大規模日本語 PDF ファイル集合からの学術論文の自動判定

石田栄美(九州大学)*　安形 輝(亜細亜大学)　宮田洋輔(慶應義塾大学)
池内 淳(筑波大学)　上田修一(慶應義塾大学)
*ishita.emi.982@m.kyushu-u.ac.jp

抄録

本研究グループでは、分野を限定せずに学術論文を収録し、全文へのアクセスも保証する検索システムの研究・開発を進めている。日本語の学術論文判定の性能向上のために、学術論文判定用の PDF ファイル集合を作成し、既存のルールを用いて学術論文の判定実験を行い、その結果の中から誤判定された例を分析することにより、ルールの改善を試みた。その結果、F 値において 0.54 から 0.62 へと性能が向上した。

1. はじめに

近年、オープンアクセスへの関心が高まっている。オープンアクセスとは、Budapest Open Access Initiative によれば、”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“[1]とされている。また、その手段として、セルフアーカイブとオープンアクセスジャーナルが挙げられている。このような流れの中で、学術情報を入手するための様々なサービスが提供されている。たとえば、学術情報を検索・入手するための代表的な検索エンジンとして、Google Scholar[2]や CiteSeer.IST[3]などが存在する。Google Scholar は多くの学術情報を収録している[4]が、全文へのアクセスを保証しているわけではない。また、CiteSeer.IST は全文へのアクセスが可能な場合が多いが、工学系の学術論文が中心である。

本研究グループでは、分野を限定せずに学術論文を収録し、全文へのアクセスも保証する検索システムの研究・開発を進めている。本検索システムでは、インターネット上に公開されている PDF ファイルを対象にしている。大多数の PDF ファイルは学術的ではなく、また学術論文であったとしても様々な分野が含まれている。そのため、PDF ファイル集合の中の学術論文を自動的に判定することが必要である。すでに、2005 年に日本語ファイルを対象に、ファイルに出現するすべての語の出現の有無をもとに判定を行う出現語アプローチと、論文のレイアウト、構造的な特徴、ファイル中に出現する特定の語を用いて判定を行うルールベースのアプローチを実験した[5]。その結果、ルールベースアプローチの性能が高く、最も性能が高い手法で、判定性能を示す基準である F 値(精度と再現率の調和平均)は 0.49 であった。本システムは日本語で書かれた論文だけでなく、英語やその他の言語を含めることも視野に入れている。そのため、2010 年にルールを英語化し、英語ファイルを対象とした実験を行った[6]。その結果、最も性能が高い手法で F 値が 0.74 であった。

学術論文の自動判定の有用性は高いが、日本語ファイルの判定性能は、英語ファイルを対象にした場合の性能に及ばない。これは英語と日本語の言語そのものに内在する違い、英語論文と日本語論文の構造に起因する違いが原因と考えられる。そこで本研究では、日本語ファイルに対する判定性能を向上させる目的で、

① 日本語PDFファイル集合を新たに構築し、
② 自動判定のための学習用集合を作成し、
③ 既存のルールを用いて実験をし、
④ 誤って判定された論文ファイルの特徴を分析(誤判定例の分析)し、
⑤ ルールの追加・修正を行い、
⑥ 追加・修正ルールを用いた判定実験を行う

という手順をとった。

誤判定例の分析とは、判定器が誤って判定したファイルを実際にみることにより、誤判定の原因、性能向上に結び付くルールの手掛かりを見つけることである。判定ルールは、複数のルールから構成されており、単純にルールを追加することが改善に結び付くわけではない。場合によっては、実験と誤判定例の分析を繰り返し(④から⑥)ながら、判定ルールを検証することが必要である。

