学術論文 PDF 検索システムの開発と評価

安形輝(亜細亜大学)*　池内淳(筑波大学)　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)
*agata@asia-u.ac.jp

【抄録】 近年、Google Scholar BETA、Scirus のように学術情報を対象とした検索エンジンが登場しているが、その検索アルゴリズムは公開されていない。本研究では公開されたアルゴリズムに基づく検索システム「アレセイア」を構築した。このシステムはインターネット上で公開されているPDFファイルを収集し、学術論文あるいはそれに準じる内容を持つ文献を自動的に識別する。本論では専門用語を用いた検索を行い、精度から評価するとともに、既存の学術情報専門検索エンジンとの比較も行った。

1. はじめに

1.1 学術情報に特化した検索エンジン

現在、インターネットの普及と学術雑誌の価格高騰といった背景から学術論文のオープンアクセス化が進展している[1)]。そうした論文はPDFファイルで公開され、即座に全文にアクセスすることができる。GoogleやYahoo！などの一般的な検索エンジンにも登録されるため、検索手段として一般的な検索エンジンを利用することはできる。しかし、学術情報のみの利用を前提すると、その検索結果には学術情報以外のウェブ上の一般的な情報が大量に含まれてしまうため、有効なアクセス手段ではない。

そこで、近年、Google Scholar BETA[i]、Windows Live Academic Search[ii]、Elsevier社の Scirus[iii]といった学術情報に特化した検索エンジンが登場してきている。しかし、これらの検索エンジンの運営は営利企業が行っているため、収録範囲については一定の情報が公開されているものの、検索アルゴリズムなどの詳細については公開されているわけではない[2)3)]。

1.2 学術情報の自動識別

オープンアクセスの実現方法としては1)オープンアクセスジャーナル、2)機関リポジトリ、3)研究者自身によるセルフアーカイビングがある。1)や2)に登録されたファイルはサイト名やドメインからある程度推定できるため、その識別は比較的容易といえるが、3)は一般的なウェブ情報との区別がつかず、その識別は難しい。また、インターネット上には学術論文ではないがそれに準じる内容を有し、研究活動に有効な情報(以下、学術的報告)も数多く存在している。本論では「学術論文」と「学術的報告」を合わせて「学術情報」と総称する。

筆者らは一連の研究[4)5)]の中で、文体や構造に着目しウェブ情報からの学術情報の識別を行う手法について検討してきた。その成果として、複数の機械学習アルゴリズムを組み合わせることで、分野を問わず学術情報(学術論文および学術的報告)の自動識別は一定の精度で可能であることを明らかにした。

1.3 本研究の目的

本研究の目的は、第一に、先行研究の成果を活かし、オープンな学術論文PDFファイルを対象として、公開されたアルゴリズムに基づく検索システム「ＡΛＨΘＥＩＡ(以下、アレセイア)」を開発し、第二に、この検索システムに対する評価を行うことである。

2. 検索システムの構築

インターネット上で公開されているPDFファイルから学術論文を自動識別するシステムの構築は、複数の段階の処理が必要となる。ここでは、PDFファイルの収集、学習用集合作成、機械学習手法による自動識別、検索システム「アレセイア」の構築に分けて説明を行う。

2.1 PDFファイルの収集

PDFファイル集合の作成は、2005年5月と半年後の2005年11月との2度にわたって行った。まず、IPAdic[iv]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、各々の語について、検索エンジンYahoo！Japanを用いて検索を行った。その際、検索対象を「PDFファイル」＋「日本語」に限定するとともに、各検索語の最大収集件数は上位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、

---

[i] Google Scholar http://scholar.google.com/
[ii] Windows Live Academic Search http://www.live.com/
[iii] Scirus http://www.scirus.com/
[iv] IPAdic http://sourceforge.jp/projects/ipadic/

それぞれ、307，514件（第1回目）、441，598件（第2回目）となった。次に、各々のURLに対してPDFファイルのダウンロードを試みた。ダウンロードが不可能であったもの、及び、0バイト・ファイル、破損ファイル、暗号化ファイル、PDFファイルでないもの等を除去した結果、第1回収集では248，314件、第2回収集では349，971件のPDFファイル集合が得られた。さらに、第1回目と第2回目の重複を除去したところ、599，673件となった。

2.2 学習用集合の作成

自動識別の教師データとなる学習用集合は、全PDFファイル集合から、20，000件を無作為に抽出し、6人の判定者が各PDFファイルについて学術論文、非論文であるかを判定したものを用いた。判定の揺れを防ぐため、12，000件を判定した時点で、判定の一致しなかった565件のファイルを改めて6人全員が再判定し、判定基準の統一を図った。

