学術論文 PDF の自動判定：学習用集合が判定性能に与える影響

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

抄録

本稿は，ウェブ上の学術論文 PDF ファイルを自動判定する検索システムを開発する過程で生じた，学習用集合の問題について報告する。機械学習手法を用いて分類する際には，学習に用いる集合（学習用集合）の構築コストと判定性能とのバランスのとれた構築手法が必要となる。そこで，本研究では構築コストの異なる，4 つの学習用集合を構築し，学術論文 PDF の分類性能を比較した。本実験からは学術雑誌ディレクトリや大学サイトを起点として収集したファイル群を学習用集合として用いる場合に高い分類性能が得られた。

1. 背景と目的

本研究グループは，ウェブ上に存在する PDF ファイルから，学術論文かどうかを，機械学習手法によって判定し，提供する検索システムを開発している[1)]。機械学習手法を用いて学術論文を自動判定するためには，学習に用いるためのラベル付された文書集合（学習用集合）が必要となる。しかし，学習用集合を作成するための人手による正誤判定は，一般に高コストである。そのため，検索システムを継続的に運営するためには，構築コストと判定性能のバランスがとれた学習用集合の構築手法が必要となる。

学習用集合の構築コストは，収集コストと判定コストからなる。収集コストは，ウェブ上から学習用集合として用いるファイル群を収集する際のコストである。たとえば，ロボットやクローラーと呼ばれるプログラムを用いて，インターネット上からファイルを収集することのコストは低いが，インターネット上から人手による検索で１つ１つ学術論文を検索する作業は高いコストをともなうと考えられる。

判定コストは，収集したファイル群から学術論文かどうか判定する際のコストである。文献ファイルが持つなんらかの属性に基づいて機械的に判定が出来る場合は，判定コストは低く，１つ１つのファイルを視認し，学術論文であるかどうかを判定する際はその判定コストは高コストになる。

本研究の目的は，学術論文の判定に適した学習用集合の構築手法を明らかにすることである。

2. 実験の概要

学習用集合の違いによる判定性能への影響を比較するために，学術論文の判定実験をおこなった。実験では，ウェブ上から収集した集合を「論文」と「非論文」とに判定した学習用集合によって学習し，判定用集合に含まれるファイルを「論文」か「非論文」に判定し，その性能を評価した。実験の流れを図 1 に示した。本章では実験の要素となる学習用集合の構築方法，機械学習による判定方法，評価手法について説明する。

2.1. 学習用集合の収集と判定

構築コストが異なる以下の 4 つの学習用集合を作成した。4つの学習用集合とは，1）日本の機関リポジトリから OAI-PMH 経由で収集した junii2 形式のメタデータで付与された資源タイプに基づいて作成した集合（機関リポジトリ集合），2）電子ジャーナル・ディレクトリに掲載された学術雑誌サイトから収集した PDF ファイルを人手で判定した集合（雑誌ディレクトリ集合），3）検索エンジンによってウェブ上から収集した PDF ファイルを人手で判定した集合（ウェブ集合），4)大学サイトを起点として収集したPDFファイルの集合（大学サイト集合）である。以下では，これら4つの集合の構築方法について述べる。表 1 に構築した学習用集合の概要を示した。

なお，人手による学術論文の判定をおこなった場合は，以下の基準に従っている。（a）論文の形態をとっている，（b）タイトル・著者名が明記されている（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。

本研究は日本語の論文 PDF ファイルを対象としている。そのため，サーチエンジンとクローラーを使用した収集で混入した英語や中国語の文献

は，論文の体裁を取っていても，「非論文」とした。

2.1.1. 機関リポジトリ集合

日本の機関リポジトリからメタデータを収集し，メタデータに付与された資源タイプに基づいて論文の判定をおこなった。

NII の機関リポジトリリスト[2]に掲載された機関リポジトリから，OAI-PMH 経由で junii2 形式のメタデータを収集した。メタデータは 2010 年 3 月 3 日に収集した。収集日現在で NII の機関リポジトリリストに掲載された機関リポジトリは 122 リポジトリであったが，OAI-PMH のベース URL がわからない・サーバにアクセスできないなどの理由で，実際にメタデータが収集できたリポジトリは合計 104 リポジトリであった。