2. 学習用集合の作成

2.1 URL の収集

分野を限定しない日本語のPDFファイル集合を構築するために、PDF ファイルの URL を以下の手順により収集した。2010 年 12 月に Yahoo! Search BOSS(Build your Own Search Service)を用いて、ファイルタイプを PDF に限定し、言語の指定を日本語とし、URL を収集した。Yahoo! Search BOSS を用いたのは、API の検索制限が緩くより多くの URL を収集できるためである。ただし、検索対象を日本(.jp)に限定しているわけではないので、後で述べるように一部のドメインを除く必要があった。検索語として日本語 WordNet と IPAdic の両方に登録されている名詞 27,384 語を用い、API からの検索結果の上位 1,000 件までを取得したところ、18,239,568URL を得た。その中から重複除去をした結果、最終的に 6,602,504 の固有の URL を得た。この URL 集合から、無作為抽出した 3 万件を 2011 年 1 月にダウンロードした。URL のサブドメインが".cn", ".tw", ".hk", ".kr" または ".sg"であったものは中国語で書かれたファイルであったため、削除した。これは、今回は日本語のみを対象としているためである。ダウンロードしたファイルから、PDF ファイルからのテキスト抽出ソフトである Apache PDFBox1.4 を用いてテキストの抽出を行い、テキストが全く抽出されなかったものは集合から外した。その結果、27,158 の PDF ファイル集合を作成した。

2.2 学習用集合の作成

上で作成した集合から、2 万件をランダムに抽出し、判定者が各 PDF について学術論文/非論文の判定を行った。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。また、今回から学位論文も論文の対象とした。

表 1 に、学術論文ファイル、非論文ファイル、全体における、ファイル数、平均文字数(バイト数)、ページ数、縦長の割合を示した。学習用集合 2 万件のうち、論文の割合は 3.22%(643)であった。用いた検索語が異なることや API の結果を上位 500 位までとしたという違いはあるが、2010 年 8 月に構築した英語を対象にした PDF ファイル集合では、2 万件のうち学術論文の割合が 10.06%(2,011)だったことを考慮すると、日本語 PDF ファイル集合に占める論文の割合は低いといえる。また、論文ファイルのほうが平均文字数も平均ページ数も大きかった。縦長の割合とは、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。論文ファイルのほうが、縦長である割合が高いことがわかる。

表 2 に実験集合中の論文・非論文ファイルの URL のセカンドレベルドメインの上位 5 位のファイル数と割合をそれぞれ示した(.com のみトップレベルドメイン)。論文のドメインの 57.7%は ac.jp であり、論文ファイルの半分以上が ac.jp である。非論文ファイルのドメインは、.com の割合が最も高く 14.3%であったが、顕著な差はみられなかった。

上で示したように、論文、非論文ファイルの間でそれぞれファイルの属性やドメインの分布に違いがみられ、これらも学術論文を判定するときの一つの特徴とすることができる。

表1. 実験集合の基本統計

	論文	非論文	全体
ファイル数	643	19,357	20,000
平均文字数	43,537	24,438	25,052
平均ページ数	14.9	9.9	10
縦長の割合	99.5%	92.1%	92.3%

表2. 実験集合ドメインの分布

論文			非論文		
ドメイン	件数	割合(%)	ドメイン	件数	割合(%)
ac.jp	371	57.7	.com	2,774	14.3
go.jp	47	7.3	co.jp	2,433	12.6
or.jp	41	6.4	ac.jp	1,862	9.6
co.jp	29	4.5	or.jp	1,504	7.8
.com	13	2	go.jp	1,299	6.7
その他	142	22.1	その他	9,485	49
計	643	100	計	19,357	100

3. 判定ルールを用いた論文自動判定実験

一回目の実験で用いた判定ルールを表 3 に示す。これは、先行研究 [6]で用いたものであり、他の先行研究を参考に著者らが学術論文判定を行う中で導き出したものである。判定ルールは、ファイルの構造、URL、出現キーワードから構成されている。ファイルの構造とは、ファイルサイズ、ページ数、レイアウト、暗号化の有無である。URL とは、ドメインが ac.jp, .com, go.jp であるか、または