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（縦形かどうか）
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．」「」（カギカッコ）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」

学術論文の判定規準として、(1)論文の体裁をとっている、(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。

なお、日本語のファイルを対象にしたが、PDFファイル収集の際に検索エンジンが誤判定した外国語のファイルも含まれていたため、これらは非論文と判定した。

結果として作成された学習用集合20，000件中の論文の割合は1．63%と低いものであり、自動判定を行うための教師データとしては、非常に偏った集合であるといえる。平均ファイルサイズやページ数は論文の方が非論文よりも多く、論文ではすべてのファイルが縦長であった。

2.3 学術論文の自動識別

学術論文の自動識別は、PDFファイル中の出現語を素性として用いた場合（以下、出現語アプローチ）、経験則による19のルールを用いた場合（以下、ルールベースアプローチ）の判定器、合わせて16の判定器によって行った。ここでは判定器にかける前処理であるテキスト抽出、形態素解析、実際の自動識別に用いた機械学習手法について簡単に述べる。

2.3.1 テキスト抽出と形態素解析

検索システムではPDFファイルを、直接扱うことはできないため、Xpdf 3.01pl2[v]を用いてPDFファイルからテキストを抽出した。

PDFファイルは表示・印刷時にレイアウトの再現が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、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 はレイアウトが指定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・空白の除去を行った場合と行わなかった場合の二つのバージョンを作成し、それぞれを用いて自動判定を行った。

日本語は膠着語であるため、テキストデータを、トークン（文字列や単語）に分割する必要がある。トークン化には、形態素解析システム MeCab[vi]とbigram（2文字単位での切り分け）を用いた。切り出したトークンからの選択は行わず、すべてのトークンを用いた。

2.3.2 論文の自動識別

論文の自動識別に用いた機械学習手法は、出現語アプローチではSVMとAdaBoostの2手

[v] Xpdf http://www.foolabs.com/xpdf/
[vi] MeCab http://mecab.sourceforge.net/

法、19のルール(表1)に基づくルールベースアプローチでは SVM、AdaBoost、NaïveBayes、決定木、Vote の5手法である。

さらに出現語アプローチについては、空白改行処理を行うか否かの2バージョン、切り分け処理が形態素解析(MeCab)か bigram での2バージョン、AdaBoost の学習ラウンド数を変えたものなどを用意したため、これらを組み合わせた16種の判定器を用いた(表2)。

表2 システムに用いた 16 種の判定器

アプローチ	手法	改行・空白	トークン化／ラウンド数
出現語	SVM	未処理	形態素
			bigram
		処理済	形態素
			bigram
	AdaBoost	未処理	Round10
			Round100
			Round1000
		処理済	Round10
			Round100
			Round1000
ルール	AdaBoost		Round10
			Round100
			Round1000
	決定木(C4.5)		
	NaïveBayes		
	Vote		

2.4 アレセイアの構築

アレセイアの基盤となる検索エンジン部分には、Apache Jakarta プロジェクト[vii]の下で開発が進められている Lucene[viii]を用いた。Lucene はJava言語で開発されている全文検索エンジンのクラスライブラリであり、標準では順位付け出力のためにベクトル空間モデルを用いる。日本語の形態素解析システムとしては MeCab のJavaポートである GoSen[ix]を組み込んだ。

アレセイアでは「学術論文らしさ」により順位付けを行うため、検索結果のPDFファイルを論文と判定した判定器数が多い順に並び変え、さらに同順位の場合にはその中で Lucene 標準の順位付けを行う出力用モジュールを独自に実装している。また、ウェブベースの検索エンジンとするために利用者インターフェースを実装している(図1)。さらに、検索結果の入手性を上げるためにキャッシュ機能を実装し利便性の向上を図っている。

図1 検索画面

3. アレセイアの評価

学術情報を対象とする検索エンジンを評価するため、学術的な専門用語から構成される専門用語コーパスを構築し、それに基づいて評価実験を行った。評価実験はアレセイアが論文と判定した検索結果にどの程度の精度で学術情報が含まれるか、と、一般的な利用を想定した環境で他の検索エンジンの検索結果と比較してどの程度の学術情報の全文が即座に入手できるか、という点から行った。