機関リポジトリから収集したメタデータから，fullTextURL 要素を抽出し全文ファイルの URL を得た。574,374 件の URL から無作為抽出し，PDF ファイルのダウンロードができた，20,000 件の全文ファイルを収録した集合を作成した。

取得した PDF ファイルをメタデータに付与された「国立情報学研究所 メタデータ主題語彙集（資源タイプ）」（NIItype 要素）によって論文と非論文とに分類した。資源タイプが「Journal Article」のものを「論文」，それ以外のものを「非論文」とした。資源タイプに，「メタデータフォーマット（junii2）各データ要素の入力内容一覧」[3]で指定された語彙以外のものが入力されているものに関しては，すべて集合から除外した。この結果，3,115 件（15.6%）の論文 PDF ファイルを得た。

機関リポジトリ集合は，標準的なプロトコルによってメタデータを収集でき，メタデータに含まれた要素に基づいて，学術論文かどうかを機械的に判定できるものである。そのため，収集コストも判定コストも低く作成できる集合といえる。

2.1.2. 雑誌ディレクトリ集合

実践女子大学図書館・短期大学図書館が運営している Directory of Open Access Journal in Japan（現「日本語学術雑誌情報源ナビ」[4]）に掲載された学会・学術雑誌に掲載されたウェブサイトから，PDF ファイルを収集し，人手によって学術論文ファイルかどうかを判定した。

ディレクトリに掲載されたウェブサイトのトップページに掲載されたリンクから，3,562 件の PDF ファイルを取得した。そこから 1 名の判定者が手作業で「論文」か「非論文」を判定し，314 件（8.8%）の論文 PDF ファイルを取得した。

この集合は，学術的な内容が掲載されたウェブページから収集したファイルを収集して作成されている。そのため，人手による判定作業ではあるものの，ウェブ全体から収集した集合に対する判定に比べると，そのコストは下がっていると考えられる。

