日本語 PDF ファイルを対象とした学術論文の自動判定

石田栄美(駿河台大学)* 池内 淳 (大東文化大学) 安形 輝 (亜細亜大学)
久野高志(作新学院大学) 野末道子(鉄道総合技術研究所) 上田修一(慶應義塾大学)
* e-mail：emi@surugadai.ac.jp

学術論文の提供手段として最も一般的な配布形式である PDF ファイル群から，本文に含まれる手がかりをもとに学術論文を判定する手法を検討した。学術論文と論文以外の PDF ファイル集合を作成し，この集合を用いて，ベイジアンフィルタを学習させ，実際の判定を行った。PDF ファイルに含まれる文字列を bigram により切り出し，それを手がかりに学習させた場合，形態素解析により切り出した場合よりも，判定結果がよかった。

1. はじめに

現在，学術論文を対象とした様々な検索サービスが提供されている。たとえば，CiteSeer.IST[1]は, 情報科学に関する英語の論文を中心に収集し，学術論文の全文をほとんど入手できる。また，国立情報学研究所では日本語を中心とした学術論文データベースの検索から本文へのリンクを提供する CiNii(NII 論文情報ナビゲータ)[2]のサービスを始めた。

このように，学術論文の全文へのアクセスがなされつつあるが，日本語の論文を掲載するウェブページの検索，提供は，部分的にしか行われていない。そこで，分野を問わず，研究者が，ウェブ上で提供している日本語学術論文を自動的に収集し，全文を対象とした学術論文の検索を行うシステムやレポジトリの構築を目的とした調査と研究を進めている。

現在，学術論文の提供手段として最も一般的な配布形式は PDF ファイルである。そこで, PDF ファイル群からの学術論文の判定を第一の課題として考えた。PDF は文書のレイアウトやデザインを維持したまま閲覧できるファイルであるため，広い用途で利用されつつある。その中から，主として本文に含まれる手がかりをもとに学術論文を判定する手法を検討した。

PDF ファイルを対象とした学術論文の自動判定は，学術論文である PDF ファイルの特徴を自動判定システムに学習させ，そのシステムを用いて，判定対象ファイルが論文かどうかを判定することで行う。自動判定の結果を評価し，手法の妥当性を検討することが必要である。本稿では，まず，学術論文の自動判定に必要な論文の PDF ファイル集合を作成した。作成方法を，ウェブ上で提供されている PDF ファイルの現状も含めて，述べる。次に，論文の自動判定を行うが，PDF ファイルから論文を判定するための手法は確立されていないため，他の目的で用いられている手法を適用し，既存の手法でどの程度判定が可能かどうかを検証する。その結果をもとに，自動判定手法の改善方法を提案する。

以下では，学術論文の PDF ファイル集合の作成，論文の自動判定手法，結果および考察について述べる。

2. 学術論文の PDF ファイル集合の作成

2.1 PDF ファイルの収集

学術論文の PDF ファイルは，ウェブ上のリンク集や研究機関，研究者が提供するページから収集することもできるが，分野による偏りを少なくするため，検索対象が PDF ファイルと限定することができるサーチエンジンを用いて収集した。ipadic2.5.1 の辞書ファイルの 6 ファイルから無作為に選定した 10,000 語の検索語を用いて, 各検索語に対し最大 100 件まで PDF ファイルの URL を収集した。重複 URL を除いた 307,514 件の URL を実際にダウンロードした。暗号化されている PDF ファイルや壊れているデータを除去し，最終的に PDF ファイル集合は，248,314 件となった。

2.2 学術論文の PDF ファイルの判定

この PDF ファイル集合から, 3,000 件を無作為に抽出し，6 人の判定者が各 500 件に対し論文かどうかを判定した。判定を迷うものに関しては，改めて 6 人が判定し 5 人以上が学術論文と判定したファイルを学術論文に含めた。学術論文の判定規準として，(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 ページ以上であるなどを用いた。

報告書や内部向けの発表は，学術論文以外とした。なお，日本語のファイルを対象にしているが，英語や中国語のファイルも一部含まれていたため，これらも学術論文以外と判定した。その結果，3,000 件中 98 件(3.3%)が学術論文と判定された。これらを使って事前調査を行った。