3.1 評価に用いた専門用語コーパス

分野を問わない学術情報に特化した検索エンジンの評価のために、学術論文中に実際に存在する、特定性の高い学術的な専門用語コーパスを用意した。このコーパスは、東京大学[x]、京都大学[xi]、慶應義塾大学[xii]の博士論文データベースにある2005年度博士論文のタイトル(東京大学772件、京都大学755件、慶應義塾大学307件)を収集し、そのリストから無作為に抽出したタイトルに含まれる専門用語(テクニカルターム)群から構成される。このコーパスからさらに無作為に180語を選択し、評価のために用いた。

[vii] Apache Jakarta
http://jakarta.apache.org/
[viii] Lucene http://lucene.apache.org/
[ix] GoSen http://itadaki.org/wiki/index.php/GoSen

[x] 東京大学学位論文データベース
http://gakui.dl.itc.u-tokyo.ac.jp/
[xi] 京都大学博士学位論文データベース
http://edb.kulib.kyoto-u.ac.jp/gakui/zenbun.html
[xii] 慶應義塾大学 OPAC
http://catalog.lib.keio.ac.jp/

3.2 学術情報出力の精度

アレセイアは収集したPDFファイルから自動的に学術情報を識別する検索システムである。そのため評価の観点の一つとしては学術情報識別についての(平均)精度が考えられる。

ここでは180語の専門用語を用いた検索を人手で行い、学術論文あるいは学術的報告が含まれる割合を調べた。判定器が1つでも論文と判定したファイルが1件以上出力されたのは126件である。表3はこの126件に関して上位20件以内に学術論文あるいは学術的報告が含まれるかを調べた結果である。

表3 学術情報の判定精度

1件以上出力	126/180
A. 学術論文数	690
B. 学術的報告数	513
C. それ以外	308
D. 出力文献数	1,511
学術論文精度	0.457
学術情報精度	0.796

表3において学術論文精度は A/D で算出し、学術情報精度は(A+B)/D で算出している。アレセイアは検索出力のあった場合にはその出力中に含まれる学術論文が5割弱、学術情報が8割弱であり、高い割合で学術情報を出力できることがわかる。

3.3 他の検索エンジンとの比較

アレセイアについて一般的な利用を想定した評価を行うために、判定器の判定数に関係なく上位10件に含まれる学術論文あるいは学術的報告の割合を調べ、他の学術情報専門の検索エンジン Google Scholar BETA、Scirus の検索結果と比較した(表4)。ここでは 3.2 での実験と異なり、アレセイアについては判定器が1つも論文と判定しなかった結果も含めて評価しており、そのため、検索できた検索語数が126件から140件と増加している。

表4で「1件以上出力」は検索語集合180件に関して1件以上の検索結果が出力された検索語数を示している。また、「アクセス不可」「リンク切れ」「リンクなし」は、出力された検索結果のアクセス可能性を示している。

表4からは0ヒットとならない検索可能な検索語としては Google Scholar の方が若干多いが、Scirus は大半の検索に失敗していることがわかる。ただし、Scirus はあらかじめ検索語を人手で分割すれば検索結果が改善する。

1件以上出力される結果については、学術論文あるいは学術情報精度のいずれもアレセイアが他の検索エンジンよりも高くなっている。その理由は、Google Scholar BETA は日本語の学術論文に関しては CiNii に依存し、リンク切れ、契約などの原因から実際の論文が入手できないことが多いが、アレセイアは公開されたファイルを収集し、キャッシュ機能を実装しているため、全ての検索結果について入手可能だからである。

表4 学術情報に特化した検索エンジンの比較

	アレセイア	Google Scholar	Scirus
1件以上出力	140/180	162/180	37/180
学術論文数	451	386	39
学術的報告数	337	242	49
それ以外	322	63	77
出力文献数	1,110	1,327	201
アクセス不可	0	296	24
リンク切れ	0	102	12
リンクなし	0	238	0
学術論文精度	0.406	0.291	0.194
学術情報精度	0.710	0.473	0.438

4. 考察と今後の展開

他の学術情報に特化した検索エンジンと比較すると、アレセイアはアルゴリズムが公開されている点で透明性が高いだけでなく、評価実験からは他に勝る精度で学術情報を識別することができ、さらに実際の学術情報の本文を入手できることが保障されている点で有用であることが明らかとなった。

【注・引用文献】
1)倉田敬子. 学術情報流通とオープンアクセス. 東京. 勁草書房. 2007, 197p.
2)Scirus White Paper. http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf
3)Noruzi, Alireza “Google Scholar : the new generation of citation indexes”. Libri, vol.55, no.4, p.170-180(2005)
4)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,p.43-63(2006)
5)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”. 情報処理学会研究報告. vol.2007, no.34, p. 33-40(2007)