学術情報に特化した検索エンジンの開発
－機械学習による英語論文の自動判定－

安形輝（亜細亜大学）*　　池内淳（筑波大学）　　石田栄美（駿河台大学）
宮田洋輔（慶應義塾大学大学院）　上田修一（慶應義塾大学）
*agata@asia-u.ac.jp

【抄録】 本研究グループでは学術情報に特化した検索エンジンの開発を目標として一連の研究を行っている。その実現には、インターネット上のコンテンツから学術論文を自動的に判定することが必要となる。先行研究では判定手法を提案し、日本語ファイルを対象とした実験を行った。本研究では、対象とする範囲を拡大し英語のファイルも含めることで、言語による違いを越えて提案手法が有効かを検証した。判定実験では、日本語ファイルを対象とした実験よりも高い性能での論文判定を行うことができた。

1. 論文の自動判定と検索エンジン
1.1 学術論文の自動判定の意義

学術論文は研究者にとってもっとも基本的な研究成果の公表メディアである。そのため、「学術論文」という概念を構成する中心的な要素としての査読制度に関する研究は多い。しかし、査読制度を離れたところで、学術論文を学術論文とする特性について分析した研究はほとんどない。また、論文の構造については経験的に序論、本論、結論などとある程度定式化されたものはあるが、実際に論文と非論文の差異という観点から論文の構造とはどのようなものかを扱った研究は少ない。例えば、学術論文の構造を情報検索に応用した事例[1]はあるが、これはあくまでテキスト構造を検索効率向上に用いたものである。さらに、近年、学術雑誌の多くが電子ジャーナル化するなかで、学術論文という概念そのものが大きく揺らいでいる。

本研究グループは、一貫して分野を限定せず、全体の特性に基づく一般的なコンテンツからの学術論文の自動判定を行ってきた。どのような特性が学術論文を学術論文としているかを詳細に検討していくことで、査読制度以外の点から、メディアとしての学術論文の特性を明らかにすることができる。

1.2 学術情報に特化した検索エンジン

オープンアクセス（以下 OA）とは、"完全に無償で制約のないアクセスによって、学術文献を世界規模で電子的に提供すること"[2]とされている。OA を実現する手段としては、「セルフアーカイビング」、「OA ジャーナル」などがある。機関リポジトリに登録された研究成果であれば、OAI-PMH[3]によってある程度体系的な収集が可能である。しかし、それ以外の手段で公開された研究成果の場合、体系的な収集方法はない。そのため、様々な手段で公開された OA の研究成果を横断的に検索する場合には、Google などの一般的な検索エンジンを用いて検索するか、Google Scholar[4]などの少数の学術情報に特化した検索エンジンを用いて検索するほか手立てはない。

一般的な検索エンジンで特定の著者、論題の研究成果を検索する場合、たとえ適合する文献が検索されても、他の膨大な検索ノイズに埋没してしまう可能性が高い。

学術情報に特化した検索エンジンとしては、Google Scholar や CiteSeer.IST[5]等が存在する。Google Scholar は分野を限定しておらず、収録範囲も広い。しかし、先行研究からはGoogleと比べクローリングの対象範囲は狭く、その周期も長いと推測される[6]。CiteCeer.IST は、計算機科学分野を中心とした限定的な収集で、規模はそれほど大きくない。

さらに、一般的な検索エンジン、学術情報に特化した検索エンジンの双方ともに収録範囲や検索アルゴリズムは公開されていない。そのため、収録範囲や出力順位に偏りがあっても知るすべはない。

1.3 本研究の目的

本研究のグループの目標は、学術論文を他のウェブコンテンツから自動判定し、検索できるようにする検索エンジン「Aletheia」を構築することである。Aletheria は、その判定アルゴリズムや検索アルゴリズムを公開することで、従来のエンジンよりも公益性の高いサービスを目指している。

すでに日本語ファイルを対象とした論文の自動判定については、複数の分類器の判定結果をプーリングすることで実用的な論文判定を行うことができることを明らかにした[7]。今回は、対象とする範囲を拡大し英語のファイルも含めることで、言語による違いを越えて判定手法が有効かを検証する。

なお、判定の対象とするウェブコンテンツは PDF ファイルである。他の形式と比べ文書のレイアウトやデザインを維持したまま閲覧でき、閲覧条件を設定することも可能であるため、論文において最も一般的な配布形式となっているためである。

2. 実験集合の作成

実験用 PDF ファイル集合は、検索エンジンで分野を限定せずに、広い範囲からURL を収集し、その中から無作為に選択した URL のうち、実際のファイルをダウンロードでき、テキストを抽出できたものから構

