. Results confirmed applicability and usefulness of pooling method in this experimental environment.

1. はじめに

インターネットの普及に伴い，学術情報流通環境は大きな変貌を遂げている．その第一波は，おそらく，一次資料のデジタル化であったと言えるだろう．ネットワーク帯域の拡充とデジタル環境の普及によって，電子図書館への期待が高まり，従来，冊子体のみで閲覧可能であった学術文献がデジタル化され，オンラインで利用されるようになった．また，それに伴って，ジャーナルの購読料の高騰といった問題が指摘されるようにもなった[1]．

そして近年では，それらデジタル化された一次資料のオープンアクセス化が急速に進展している．オープンアクセスの定義は様々であるが，その主な実現方法として，(a) オープンアクセスジャーナル，(b) 機関リポジトリ，(c) 研究者等によるセルフアーカイビング等が挙げられる．

オープンアクセスは，一次資料を利用するための金銭的・時間的コストを大幅に削減し，学術情報の主要な生産者であり，かつまた，消費者でもある研究者らに大きなメリットをもたらしている．その一方で，ジャーナルの購読料を運営の資金源としてきた出版社や学協会は必ずしも順境にあるとは言えない．いずれにせよ，オンラインでアクセス可能なフリーの学術文献の数が増加傾向にあることは明らかである．

こうした背景を受けて，Google Scholar* や Windows Live Academic Search**，あるいは，Elsevier Science 社による Scirus*** といった学術文献に特化されたサーチエンジンが公開されるようになっている．これらは，必ずしも，フリーの学術文献のみを収集対象としている訳ではないが，膨大な数の学術文献を索引化しており，(a) 引用索引，(b) リンクリ

* http://scholar.google.com/
** http://academic.live.com/
*** http://www.scirus.com/

ゾルバ，(c) 図書館や出版社との連携といった多様な機能を備えることによって，利用者の裾野を拡大している．

こうした学術文献サーチエンジンのもう一つの主要なメリットは，ウェブの検索においてしばしば惹起される検索ノイズの問題に悩まされることなく，予め一定水準のクオリティコントロールがなされたコンテンツのみを検索対象とすることが可能な点にあると言えるだろう．近年，ウェブマイニングにおいて，特定のコンテンツ（書き込み，データ，ページ，ファイル，サイト等）を自動的に発見する，または，フィルタリングするといった研究が盛んに行われているが，それらの中でも，学術文献の自動検出は，単に技術的側面だけではなく，社会的有用性という観点からも関心が高いと言えるだろう．

しかしながら，サーチエンジンの運営主体が基本的には営利企業であるためか，収録対象等について一定の情報公開はなされているものの，そのクローリングの範囲やランキングアルゴリズムの詳細については公開されていないといった点が懸念されている[2]．また，こうしたシステムを構築するために，どこまでが自動化可能であり，どの部分を手動で行わなければならないのかなど関心は尽きない．

2. 研究目的

筆者らの目的は，広大なウェブ空間の中に存在する膨大なコンテンツの中から，日本語学術論文ファイルのみを自動的に検出し収集することにある．本章では，筆者らがこれまで行ってきた一連の研究を概観するとともに，本稿における問題意識と研究目的について明らかにする．

2.1 学術論文の計量文体学的特性

ウェブ上において，フリーでアクセスすることが可能な学術論文の数が増加傾向にあることは疑いないものの，ウェブコンテンツ全体に占める比率は極めて少ないことが予想される．それら大半の学術論文以外のものと，僅かな学術論文を判別するための手がかりとして，第一に，その計量文体学的特性に着目した[3]．

具体的には，「論文（11 分野，オープンアクセス可）」，「新聞記事（毎日新聞）」，「日記・blog（はてなダイアリー）」の三つのカテゴリから抽出した記事について，(a) 一記事の文数，(b) 一文の長さ，(c) 一記事の単語数，(d) 文字種の比率，(e) 品詞の比率，(f) 文頭表現の頻度，(g) 文末表現の頻度，(h) 接続詞の頻度といった八つの観点から比較を行った．