URL に paper, article, research という語が含まれているかを基準とした。

出現キーワードとは PDF ファイルから抽出したテキストファイルに、特徴素として指定した語が出現しているかが基準となる。実際の論文では、「引用文献」の代わりに、「参考文献」、「参照文献」と書かれることがあり、同じ役割を示す語でも様々な表現がある。そのため、出現キーワードに関しては、まず、カテゴリを作成し、その中で実際に出現する語を決め、特徴素の出現の有無はカテゴリごとに判断した。たとえば、実験、調査、分析、アンケートという語は「調査法」というカテゴリに属しており、これらの語のいずれかが出現していれば、「調査法」カテゴリが出現したとみなした。これら 23 の特徴素を用いて、それぞれのファイルごとに出現の有無を調べ、その情報を判定器に入力した。

表3. 実験に用いた判定ルール(修正前)

	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ac.jpか	
	ドメインがcomか	
	ドメインがgo.jpか	
	paper, article, researchが出現するか	
出現キーワード	**カテゴリ**	特徴素
	研究	研究
	論文	研究報告|論文
	引用文献	(引用|参考|参照)文献
	抄録	抄録|要約
	媒体	紀要|学術雑誌
	調査法	実験|調査|分析|アンケート
	被験者	被験者
	所属	大学|研究所|研究センター
	図表	図|表
	本論文	本(論文|研究)
	結果	成果|結果
	考察	議論|考察|結論
	コード	doi:| doi |ISSN
	査読者	査読者
	挨拶	おはよう|こんにちは|こんばんは

判定器には、Weka3.6.4 に組み込まれている AdaBoost, Decision tree(C4.5), Naïve Bayes, RandomForest, SVM, Vote を用いた。SVM についてはPUKカーネルを用いている。どのような判定器を用いるかを検討することも分類性能向上のためには重要であるが、本発表では、選定する特徴素に重点をおいているため、一般的に性能が高いと言われている判定器を用いることとした。

判定結果を表 4 に示す。判定性能が高いのは Vote であるが、F 値で 0.540 にとどまっている。

表4. 日本語ファイルの判定結果(修正前)

判定器	精度	再現率	F値
AdaBoost	0.646	0.330	0.437
決定木(C4.5)	0.660	0.401	0.499
ナイーブベイズ	0.288	0.792	0.422
RandomForest	0.564	0.495	0.527
SVM	0.789	0.198	0.316
分類器投票(Vote)	0.623	0.477	0.540

4. 誤判定例の分析

誤判定例の分析では、6つの判定器すべてにおいて、非論文と判定された学術論文ファイル(論文誤判定)と論文と判定された非論文ファイルを対象(非論文誤判定)に、すべてについて人手により確認し、誤判定の理由、改善への手掛かりを分析した。修正前の実験結果で対象となるファイルは、論文誤判定が 124 件であり、非論文誤判定が 16 件であった。これらをすべて人手により確認し、まず、判定ルールの出現キーワードを中心に誤判定された原因を分析した。主な原因としては、1)テキスト抽出失敗、2)特徴素の異表現、3)出現キーワードなし、が確認された。

テキスト抽出失敗としたものは、テキストの一部しか抽出されていない例、半角英数字の部分は抽出されているが日本語の部分は文字化けしている例などである。機関リポジトリのファイルの中には、機関リポジトリ用の表紙はテキストが抽出されているが、本文の抽出が失敗している例もあった。縦書きのものは一文字ずつで改行されている例もあった。これらのものは、PDF ファイル上では出現キーワードの特徴素に該当しているようにみえるが、テキストファイルを元にしているため、ルールに該当しないと判断されたと考えられる。これに該当するのは 37 件であり、全体の 3 割と大きな割合を占めているが、PDFファイルの性質やPDFファイルからテキストを抽出するためのソフトに依存する部分が大きいため、本分析では対象外とした。