2.1.3. ウェブ集合

ウェブ上に存在する日本語 PDF ファイルを収集し，人手によって論文の判定をおこなった。

ウェブでの 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 2 度に亘って

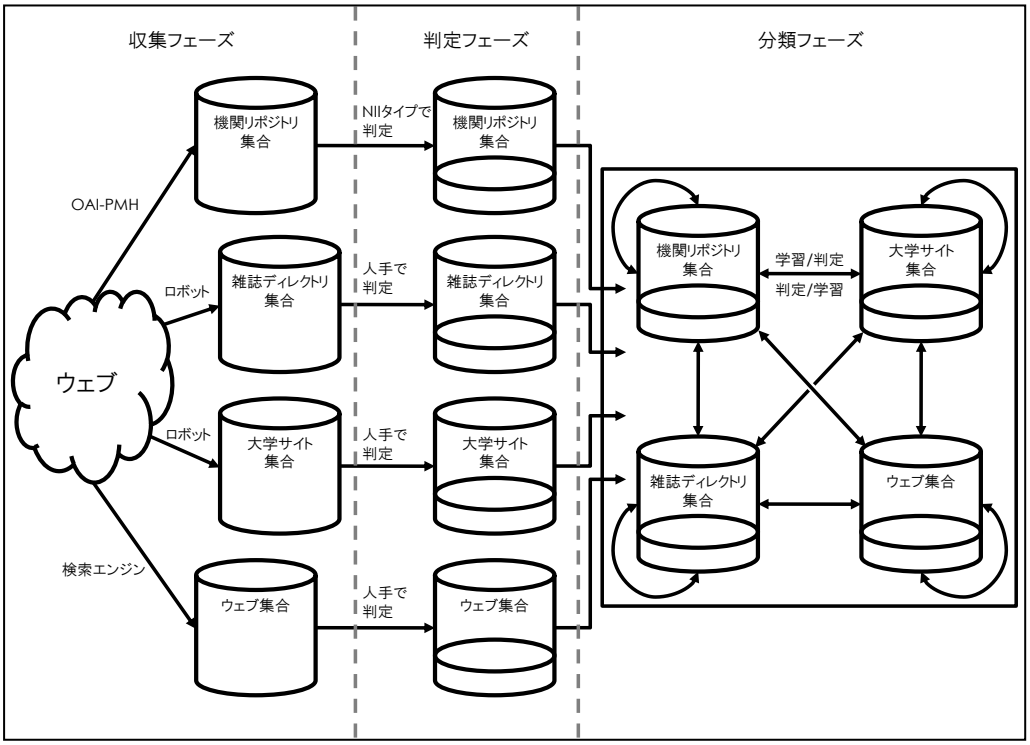

図 1 判定実験のプロセス

おこなった。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 6 つの名詞辞書ファイル（計213,020語）から，それぞれ，9,750語（第1回目），10,250 語（第 2 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 1 回目），441,598 件（第 2 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。その結果，それぞれ，248,314 件（第 1 回目），349,971 件（第 2 回目）の集合が得られた。さらに，2 つの集合を重複除去した結果，544,096 件の日本語PDF ファイルを得た。

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，「論文」と「非論文」の判定をおこなった。判定作業の結果，学術論文と認められるものは 371 件（1.9%）となった

ウェブ集合は，機械的な収集によって収集コストは低いものの，学術論文の割合は相対的に小さく，十分な集合を得るためには，その判定コストは高くなりがちである。しかし，一方で，幅広い範囲から収集し学術論文以外の種々の内容を含んでいるという特徴も持っている。

2.1.4. 大学サイト集合

大学のウェブサイトも学術的な情報を提供していると考えられる。そこで，大学サイトを起点として，PDF ファイルを収集し，学習用集合を構築した。

クローラーを作成し，2009 年 8 月に，筑波大学ウェブサイト（http://www.tsukuba.ac.jp/）を起点として，PDF ファイルを収集した。収集できたPDF ファイルは 29,595 件である。

重複ファイルを除去しこの PDF ファイル群から，2,000 件を無作為抽出し，3 名の判定者によって，人手で「論文」と「非論文」とに判定した。その結果，17 件（0.9%）の論文 PDF ファイルを取得した。

大学サイト集合は，雑誌ディレクトリ集合よりも幅広い内容を持ち，ウェブ集合よりも学術的なコンテンツが多く含まれる集合と考えられる。

表 1 学習用集合の概要

	集合サイズ	論文数	論文%
機関リポジトリ	20,000	3,115	15.6%
ウェブ	20,000	371	1.9%
雑誌ディレクトリ	3,562	314	8.8%
大学サイト	2,000	17	0.9%

2.2. 学術論文 PDF の判定実験

2.2.1. 判定用集合

本研究では，学習に用いた 4 つのデータを判定用集合としても用いて，判定実験をおこなった。つまり，4 つの学習用集合と，4 つの判定用集合で，合計 16 セットの実験をおこなった。

なお，学習用集合と判定用集合が同じ場合には，学習用集合と判定用集合に分割し 4 交差検定の結果で評価した。

2.2.2. 判定に用いる特徴

ヒューリスティックなルールに基づいてPDFファイルから，判定に用いる特徴を抽出した。判定ルールは，論文ファイルの持つ形態的特徴と，テキストの持つ文体的特徴の両方が含まれている。表 2 に判定に用いたルールの一覧を示した。

PDF ファイルの特徴抽出には iText 5.0.1 を用いた。PDF ファイルからのテキスト抽出には，PDFBox 1.1.0 を用いた。

表 2 ルールベースに用いた属性[5)]