その結果，(f) 文頭表現の頻度については，「本研究では」，「その結果」，(g) 文末表現の頻度については「を行った」，「が分かる」，(h) 接続詞の頻度については「すなわち」，「あるいは」等が，新聞や日記と比較して，論文に特有で，かつまた，頻繁に用いられる文体的特性であることが明らかになった．

2.2 ベイジアンファイルタによる非論文の検出

第二段階では，実際に，論文と非論文の自動判定実験を行った[4]．まず，サーチエンジンを用いて収集した約 25 万件の日本語 PDF ファイルの中から，12,000 件を無作為抽出し，それら全てについて，予め，「論文」か「非論文」かを人手によって判定し，ラベル付けを行った．その結果，論文は 345 件（2.9%），非論文は 11,655 件（97.1%）であった．

次に，PDF ファイルからテキストを抽出し，形態素解析，及び，bigram の二通りの手法によってトークン化を行った．また，ここでは，ベイジアンフィルタ（Gary Robinson-Fisher 法）を用いて分類を行った．ベイジアンフィルタは，ベイズの定理を応用した分類器であり，近年，スパムメールのフィルタリングツールに応用されたことから，一般にも広く知られるようになっている．

実験の際には，12,000 件を 3,000 件ずつグループ化して，4 交差検定を行った．ベイジアンフィルタにおいては，スパム確率（ここでは非論文確率）をパラメータとして，既定しておかなければならないが，ここでは，0.6 と 0.9 の二パターンを試行した．その結果，「bigram」＋「0.6」の組合せのとき，98.6%と最も高い再現率を示した．

2.3 複数の手法を用いた学術論文の判定性能比較

第三段階では，論文と非論文の自動判定を行う際に，より優れた分類器は何であるかを明らかにするために，複数の手法を用いた実験を行い，それらの性能を比較することを試みた[5,6]．

まず，サーチエンジンを用いて，約 35 万件の日本語 PDF ファイルを収集し，前回実験の際に用いた 25 万件とマージして，約 54 万件の PDF ファイル集合を得た．次に，そこから，20,000 件を無作為抽出し，やはり，予め，人手によって判定作業を行った．但し，前回実験において，「論文」と「非論文」との判別がつき難いものや，論文ではないが学術的観点から有益な資料も存在したことから，「論文」，「準論文」，「非論文」の三つのカテゴリを設けた．その結果，論文は 326 件（1.6%），準論文は 624 件（3.1%），非論文は 19,050 件（95.3%）であった．

また，この実験では，前回と同じ出現語を用いるアプローチだけではなく，新たに，「ルールベース」と呼ばれるアプローチを採用した．出現語を用いた場合，実験集合が大きくなるにしたがって，当然ながら素性

数と計算量が膨大になることから，一定の限界の存在が想定される．殊に，ウェブへの応用を視野に入れた場合，その問題は看過できない．

それに対して，ルールベースでは，ヒューリスティックに導かれた19の属性（→表1．ルールベースにおける判定属性）のみによって判定を行うことから，コスト面でより有利なアプローチであると言える．また，テキスト情報だけではなく，ファイルの形態的特性も勘案することができる．ちなみに，これら19の属性を導くに当たっては，様々な属性を加除しながら，プレテストを繰り返し，最も性能の良かった属性の組合せを採用している．

表1．ルールベースにおける判定属性

カテゴリ	属性
構造	ファイルサイズ
	ページ数
	ページの形
入手元	URLがac.jpであるか
	URLがgo.jpであるか
文体	文体が「である」調か「ですます」調か
	会話が出てくるか（文末に「ね．」「」が使われているか）
	ひらがなが出現するか（外国語か）
出現語彙	「研究」
	「文献」
	「被験者」
	「調査」「分析」「実験」
	「紀要」「研究報告」「研究ノート」
	「図」「表」
	「本稿」「本研究」「本論文」
	「研究成果」「研究結果」
	「考察」「考慮」
	「引用文献」「参照文献」「参考文献」
	「大学」「研究所」「研究センター」