さらに，PDF ファイル集合から無作為に抽出した 10,000 件を対象に，先に判定した 3,000 件と同じ手順で学術論文の判定を人手により行った。その結果，284 件(2.8%)が学術論文と判定された。現在，ウェブ上で提供されている PDF ファイルの中では，論文の割合は非常に小さいことがわかる。

2.3 実験集合

本実験では，事前調査に用いた 3,000 件の集合と新たに判定した 10,000 件の集合から，12,000 件を無作為に抽出し，これを 3,000 件ずつ 4 分割し，3,000 件を評価用集合，残りの 9,000 件を学習用集合とし，4 交差検定で実験を行った。

学術論文と論文以外（以下，「非論文」と示す）のファイル数，ファイルサイズ，ページ数を表 1-1 に，ファイルの URL のサブドメインの上位 5 位までを表 1-2 に，論文ファイルの分野の分布を表 1-3 に示す。表 1-1 から，PDF ファイル集合 12,000 件中の論文の割合は 2.9%と低かった。平均ファイルサイズは論文の方が多いが，平均ページ数は論文と非論文に差はなかった。表 1-2 から，サブドメインが co の論文ファイルが最も多かった。一方，サブドメインが ne，or の非論文ファイルは，論文の割合と比較して多かった。N/A は，サブドメインがないものである。非論文の特徴として，市町村の公報や申請書類などが多く見られた。最近では，市町村のお知らせなどを PDF 形式で提供している現状がうかがえる。表 1-3 をみると，論文の分野には自然科学が多く，ついで技術，社会科学の順になっている。ウェブ上で提供されている論文は，自然科学分野が人文科学分野に比べて多いということを示している。

表1-1 ファイル数･サイズ･ページ数

	論文	非論文
ファイル数	345	11,655
平均ファイルサイズ(バイト)	408,981	310,331
平均ページ数	8.17	8.08

表1-2 サブドメインの分布(上位5位)

	論文		非論文	
サブドメイン	件数	比率	件数	比率
co	51	21.70%	1,441	22.13%
ac	34	14.47%	1,769	27.17%
go	13	5.53%	293	4.50%
ne	5	2.13%	653	10.03%
or	3	1.28%	753	11.57%
N/A	85	36.17%	375	5.76%

表1-3 論文の分野

分野	論文数	比率	分野	論文数	比率
総記	16	4.6%	技術	68	19.7%
哲学	21	6.1%	産業	45	13.0%
歴史	9	2.6%	芸術	4	1.2%
社会科学	63	18.3%	言語	10	2.9%
自然科学	104	30.1%	文学	5	1.4%
			計	345	100.0%

3. 論文の自動判定実験の概要

3.1 テキストファイルへの変換

論文と非論文の PDF ファイル集合を用いて，ベイジアンフィルタを用いて，学術論文の自動判定を行った。

ベイジアンフィルタでは，学術論文を判定する手がかりとして，ファイル中で使用されている文字列を用いる。PDF ファイルそのものを扱うことができないため，テキストファイルに変換する必要がある。本実験では，Xpdf [3]を用いて，PDF ファイルからテキストファイルへの変換を行った。

PDF ファイルは表示･印刷時にレイアウトが再現可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 ファイルでは単にレイアウト情報しか持たない。そのため，テキストファイルへの変換を行うと，Xpdf はレイアウトの指定がされている箇所を改行・空白へと変換することが多い。意図された改行・空白と Xpdf によって変換された改行・空白を判別することは困難であるため，ここでは改行・空白の除去等の特別な後処理は行わず，変換したファイルをそのまま用いた。

テキストファイルからは，トークン(文字列や単語)を抽出する必要がある。これには，形態素解析システム MeCab0.81[4]を用いて，形態素に切り出したものと，bigram によって切り出したものをトークンとして用いた。切り出したトークンからの選択は行わず，すべて用いた。なお，bigram による切り出しは，文字種ごとに分けたのち，漢字に対してのみ bigram を適用している。