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」

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

2.2.3. 判定手法

これまでの研究で高い再現率を得られていたナイーブベイズによって論文を判定した。判定には，テキストマイニングツールの Weka 3.6.2 を用いた。

2.2.4. 判定結果の評価

判定結果の評価には，「論文」を判定する際の F 値と，判定精度（accuracy）を用いた。

F 値は，判定の正確さ（精度）と判定の網羅性（再現率）を組み合わせた尺度である。F 値では精度と再現率の重みを変えることができるが，本実験では同じ重みとして扱った。F 値は以下の式で算出した。

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

判定精度は，非論文も含めて，判定対象の文書をどれだけ正しく判定できたかを示す。F 値では「非論文」の判定性能は考慮されていないが，判定精度では，非論文も含めた分類の正確さを示している。判定精度は以下の式で算出する。

$$\text{判定精度} = \frac{\text{正解論文数} + \text{正解非論文数}}{\text{判定した文書の総数}}$$

3. 実験結果

はじめに，各実験セットでの論文判定の F 値を表 3 に示した。F 値の結果から，学習に大学サイト集合を用いた場合が，平均で 0.235 と最も性能が高い。一方，機関リポジトリ集合は，機関リポジトリ集合自体の判定を除いては，著しく性能が低い。

表 3 各実験セットにおける論文判定の F 値

		学習			
		機関リポジトリ	ウェブ	雑誌ディレクトリ	大学サイト
判定	機関リポジトリ	.344	.115	.130	.107
	ウェブ	.001	.400	.301	.412
	雑誌ディレクトリ	.006	.292	.442	.263
	大学サイト	.008	.093	.051	.157
	平均	.090	.225	.231	.235

つぎに，各実験セットでの判定精度の結果を表 4 に示した。判定精度においても，大学サイトが平均で 0.874 と最も高い。次に，ウェブ集合・雑誌ディレクトリ集合が続いている。一方，機関リポジトリ集合は，ほかの学習用集合に比べて，判定精度においても，ほかの集合に比べてかなり低い性能を示している。

表 4 各実験セットにおける判定精度

		学習			
		機関リポジトリ	ウェブ	雑誌ディレクトリ	大学サイト
判定	機関リポジトリ	.597	.625	.575	.732
	ウェブ	.261	.950	.931	.964
	雑誌ディレクトリ	.303	.775	.791	.843
	大学サイト	.413	.902	.851	.957
	平均	.393	.813	.787	.874

4. 考察

本研究では，機械学習手法に基づく自動分類システムの構築・維持における学習用集合構築コストの問題に対して，構築コストの異なる 4 つの学習用集合を作成し，学術論文判定実験における，それらの性能の比較を試みた。

実験の結果，論文の判定に対しては雑誌ディレクトリ集合が，非論文の判定も含んだ判定精度ではウェブ集合・雑誌ディレクトリ集合・大学サイト集合で概ね同程度の性能が得られた。機関リポジトリ集合は F 値と判定精度のいずれでもやや低い傾向であった。

これらの集合の中でウェブ集合が最も学習用集合の構築コストが高いことを考慮すると、学術雑誌ディレクトリや大学サイトのような学術的なコンテンツを持つサイトを起点として，PDF ファイルを収集し，学習用集合とするのが，分類性能とコストのバランスがとれると考えられる。

引用文献

1) 安形輝ら. "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. 鶴見大学, 2007-10-13/14. 日本図書館情報学会, 2007, p. 57-60.
2) 国立情報学研究所学術機関リポジトリ構築連携支援事業. 機関リポジトリ一覧. http:// www. nii. ac. jp/ irp/ list/ （参照 2010-04-25）
3) 国立情報学研究所. メタデータフォーマット（junii2）各データ要素の入力内容一覧. http://www.nii.ac.jp/irp/archive/system/pdf/junii2_elements_guide_ver2.pdf, (参照 2010-04-25)
4) 伊藤民雄. 日本語学術雑誌情報源ナビ. http:// jcross. jissen. ac. jp/ atoz/ index. html, （参照 2010-04-25）
5) 池内淳ら. "プーリング手法を用いた学術論文の自動判別実験". 研究報告-情報学基礎（FI）. 情報処理学会, 2007-3-27. 情報処理学会, 2007, p. 33-40.