次に，PDFファイルからテキストを抽出し形態素解析，及び，bigramの二通りの手法によってトークン化を行った．その際，改行・スペースを除去しないものと，改行・スペースを除去するものの二つのバージョンを作成した．実験に際しては，出現語アプローチについては，(a) SVM (Support Vector Machine)，(b) AdaBoost，(c) ベイジアンフィルタの三種類，ルールベースについては，(a) SVM，(b) AdaBoost，(c) ナイーブベイズ，(d) 決定木 (Decision Tree: C4.5)，(e) メタ判定機であるVoteの五種類を用いるとともに，20,000件を5,000件ずつグループ化して，4交差検定を行うことによって，判定性能の比較を行った．その結果，精度ではSVMが最も優れており，再現率ではナイーブベイズの最も優れていることが明らかになった．

2.4 大規模データ集合を対象とした評価実験

2.4.1 機械学習における準教師付き学習

さて，機械学習における自動分類実験を行うためには，予め分類されたテストコレクションを必要とするが，その作業は概ね人手によってなされなければならない．教師付き学習 (Supervised Learning) におけるテキスト分類 (Text Categorization) のテストコレクションとしては，Reuters-21578[7]が代表的であり，しばしば分類性能のベンチマーキングに用いられている．

しかしながら，こうしたテストコレクションの作成には多大なコストを要することが指摘されてきた．その一方で，分類されていないデジタル化されたデータを入手することは極めて容易である．そこで，近年，少量のラベル付きデータと大量のラベルなしデータを併用して，より高度な分類器を構築しようとする準教師付き学習 (Semi-Supervised Learning) に関する研究が多くなされるようになっている[8]．例えば，教師付き学習のための分類器と，欠損値の最尤推定を行うEMアルゴリズムとを組合せるといったアプローチが採られている[9]．

2.4.2 情報検索におけるプーリング手法

一方，情報検索研究においては，評価対象となるデータセットが大規模であり，予め，全ての文献について，適合／非適合の判定を行うことが困難である場合，プーリング手法が用いられている．すなわち，特定の検索課題に対して，複数の検索システムが，予め既定された上位何件かまでに出力した文献をプーリングしておき，それらの集合についてのみ，実際に，人手による判定作業を行うというものである．こうしたアイディアは1970年代に，既に提案されていたものであるが[10]，その実用性と有効性が認知されるようになったのは，おそらく，1990年代以降，TREC*やNTCIR**といった情報検索コンテストの評価において用いられるようになってからであると推察される．

また，プーリング手法による大規模テストコレクションの評価が，コストの面から効率的であることは明かであるが，プーリングされなかった適合文献の存在といった問題が指摘される．この点に関して，Kuriyamaら[11]は，NTCIR-1におけるテストコレクションを対象とした検証調査を行っており，プーリン

* http://trec.nist.gov/
** http://research.nii.ac.jp/ntcir/

グによる適合文献の網羅性と，評価実験の信頼性の高さを明らかにしている．

2.4.3 本研究におけるアプローチ

冒頭で述べたように，筆者らの目的は，ウェブ上にある日本語学術論文ファイルを自動収集することにある．これを機械学習における二値分類の問題として捉えたとき，これまで行ってきた実験環境の範疇においては，一定の有効性を示すことができた．しかしながら，現実の評価集合となるウェブ上のコンテンツは，概ね無限母集団とみなすことができるため，学習集合には出現しない膨大な数の未知の素性が含まれることは言うまでもない．

本稿では，プーリング手法を用いて，論文／非論文のラベル付けがなされた 20,000 件の学習集合を用いて構築された分類器から，約 52 万件のラベルなし集合の判別実験を行う．ここでの目的は，上述のような環境においてもなお，各々の分類手法が，これまでと同等の判定性能を示すことができるのかを確認することであり，かつまた，大規模なテキスト分類実験におけるプーリング手法の適用可能性を検証することである．

3. 実験集合の作成

3.1 日本語 PDF ファイルの収集

ウェブ上において，学術論文の全文を公開するためのファイル形式としては，(a) プレーンテキスト，(b) HTML，(c) PDF，(d) PostScript，(e) TeX，(f) MS Word（互換形式）など様々である．ここで，三根による調査[12]によれば，2006 年 5 月時点で，オープンアクセス論文に占める PDF ファイルの比率は 80.9%に達しており，現時点での標準的なフォーマットであるとみなしてよいだろう．一方，HTML ファイルの数は他のフォーマットと比較して（プレーンテキストそのものを除けば），テキスト情報の抽出も容易であるといったメリットが存在する．しかしながら，上述の調査では，HTML ファイルの比率は 5.1%と低く，かつまた，1 論文が 1 ファイルで構成されていない例の多いことが予想される．以上の観点から，PDF ファイルのみを収集対象とした．