特徴素の異表現とは、各カテゴリで予め決めた

特徴素以外にも表現があった場合である。たとえば、「引用文献」カテゴリには、引用文献、参考文献、参照文献を含めていたが、実際には「References」(5件)、「参考資料」(3件)なども用いられていた。抄録では、「抄録」という語の他に「要旨」(6 件)、「Abstract」(6 件)、「Summary」(3 件)などの例が見られた。

出現キーワードなしとは、「本稿」「本研究」「考察」などのカテゴリに属する特徴素が本文中に出現しない場合である。これらに関しては、学術論文としての特徴を表現するための新しいルールを考察した。その結果、章やセクションに「はじめに」が用いられている(50 件)、「Key words」(20 件)、「キーワード」(6 件)という語が出現している例などが見られた。

以上のような分析を行い、ルールを追加・修正し実験を行った、さらにその結果をもとに同様の手順で誤判定された例を再び分析し(論文誤判定100 件、非論文誤判定 9 件)、ルールの追加・修正を行った。

誤判定例の分析によって追加・修正したルール(追加・修正分のみ)を表 5 に示した。引用文献、抄録、所属のカテゴリには特徴素を追加し、新たに「キーワード」「見出し」「謝辞」「報道資料」というカテゴリと特徴素を追加した。報道資料は非論文の例に多く見られた。

追加・修正したルールを用いて実験した結果を表 6 に示す。これらの結果から、判定器によっては精度、または再現率が下がっている例がみられるものの、すべての判定器における F 値は上昇している。特に、修正前に最も性能が高かった Vote では F 値が 0.540 から 0.621 へと大きく性能が向上した。

5. おわりに

本研究では、学術論文判定用の PDF ファイル集合を作成し、既存のルールを用いて学術論文の判定実験を行い、その結果の中から誤判定された例を分析することにより、ルールの改善を試みた。その結果、高い性能向上を示すことができた。誤判定の例を分析したことで、テキスト中の出現回数は少ないが、学術論文の判定に有効な特徴素を導き出すことができた。また、分野により論文のスタイルがあるが、それぞれのルールを導き出すことにより、分野を問わない分類がより可能になった。今後は、この改善したルールを別の日本語 PDF ファイル集合に適用し、ルールの汎用性と適用可能性を検証する予定である。

また、本研究でルールベースを用いているのは、「抄録」「要約」などの語と他の特徴(URL の構造、ファイルの特徴)を同時に扱うことが比較的容易なことや判定器に入力するコストも小さいことなどからである。しかしながら、今後はより高度な出現語アプローチとの比較、または組み合わせの可能性を検討していくことが必要である。

表5. 追加・修正したルール

	カテゴリ	特徴素
出現キーワード	引用文献	((引用|参考|参照)(文献|資料)|reference)
	抄録	抄録|要約|要旨|abstract|summary
	所属	大学|研究所|研究センター|†|‡
	キーワード	keyword|key word
	見出し	[¥..　]+はじめに[　]*
	謝辞	謝辞|acknowledgement
	報道資料	プレスリリース|報道資料

表6. 日本語ファイルの判定結果(ルール修正後)

判定器	精度	再現率	F値
AdaBoost	0.678	0.314	0.429
決定木(C4.5)	0.670	0.502	0.574
ナイーブベイズ	0.324	0.820	0.464
RandomForest	0.647	0.558	0.599
SVM	0.698	0.421	0.526
分類器投票(Vote)	0.676	0.574	0.621

【注・引用文献】

1) Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess/read.shtml
2) "Google Scholar Beta" <http://scholar.google.com/>
3) "CiteSeer.IST" < http://citeseer.ist.psu.edu/cs>
4) Meier, J.J. and Conkling, T.W. Google Scholar's Coverage of the Engineering Literature: An Empirical Study. Journal of Academic Librarianship, 2008, 34(3), 196-201.
5) 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
6) 安形輝ほか. "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"2010 年日本図書館情報学会研究大会発表要綱, 藤大学, 2010-10-9/10