3.2 ベイジアンフィルタ

ベイジアンフィルタとは，ベイズ理論に基づく分類器であり，どのような分類にも応用可能である。現在では，主として電子メール

の中からスパムメールを検出するシステムで用いられており，特に“A plan for spam”[5]が発表されて以降，多くのシステムが開発されている。

このベイジアンフィルタをスパムメールに応用する場合，非スパムメールとスパムメールに出現するトークンに対するスパム確率を学習し，そのスパム確率をもとに，新たに受信した電子メールに対して，スパムメールの検出を行う。

スパムメールは，内容からも判定することは可能であるが，内容だけでなく，件名の書き方などそのスタイルが判定に有効であるといわれている。メールと論文というスタイルの違いはあるが，本実験では，分野を問わない論文の判定に，このスパムメール判定の方法が適用できるのではないかと考え，用いることにした。

実際に用いたベイジアンフィルタは，日本語にも対応可能である bsfilter[6]を用いた。トークンに対するスパム確率の算出方法は，スパムメール判定に最も精度が高いとされる Gary Robinson-Fisher 方式を用いた。

bsfilter は，各ファイルに対してスパム確率を算出する。スパムメール判定に用いる場合には、この確率が高いとスパムメールであると判定されるが，本実験では，「非論文」として判定する。そこで，以下では，スパム確率ではなく「非論文確率」として扱う。

この確率はシステムが算出する確度であるため，論文－非論文の境界値として，どのような値も設定可能であるが，ここでは，非論文確率 0.6 以上，または 0.9 以上の二つの場合についての結果を出した。

4. 評価尺度

評価尺度には，再現率，精度を用いた。表 2 は，評価用集合中の論文と非論文ファイルに対するシステムが判定した論文と非論文の件数を示している。システムが論文と判定したファイル中，評価用集合の中で人手により判定された論文はA件あったということを示している。判定結果を表 2 の表にあてはめ，以下の式に示すようにそれぞれの評価尺度を求めた。

表2 判定結果に関する分割表

		評価用集合	
		非論文	論文
システムが判定	非論文	A	B
	論文	C	D

・再現率 ＝ A / (A+C)
・精度 ＝ A / (A+B)

再現率と精度は，非論文を中心にみた場合，非論文ファイルをシステムが非論文と正しく判定できたかということに注目している。

本実験では，学習用・判定用データを分割し，4 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も算出した(macro-averaging)。

5. 実験結果

判定集合に用いたデータセット名を 0 から 3 までとして，それぞれのデータセットでの結果および平均の再現率，精度を表 3 に示す。

非論文確率 0.6，0.9 の場合とも bigram によって切り出された場合の再現率のほうが高かった。MeCab を用いた場合の精度は非常に高いが，再現率は bigram に比べて低かった。全体的にみると，文字列の切り出しに bigram を用い，非論文確率 0.6 が最もよい判定結果である。

なお，ベイジアンフィルタには非論文確率の計算方法に Paul Graham 方式があるが，Gary Robinson-Fisher 方式の方が高い結果を示した。また，Support Vector Machine による手法も比較したが，事前調査では，ベイジアンフィルタの有効性が示されている。

表3 非論文の自動判定結果

		切り出し	bigram		MeCab	
		確率	0.6	0.9	0.6	0.9
データセット	0	再現率	98.54%	97.96%	74.50%	65.10%
		精度	98.10%	98.12%	99.73%	99.79%
	1	再現率	98.76%	97.66%	72.41%	60.86%
		精度	97.56%	97.87%	99.91%	99.94%
	2	再現率	97.96%	96.62%	73.99%	63.02%
		精度	97.36%	97.56%	99.91%	99.89%
	3	再現率	99.14%	98.83%	74.01%	63.01%
		精度	97.07%	97.10%	99.81%	99.84%
再現率			98.60%	97.77%	73.73%	63.00%
精度			97.52%	97.66%	99.84%	99.87%

6. 結果に対する考察

6.1 トークンの切り出し手法

ベイジアンフィルタでは，ファイルに出現するトークンに対する非論文確率を学習する。データセット 1,2,3 で学習させた場合，対象となるトークンは 44 万語以上と非常に多かった。非論文確率の高い語の特徴としては，カタカナは連続した文字列で切り出されるので意味をもつ語が多いが，漢字に関しては bigram による語の切り出しをおこなっていることもあり，単独では意味をなさない語も多かった。全体的にみると，固有名詞が多く