実際の PDF ファイル群の収集は，2005 年 5 月と半年後の 2005 年 11 月との二度に亘って行った．ここでは，クローリングではなく，サーチエンジンを利用することとした．まず，ipadic2.5.1*の六つの名詞辞書ファイル（計 213,020 語）から，それぞれ，9,750 語（第一回目），10,250 語（第二回目）を無作為抽出し，それらのキーワードとして，Yahoo!で検索を行い，URL を収集した．その際，言語を「日本語」，ファイル形式を「PDF ファイル」に限定するとともに，出現頻度の高い特定の語彙が含まれるファイルに偏った収集となることを避けるために，キーワードごとの URL の最大収集件数を上位 100 件までとした．

出力結果の重複除去後の異なり URL 件数は，それぞれ，307,514 件（第一回目），441,598 件（第二回目）となった．さらに，各々の URL から PDF ファイルのダウンロードを試みた．そして，(a) ダウンロード不可能であったもの，(b) 0 バイトファイル，(c) 破損ファイル，(d) 暗号化ファイル，(e) 拡張子が.pdf でるにも拘わらず PDF ファイルではなかったもの等を除去した結果，それぞれ，248,314 件（第一回目），349,971 件（第二回目）の集合が得られた．さらに，二つの集合を重複除去した結果，544,096 件の日本語 PDF ファイルが得られた．

3.2 学術論文と非論文の判定作業

次に，PDF ファイル集合全体から 20,000 件を無作為抽出し，6 人の判定者が，各々について，「学術論文」と「非論文」の判定を行った．学術論文の定義は必ずしも明確ではないが，ここでは，(a) 論文の形態をとっている (b) タイトル・著者名が銘記されている（所属機関，抄録等のあることが望ましい），(c) 引用文献や参考文献がある，(d) 1 論文が 1 ファイルで構成されている，(e) 2 ページ以上である，とした．したがって，学術論文の一部や 1 ファイルに複数の学術論文が含まれる場合は「非論文」と判定される．また，サーチエンジンでは，日本語のみを検索対象としたものの，一部，英語や中国語の論文が混入していたが，これらも「非論文」となる．

ここまでは，「2.3 複数の手法を用いた学術論文の判定性能比較」で触れた実験の手続きと重複しているが，この後，再判定を行ったところ，学術論文と認められるものは 371 件（1.9%）となった．なお，今回の実験では，「準論文」カテゴリは設けていない．

3.3 テキスト抽出とトークン化

3.3.1 PDF ファイルからのテキスト抽出

判別実験に用いる分類器は，PDF ファイルを直接扱うことができないため，PDF ファイルからテキストデータを抽出しなければならない．そのためのツールとしては様々なものが存在するが，自動化の可否，あるいは，日本語を扱えるといった観点から，Xpdf3.01p12**を用いることとした．

PDF ファイルは表示・印刷時にレイアウトの再現可

* http://chasen.naist.jp/stable/ipadic/

** http://www.foolabs.com/xpdf/

表2. 学習集合と評価集合のトークン数の比較とその一致率

実験集合	空白・改行	トークン化	トークン数			D.合計	A／B	C／D
			A. 学習用のみ	B. 評価用のみ	C. 両方に出現	(A+B+C)		
20,000件	未処理	形態素	391,189	113,766	167,551	672,505	343.85%	24.91%
		bigram	580,098	146,998	490,736	1,217,832	394.63%	40.30%
	処理済	形態素	460,983	136,226	173,788	770,996	338.40%	22.54%
		bigram	806,651	202,223	681,235	1,690,109	398.89%	40.31%
544,096件	未処理	形態素	189,178	7,007,562	483,673	7,680,413	2.70%	6.30%
		bigram	60,590	3,106,734	1,157,242	4,324,566	1.95%	26.76%
	処理済	形態素	278,091	9,458,376	493,251	10,229,718	2.94%	4.82%
		bigram	82,164	4,682,888	1,760,213	6,525,265	1.75%	26.98%