築した。さらに、判定実験を行うために人手により論文の判定を行った。

2.1 PDF ファイル URL の収集

URL 収集には米国の Yahoo!を用いた。選択の理由は、1)検索エンジン API が公開されていること、2) PDF ファイルを指定した検索が可能なこと、3)単位時間あたりの検索数の制限が少ないこと、4)検索結果の取得数の制限が少ないことである。実際に用いた検索エンジン API は Yahoo! Search BOSS(Build your Own Search Service)[8]である。

検索エンジン API の検索式として入力した語句は WordNet3.0 に付属の名詞句辞書(index.noun)に収録された 117,797 語句である。複数語から構成される句はそのまま検索式として用いている。1 検索式あたりの検索結果のうち上位 500 件までを収集した。英語の検索式を用いたが、非英語ファイルにも一部に英語を記述することがあるため、収集された集合は英語が中心であるが、非英語ファイルも含まれている。

PDF ファイルの URL の収集は 2010 年 7 月 4 日から 7 月 29 日までに実施した。合計で 22,591,139 件の重複のない PDF ファイルの URL を収集することができた。Google に登録された PDF ファイル数が約 1 億 6 千万件であること(2010 年 8 月 30 日現在)を考慮するとインターネット全体のサンプルとして十分な規模といえる。

2.2 ファイルのダウンロードとテキスト抽出

収集された URL 集合から無作為に 3 万件の URL を抽出し、ファイルを 2010 年 8 月 9 日にダウンロードした。ダウンロードできたものは 29,115 件であった。そこから Apache PDFBox1.2.1[9](PDF ファイルを操作するライブラリ)によって一部分でもテキスト抽出ができたものは 27,848 件であった。

テキスト抽出できた PDF ファイル群からさらに 2 万件を無作為に抽出し、これを実験用集合とした。

2.3 人手による判定

実験用集合に含まれる PDF ファイル全てについて、人手による論文の判定を行った。学術論文の判定規準としては先行研究と同様に、(1)論文の形態をとっている、(2)タイトル、著者名、所属機関が明記されている、(3) 引用、参考文献がある、(4)1 論文が 1 ファイルで構成されている、(5)2ページ以上である、を用いた。基準の統一をはかるために、最初の判定で論文と判定されたものに関しては、他の判定者が改めて判定した。

また、抄録などは英語であるが英語以外の言語で書かれた論文については、非英語という属性を付与した。

2.4 実験集合の特性

(1) 実験集合の基本的な属性

学術論文と非論文のファイル数、ファイルサイズ、ページ数、縦型の割合を表1に示す。表1から、実験用集合 20,000 件中の論文の割合は 10.05%と低いが、日本語を対象とした先行研究の論文の割合(1.63%など)よりも高い。この理由としては、収集に際して検索エンジンを用いたこと、インターネット上の PDF ファイル数が多くなったため、比較的検索結果上位にくる傾向にある論文の割合が高くなったためと考えられる。

表1 実験集合の基本統計

	論文	非論文
ファイル数	2,011	17,989
平均文字数	50152.4	48219.9
平均ページ数	13.1	17.9
縦長レイアウトの割合	99.9%	93.7%

縦長レイアウトの割合は PDF ファイルの 1 ページ目の縦の長さと横の長さを比較したときに縦が長いものの割合を示している。この割合は先行研究では論文において 100%であったが、今回は論文ではあるが横長のレイアウトのものが 2 件含まれていた。理由としては、2 ページを 1 ページに収めたレイアウトであったためである。

(2) ドメインの分布

表2は実験集合中の URL のトップレベルドメインの上位 10 位までのファイル数とその割合を論文と非論文の別に示したものである。論文のドメインは edu が多く、1／4を占めることがわかる。一方、非論文のドメインは com のものが最も多い。

表2 実験集合のドメインの分布

論文			非論文		
ドメイン	件数	割合	ドメイン	件数	割合
edu	502	25.0%	com	5411	30.1%
org	360	17.9%	org	3658	20.3%
com	209	10.4%	edu	1967	10.9%
uk	101	5.0%	uk	1070	5.9%
de	71	3.5%	gov	939	5.2%
ca	67	3.3%	us	679	3.8%
jp	62	3.1%	au	594	3.3%
fr	48	2.4%	ca	528	2.9%
es	45	2.2%	net	490	2.7%
net	42	2.1%	de	334	1.9%
その他	504	25.1%	その他	2319	12.9%
計	2011	100.0%	計	17989	100.0%

3. 実験環境

3.1 判定に用いた属性

学術論文の判定はテキスト分類の課題の一つである。まずテキストの内容を考慮した場合に、PDF ファイル中に出現する語を手がかりとなる属性として用いることが考えられる。この出現語によるアプローチは、従来の研究成果も多く、実績のあるテキスト分類の分類器を用いることができる。ただし、このアプローチでは、素性数が爆発的に増加するため、大規模集合に対して機械学習手法で短時間に処理を適