含まれていた。

先に述べたように PDF ファイルからテキストファイルの変換を行うと，レイアウトの指定は改行や空白に変換されることが多い。このため，変換されたテキストファイルは，オリジナルの文章を維持しているものばかりではなく，途中に空白や記号が含まれてしまう例も多かった。つまり，形態素解析システムを用いても，オリジナルの意味を維持したままでトークンを抽出することが出来なかった可能性がある。bigram を用いた場合の結果が高かった理由の一つとしては，ひとつひとつのトークンよりも，ファイル中に用いられているトークン全体で非論文を判定していることが考えられる。

6.2 誤り分析

ベイジアンフィルタで論文として判定された非論文，また非論文として判定された論文にはどのような特徴があるかを調べた。評価用集合がデータセット 0 の場合を例として述べる。非論文を論文として判定した件数は 47 件，論文を非論文として判定した件数は 8 件あった。それぞれのファイルには，以下のような特徴があった(重複あり)。

非論文を論文として判定したファイルの特徴には以下のようなものがあった。

(1) テキストファイルへの変換トラブル(18 件)
(2) ファイル中の文章のほとんどが外国語(英語，中国語)(11 件)
(3) 発表用のスライド(3 件)
(4) 報告書，研究要旨(8 件)
(5) 授業用のテキスト(3 件)

(1)は制御コードなどの影響により，変換後のテキスト中が空白や改行だけのファイルである。(3)は，研究要旨と発表文献をまとめたものや，テクニカルレポートなどの報告書が含まれる。また，全体的に，英語や数字，記号が多く含まれる場合は，非論文として検出されない傾向があった。

論文を非論文と判定した特徴には以下のようなものがあった。英語の抄録がある(5 件)または，引用文献の中で英語の文献が多い(5 件)などである。これらの特徴は必ずしも誤判定された論文だけに共通しているとは言いがたく，論文を非論文と判定した明らかな特徴については見出せなかった。最も妥当な理由としては，用いられている語が，論文に多く出現する語が少なかったということであろう。

非論文を論文と誤判定した理由を分析した結果, いくつかの改善点を提案する。まず, (1), (2)についてはベイジアンフィルタにかける前の処理が必要である。日本語ファイルの判定方法を精緻化し，変換トラブルを持つファイルに関しては，ファイルの大きさなどから，事前に何らかの処理を行う必要がある。(3)については, 学術的な用語や英語を用いている場合が多いが，画像が多い，ファイルサイズが小さいという特徴があるので，文字列だけではない他の情報も判定の要素として含めることが考えられる。(4)，(5)については，言葉使いなどは論文とほとんど変わらないため，文字列だけで判定するのは非常に難しいと考えられる。これについては，例えば，文字列内の特定の語, 論文であれば,「受付」「受理日」などが明記される場合もあるので，このような特定の語を判定の手がかりにすることが考えられる。

7. おわりに

本稿では，論文の PDF ファイル集合を作成し，論文の自動判定を行った。その結果，ある程度の確率で，既存のベイジアンフィルタを用いても, 論文判定を行うことができた。今後は，誤り分析から得られた改善案を適用する予定である。

引用文献

1) CiteSeer.IST. “Computer and Information Science Papers CiteSeer Publications Research Index”. <http://citeseer.ist.psu.edu/> [2005-09-22]
2) 国立情報学研究所. “CiNii Home(NII 論文情報ナビゲータ)” <http://ci.nii.ac.jp/> [2005-09-22]
3) “Xpdf” <http://www.foolabs.com/xpdf/> [2005-09-22]
4) “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”. <http://chasen.org/~taku/software/mecab/> [2005-09-22]
5) Paul Graham. “A Plan for Spam”. <http://www.paulgraham.com/spam.html> [2005-09-22]
6) “bsfilter / bayesian spam filter / ベイジアンスパムフィルタ”. <http://bsfilter.org/> [2005-09-22]
7) “TinySVM: Support Vector Machine”. <http://chasen.org/~taku/software/TinySVM/> [2005-09-22]