能なデータ形式であり，内部的には文書構造の情報をも保持することが可能である．しかしながら，多くのPDFファイルは単にレイアウト情報しか持たない．そのため，テキストデータの抽出を行うと，Xpdfはレイアウトの指定がなされている箇所を改行・空白へと変換することが多い．したがって，オリジナルのPDFファイルの作成者によって意図された改行・空白と，レイアウトの指定が，Xpdfによって変換された結果としての改行・空白とを識別することは不可能に近いと言える．

そこで，出現語を属性として用いる場合は，(a) 改行・空白の除去といった後処理を行わないものと，(b) 改行・空白を，英数字の前後では空白1文字に変換し，英数字の前後以外の箇所では除去するものとの二つのバージョンを作成した．ちなみに，英数字の前後で空白を1文字に変換したのは，英単語の連結を防ぐためである．

3.3.2 抽出されたテキストのトークン化

日本語は単語間を容易に識別する空白等のデリミタを持たない言語であるから，分類実験を行う前に，テキストデータをトークンに分割しなければならない．ここでは，形態素解析とbigramの二つのアプローチを採用した．なお，形態素解析には，MeCab0.81を*用いた．

表2は，ラベル付きの20,000件の実験集合と，544,096件の実験集合について，それぞれ，「A. 学習集合のみに出現するトークン数」，「B. 評価集合のみに出現するトークン数」，「C. 学習集合と評価集合の両方に出現するトークン数」，「D. 合計トークン数」，並びに，「学習用のみと評価用集合のトークン数の比率 (A／B)」，「全体に占める一致トークン数の比率 (C／D)」を示したものである (→表2. 学習集合と評価集合のトークン数の比較とその一致率)．

ここで，20,000件の実験集合におけるトークン数は，4交差検定のために分割した四つのグループの平均値を示している点に留意されたい．また，544,096件の実験集合における学習集合とは，ラベル付きの20,000件のデータのことであり，実験集合とは，残りのラベルなしの524,096件のデータのことである．

20,000件の実験環境と，544,096件の実験環境との相違は，この表から明らかであろう．すなわち，表中のA，B，Cのトークン数を比較したとき，20,000件では，いずれも「B. 評価用のみ」が最も少なく，「A. 学習用のみ」とは3倍～4倍の差がある．一方，大量のラベルなしデータを含む544,096件では，逆に，「B. 評価用のみ」が最も多く，「A. 学習用のみ」に対して，1.75%～2.94%の比率を占めるに過ぎない (この逆数を算出すると，34倍～57倍の開きがあることが分かる)．これは，評価集合の中に，膨大な数の学習集合にとって未知のトークンが含まれているということを意味している．

4. プーリング手法の適用可能性の検討

4.1 予備実験

「2.3 複数の手法を用いた学術論文の判定性能比較」で触れたように，既に，20,000件のラベル付きデータを対象とした学術論文の自動判別実験は行っているものの，その後，人手による学術論文の再判定を行い，実験集合に変化が生じたこと，並びに，プーリング手法の適用可能性を確認するために，改めて，同様の実験を行った．

ここでも，上述の「出現語アプローチ」と「ルールベースアプローチ」を採用し，出現語アプローチについては，(a) SVM，(b) AdaBoostの二種類，ルールベースについては，(a) SVM，(b) AdaBoost，(c) ナイーブベイズ，(d) 決定木 (C4.5)，(e) Voteの五種類を用いるとともに，20,000件を5,000件ずつグループ化して，4交差検定を行った．

出現語アプローチのうち，SVMについては，JoachimsによるSVM^{light}6.01**を用いた (線型カーネ

* http://chasen.org/~taku/software/mecab

** http://svmlight.joachims.org/

表3. 20,000件の実験集合における学術論文の判定性能の比較(再判定)