用するには工夫が必要である。

また、筆者らによる先行研究では、出現語だけからのアプローチは判定性能の面からも不十分なことが示唆された[10]。そこで、経験的に得られた素性を用いたアプローチを採用した

人があるファイルが論文かを判断する場合には、ファイルの内容だけでなく、ファイルのレイアウト、URL 等のさまざまな要素を総合的に手がかりとする。そこで、本研究でも様々な要素を考慮に入れた 3 カテゴリ、24 素性を提案し、機械学習に用いた(表3)。

表3 判定に用いた素性

カテゴリ	属性	
構造	ファイルサイズ	
	ページ数	
	レイアウト(縦型か横型か)	
	暗号化されているか	
URL	ドメインがeduか	
	ドメインがcomか	
	ドメインがgovか	
	paper, article, researchが出現するか	
出現キーワード	研究	research
	論文	paper, article
	引用文献	bibliography, reference
	抄録	abstract
	媒体	journal, bulletin
	調査法	investigation, survey, experiment, analysis
	被験者	subject
	所属	university, laboratory, research institute
	図表	figure, table
	本論文	this (article|paper|research)
	結果	finding, result
	考察	discussion,conclusion,considration
	コード	doi, issn
	査読者	referee, reviewer
	挨拶	hello, good (morning|evening|afternoon)

これらの素性は、1)本研究グループの先行研究で用いた素性[11]、2)人手で判定を行う中で明らかとなった論文と非論文の差異、3)判定集合での統計に基づいて考案したものである。これらは論文と非論文の判定を行うさいの手がかりとなるものであり、非論文で多く出現するような特徴も含めている。つまり、これらを満たせば論文となるような条件とはなっていない。今回新たに追加したものとしては、URL に論文に関する語(paper,article,research)が含まれるか、DOI や ISSN が含まれるか、挨拶が含まれるか、などである。挨拶が含まれるかは非論文に多く出現するであろう要素として加えている。

3.2 判定手法とその実装

判定性能の向上を図り、各判定手法の特性比較のため、できるだけ幅広い観点から採用する判定手法を検討した。結果として、テキスト分類において定評のある SVM、AdaBoost、Random Forest に加えて、観点の異なるナイーブベイズ、決定木(C4.5)、メタ分類器として分類器投票(Vote)からの判定を行った。

各判定手法を実装したシステムとして、Weka 3.5.6(Waikato Environment for Knowledge Analysis)を用いた。Weka は数多くの機械学習に基づく分類器を実装している。

(1) SVM

SVM(サポートベクターマシン)は、Vladimir N. Vapnik によって提案された 2 クラス分類器の一種である[12]。高い汎化性能を持ち、カーネル法により非常に高次元のデータを扱うことができる点が特徴であり、投入する属性数が多くなりがちなテキスト分類において、多くの応用事例がある。

(2) AdaBoost

ブースティング(Boosting)法は、精度がそれほど高くない複数の弱学習器の重み付けを学習することで性能を高める手法である。AdaBoost は初期のブースティング法を改良したもので、Schapire と Singer[13]による実験では、単語の有無による弱学習器を AdaBoost によって組み合わせた分類器が最近傍法(k-NN 法)やナイーブベイズ法による分類器よりも高い判定性能を示している。

(3) Random Forest

Random Frorest は集団学習(ensemble learning)であり、AdaBoost と同様に精度がそれほど高くない複数の弱学習器の組み合わせ方、重み付けを学習することで性能を高める手法である[14]。

(4) ナイーブベイズ

ナイーブベイズ分類器(naive Bayesian classifier)は、ベイズの確率モデルに基づく、単純なシステムである。この分類器を論文判定に応用した先行研究では、精度は低いが再現率が高いという他の分類器とは異なる結果を示した。

(5) 決定木(C4.5)

決定木(decision tree)は可読性の高い分類器であり、近年では AdaBoost などの集団学習の弱学習器として使われることが多い。ここでは Weka に実装されている C4.5[15](モジュール名は J48)を用いた。

(6) 分類器投票(Vote)

複数の分類器を組み合わせたメタ分類器も用いた。これは、学術論文の判定という難しい課題に対して、多くの観点からの予測を用い判定性能向上を目指すという理由から行った。組み合わせた分類器は決定木、ナイーブベイズ、RandomForest である。

3.3 評価尺度

この研究では精度(P)、再現率(R)、F 値(F)を評価のために用いた。精度はどれだけ正確に判定できた

かを、再現率はどれだけ網羅的に判定できたかを示す。ただし、原則的に精度と再現率は反比例の関係にあるため、精度だけあるいは再現率だけから評価することはできない。そこで、総合的な指標として F 値を用いた。F 値はパラメータ α の値によって、精度と再現率の重みを変えることができる。ここでは最も一般的な α=0.5 とした場合を用いた。

$$P = \frac{\text{システムが判定した正解件数}}{\text{システムが論文と判定した件数}}$$

$$R = \frac{\text{システムが判定した正解件数}}{\text{全論文件数}}$$

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}}$$

本実験では、学習用・判定用データを分割し、10 交差検定を行ったが、各データセットにおいて、各評価尺度の値を求め、それらを平均した値を算出した(macro-averaging)。

4. 判定結果

4.1 学術論文を対象とした判定結果

学術論文を対象とした自動判定の判定結果を表 4 に示した。先行研究における日本語論文の判定では精度・再現率の両方が同時に 5 割を越える手法がなかったが、今回は多くの分類器が単独で 7 割前後の性能を出している。また、ナイーブベイズだけは、他の分類器と異なり、再現率が高いが精度は低いという先行研究[11]と同様の傾向が見られた。このことは、ナイーブベイズを他の分類器と組み合わせることで性能を向上させることが可能なことを示唆している。

表4 学術論文を対象とした判定結果

分類器	精度	再現率	F値
AdaBoost	0.717	0.652	0.683
決定木(C4.5)	0.731	0.688	0.709
ナイーブベイズ	0.447	0.896	0.597
RandomForest	0.713	0.710	0.712
SVM	0.729	0.632	0.677
分類器投票(Vote)	0.695	0.773	0.732

4.2 英語論文のみの判定結果

実験用集合から非英語の論文を外した英語論文のみの判定実験の結果を表5に示した。出現キーワードに関する素性は英語に特化したものであるため、若干ではあるが、性能が向上している。

4.3 考察

先行研究[11]における日本語論文の自動判定実験においては、F 値が5割を超えるものがなく、複数の分類器を組み合わせることで性能向上をはかる必要があった。しかし今回は、英語論文に関しては単独の分類器でも7割を超える F 値を得ることができた。理由としては、1)英語論文判定に関する素性の選択が良かったこと、2)以前と比較し DOI などの論文のみに出現する手がかりが付与されるようになったこと、などが考えられる。

今後は 1)今回行わなかった出現語アプローチに基づく自動判定実験を行う、2)実験用集合に用いた以外の大規模 PDF ファイルの URL 集合を収集し、それらを用いて、十分な規模を持つ実用的な検索エンジンの構築を行う。

表5 英語論文のみの判定結果

分類器	精度	再現率	F値
AdaBoost	0.686	0.653	0.669
決定木(C4.5)	0.735	0.712	0.723
ナイーブベイズ	0.455	0.917	0.609
RandomForest	0.729	0.731	0.730
SVM	0.728	0.654	0.689
分類器投票(Vote)	0.699	0.807	0.749

【注・引用文献】

1) 神門典子. "複数領域における日本語原著論文の機能構造分析--構成要素カテゴリの自動付与". Library and Information Science. No.31, 1993, p.25-38
2) Budapest Open Access Initiative. 2002. <http://www.soros.org/openaccess/read.shtml>
3) Open Archive Initiative Protocol for Metadata Harvesting.<http://www.openarchives.org/OAI/openarchivesprotocol.html>
4) "Google Scholar Beta" <http://scholar.google.com/>
5) "CiteSeer.IST" <http://citeseer.ist.psu.edu/cs>
6) 安形輝ほか. "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, 2007, 鶴見大学, 2007-10-13/14,p.57-60
7) 池内淳ほか. "プーリング手法を用いた学術論文の自動判別実験". 情報処理学会第 82 回情報学基礎研究会, 情処研報 Vol.2007 No.34, p.33-40.
8) "Yahoo! Search BOSS" <http://developer.yahoo.com/search/boss/>
9) "Apache PDFBox" http://pdfbox.apache.org/
10) 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 2005, 慶應義塾大学, 2005-10-22/23, p.165-168
11) 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, 2006, p.43-63
12) Vladimir N. Vapnik. The nature of statistical learning theory, 2nd ed. New York, Springer, xix, 314p., 2000
13) Schapire, R.E.; Singer, Y. "BoosTexter : A Boosting-based System for Text Categorization", Machine Learning, Vol. 39, Number 2/3, p.135-168 (2000)
14) Breiman, L. Random Forests, Machine Learning, No. 45, 2001, p. 5-23.
15) J. R. キンラン著(古川康一監訳). AI によるデータ解析. 東京, トッパン, 1995, 293p.