アプローチ	手法	改行・空白		精度	再現率	F値	判定論文数	正解論文数
出現語	SVM	未処理	形態素	.792	.420	.549	197	156
			bigram	.805	.434	.564	200	161
		処理済	形態素	.799	.450	**.576**	209	167
			bigram	**.811**	.439	.570	201	163
	AdaBoost	未処理	Round10	.508	.447	.476	327	166
			Round100	.615	.477	.537	288	177
			Round1000	.651	.442	.526	252	164
		処理済	Round10	.577	.518	.545	333	192
			Round100	.572	.472	.517	306	175
			Round1000	.674	.423	.520	233	157
ルール	AdaBoost		Round10	.484	.480	.482	368	178
			Round100	.537	.469	.501	324	174
			Round1000	.559	.445	.495	295	165
	決定木(C4.5)			.541	.340	.417	233	126
	ナイーブベイズ			.259	**.892**	.402	1276	331
	Vote			.506	.596	.547	437	221

ル関数を使用). また, AdaBoost については, Schapire らによる BoosTexter[13]の AdaBoost.MH を用いた. なお, ブースティング (Boosting) は, バッギング (Bagging) と同様に, 複数の判定器 (弱学習器) を組み合わせる集団学習 (Ensemble Learning) と呼ばれる枠組みを用いた機械学習手法の一つであり, BoosTexter では, 弱学習器として, 決定木アルゴリズムの一種である決定株 (Decision Stumps) を採用している. また, ここでは, 学習の繰り返しラウンド数を10回, 100回, 1000回と変化させて, 結果の比較を行っている.

一方, ルールベースアプローチでは, いずれの手法についても, Weka3.4.7 (Waikato Environment for Knowledge Analysis) *を用いている. Weka は, Waikato 大学 (ニュージーランド) の機械学習センターを中心に Java 言語で開発が行われているデータマイニングツールであり, 数多くの機械学習に基づく判定器を実装している.

表2にあるように, 今回の実験に用いた素性数は決して少なくはないものの, ここでは, 素性選択等の処理を行っていない. これは, 多様な主題を包含する学術論文の判定においては, 特定の主題語や内容語だけではなく, 機能語の中に論文固有の特性が現れる可能性があることから, それらを除外してしまうリスクに配慮したためである.

また, 通常のテキスト分類実験では, tf・idf 等によって, 索引語や素性の重み付けを行うことが一般的であるが, 今回, 最も精度の高かった SVM について, 20,000 件の実験集合を対象に, 二値的な語彙の出現情報 (0/1) を用いた場合と, 語彙の出現頻度による重み付け (tf・idf) を行った場合とで, 4 交差検定による比較を試みたところ, 後者の場合, 学術論文の判定数が0件となるものがあったり, 精度・再現率の観点から明らかに性能が低下したことから, ここでは, 前者を採用することとした.

実験結果を表3に示す (→表3. 20,000 件の実験集合における学術論文の判定性能の比較). なお, ルールベースの SVM では, 全てを非論文と判定し, 学術論文と判定されたファイルがなかったことから, 表3からは割愛している. これによれば, やはり, 精度については, SVM (とくに bigram の場合) が高く, 再現率については, ナイーブベイズの高いことが分かる. また, 両者の調和指標である F 値については, 必ずしも大きな差は見られないものの「SVM+形態素解析+改行空白処理済」の組合せのとき, 0.576 と, 最も高い値を示した.

4.2 プーリング手法の網羅性とコスト

実際に, 分類実験の評価のためにプーリング手法を適用するためには, 以下の二点について検討しておかなければならないだろう.

(1) プーリングされる学術論文の網羅性
(2) 人手による判定作業のコスト

4.2.1 プーリング手法における網羅性の検討

表4は, 20,000 件の実験環境において, プーリング手法を適用したと仮定した際に, プーリングされる学術論文の網羅性を検証したものである. 具体的には, 表3に示した 16 の分類手法を単一の分類器とみなして, 4 交差検定における各評価集合 (set1～set4), ならびに, 全体 (total) において, 全学術論文のうち, どれだけを検出することができたかを確認している.

情報検索実験におけるプーリングでは, 上位何件かまでをプーリングすることが通例であるが, ここでは,

* http://www.cs.waikato.ac.nz/~ml/weka/

論文と非論文の二値判定を行っていることから，各システムが学術論文と判定したもの全てをプーリングしている．これによれば，16手法を組み合わせた場合の再現率は0.929〜0.975（全体で0.951）と高い水準にあると言える．しかしながら，数は少ないものの，検出されなかった学術論文が存在することも看過してはならないだろう．

表4.2 万件の実験環境におけるプーリングの網羅性

	set1	set2	set3	set4	total
論文数	81	93	99	98	371
判定数	346	343	348	336	1,373
正解数	79	88	95	91	353
未検出数	2	5	4	7	18
精度	.228	.257	.273	.271	.257
再現率	.975	.946	.960	.929	.951

表5. 手法間の判定論文数の一致度と正解率の関係

手法間一致	正解数	不正解数	正解率
16	23	1	95.8%
15	25	4	86.2%
14	20	7	74.1%
13	15	2	88.2%
12	17	4	81.0%
11	20	4	83.3%
10	22	11	66.7%
9	25	22	53.2%
8	18	19	48.6%
7	13	20	39.4%
6	31	42	42.5%
5	28	52	35.0%
4	21	36	36.8%
3	22	80	21.6%
2	22	102	17.7%
1	31	614	4.8%
0	18		
合計	371	1,020	

次に，表5では，16の手法による学術論文の判定がどれだけ一致しているのかを，正解と不正解に分けて示すとともに，一致度の正解率との関係を確認している．これによれば，手法間の一致数が高ければ高いほど，判定の不正解数は減少し，正解率も高くなっていることが分かる．とくに，16全ての手法が学術論文と判定した場合の正解率は95.8%と極めて高く，11以上の手法が一致していれば，74.1%〜95.8%の正解率を示している．

4.2.2 プーリング手法におけるコストの検討

表6は，524,096件の実験環境において，プーリング手法を適用した際に，プーリングされる学術論文を予め人手によって判定する際のコストを検証したものである．これは，表4に示した20,000件を対象とした「論文数」，「判定論文数」，「正解論文数」，「未検出論文数」を，524,096件の場合に，機械的に当てはめて算出した推定値である．したがって，「3.3.2 抽出されたテキストのトークン化」で触れた，実験集合と評価集合の素性数の相違といった実験環境の変化を考慮に入れていないことから，この段階では，単なる予測値に過ぎない点に留意されたい．

これによれば，16の手法がプーリングするであろうユニークな論文数は，35,219〜36,477件であると推定される．これは，コスト面を考慮して，人手による判定作業を行うのに，必ずしも不可能な数値ではないと言えるだろう．また，論文数の推定値は8,490〜10,377件であるから，プーリングされたもののうち23.4%〜29.2%は学術論文ということになる．さらに，推定未検出数は210〜734件となっており，判定漏れは2.5%〜7.1%程度であるから，プーリング手法を適用することに，一定の有効性が認められるものと言える．

表6. 52万件の実験環境におけるプーリングの推定値

	set1	set2	set3	set4	total
論文数	8,490	9,748	10,377	10,272	9,722
判定数	36,267	35,953	36,477	35,219	35,979
正解数	8,281	9,224	9,958	9,539	9,250
未検出数	210	524	419	734	472

5. 学術論文判別実験とプーリングを用いた評価

さて，本章では，実際に，20,000件のラベル付きデータを学習集合とし，524,096件のラベルなしデータを実験集合として，学術論文の自動判別実験を行った．ここでも，出現語アプローチについては，(a) SVM，(b) AdaBoost (R10, R100, R1000) の二種類を用い，MeCab による形態素解析と bigram によるトークン化と，改行・空白未処理と処理済みの組合せにより，四つのバージョンが存在している．また，ルールベースについては，学習集合において学術論文を判定しなかった SVM を除き，(a) AdaBoost (R10, R100, R1000)，(b) ナイーブベイズ，(c) 決定木 (C4.5)，(d) Vote の四種類を用いた．

これら16の手法によって，プーリングされた異なり判定論文数は36,857件であり，表6に示した推定値（35,219〜36,477件）と概ね一致していることが分かる．これらについて，6人の判定者が，「3.2 学術論

表7. 544,096件の実験集合における学術論文の判定性能の比較

アプローチ	手法	改行・空白		精度	再現率	F値	判定論文数	正解論文数
出現語	SVM	未処理	形態素	.826	.380	.520	6,189	5,110
			bigram	**.849**	.377	.523	5,972	5,073
		処理済	形態素	.824	.389	.529	6,354	5,233
			bigram	.835	.388	.530	6,249	5,219
	AdaBoost	未処理	Round10	.669	.498	.571	10,013	6,699
			Round100	.758	.483	.590	8,576	6,497
			Round1000	.777	.432	.555	7,477	5,809
		処理済	Round10	.679	.484	.565	9,577	6,507
			Round100	.751	.482	.587	8,628	6,482
			Round1000	.774	.428	.551	7,433	5,755
ルール	AdaBoost		Round10	.601	.505	.549	11,279	6,784
			Round100	.698	.465	.558	8,951	6,246
			Round1000	.729	.429	.540	7,913	5,768
	決定木(C4.5)			.728	.387	.505	7,145	5,202
	ナイーブベイズ			.366	**.936**	.527	34,350	12,589
	Vote			.653	.581	**.615**	11,972	7,817

文と非論文の判定作業」で述べた基準にしたがって，実際に判定作業を行ったところ，学術論文数は13,446件（36.5%），非論文数は23,411（63.5%）となった．

表7は，544,096件の実験集合における各分類器の判定性能の比較である（→表7. 544,096件の実験集合における学術論文の判定性能の比較）．これを見ると，表3に示した20,000件の実験環境と比較したとき，SVMで再現率が低下したことを除けば，評価指標の値は，いずれも高まっていることが分かる．

再現率のみが向上している場合，プーリングされなかった学術論文が存在したためである可能性が高いが，いずれの手法についても，精度が向上していることから，全体的な性能が良くなっていると捉えてよいだろう．この理由としては，学習集合の数が増加したこと（15,000件→20,000件），実験集合中の学術論文数の比率が高まったこと（29.2%→36.5%）等が考えられる．また，各々の手法の判別性能を順位付けした場合，SVMの精度が高く，ナイーブベイズの再現率が高いなど，54万件においても，2万件の場合とほぼ同内容の結果が導かれている．以上の結果から，より大きい評価集合への適用可能性，ならびに，プーリングによる評価について，一定の有効性が示されたと言えるであろう．

参考文献

1) Mayor, Susan. "Libraries Face Higher Costs for Academic Journals". http://www.bmj.com/cgi/content/full/326/7394/840/a.

2) "Google, 学術情報専用の検索エンジンを発表". カレントアウェアネス-E, http://www.dap.ndl.go.jp/ca/modules/cae/item.php?itemid=279.

3) 石田栄美ほか. "文体からみた学術的文献の特徴分析". 三田図書館・情報学会研究大会発表論文集, pp.33-36(2004)

4) 石田栄美ほか. "日本語PDFファイルを対象とした学術論文の自動判定". 日本図書館情報学会, 三田図書館・情報学会合同研究大会発表要綱., pp.165-168(2005)

5) 安形輝ほか. "オープンアクセスを想定した日本語学術論文ファイルの自動判定". 情報処理学会研究報告, 2006-FI-82, Vol.2006, No.33, pp.55-62(2006)

6) 安形輝ほか. "日本語学術論文 PDF ファイルの自動判定". Library and Information Science, No.56, pp.43-63(2006)

7) Lewis, David D. Reuters-21578 Text Categorization Test Collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/

8) Chapelle, Olivier, et al. Semi-Supervised Learning. MIT Press, 2006, 508 p.

9) Nigam, Kamal, et al.. "Text Classification from Labeled and Unlabeled Documents Using EM". Machine Learning, Vol.39, No.2/3, pp.103–134(2000)

10) Gilbert , H., et al. Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection. Computer Laboratory, University of Cambridge, BL R&D Report 5481, 1979.

11) Kuriyama, Kazuko, et al. "Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop". Information Retrieval, Vol.5, Issue 1, pp.41-59(2002)

12) 三根慎二. "オープンアクセス資料のファイル形式", http://www.openaccessjapan.com/archives/2006/05/oa_1.html

13) Schapire, Robert. "BoosTexter: A Boosting-based System for Text Categorization", Machine Learning, Vol.39, No.2/3, pp.135-168(2000)