様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月４日現在

機関番号：３２６１２
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：２１３０００９５

研究課題名（和文）ウェブ上の文書から学術論文を自動判定し，検索するシステムの設計開発

研究課題名（英文）

The Development of a Search Engine for Academic Papers in Web

研究代表者
上田　修一（UEDA SHUICHI）
慶應義塾大学・文学部・教授
研究者番号：50134218

研究成果の概要（和文）：
研究の目的は,分野を限定せず,日本語及び英語の学術論文への直接的なアクセスを保証し,公開された検索アルゴリズムを用いた学術論文に特化した検索エンジンの構築と評価である。ウェブクローリングを行うために機関リポジトリ収載ファイルを調査し，深層ウェブの存在などウェブ構造を明らかにした。また，日本語および英語で書かれた全分野の学術論文の構成要素と構成を調査し，その結果に基づいて，学術論文の自動判定を行うための判定ルールを構築した。次いでウェブから約 300 万件の日本語 PDF ファイルを収集し，Solr による検索エンジンの構築を行った。既存の検索エンジンと比較評価を行った結果，構築した検索エンジン「アレセイア」は，論文へのアクセスの点で優れており，高い確率で学術論文を自動判定できることが明らかになった。

研究成果の概要（英文）：
Open access scientific papers available on the Web could be searched through several search engines. For example, Google scholar has higher coverage of literature, although it does not necessarily guarantee free access to full text. We have developed and evaluated the “Aletheia” search engine for full text academic papers. The system obtains PDF files on a broad range of topics and automatically detects academic papers using classifiers based on text and structure features. We have built PDF database collection containing 3 million Japanese PDF files, five types of Weka classifiers (AdaBoost, Decision Tree(C4.5), Naive Bayes, Random Forest, and Support Vector Machine) were separately trained for 20,000 test collection using 10-fold cross-validation to automatically detect academic papers. The features were generated using hand-built rules and consisted by the three types of features: structure, URL, and content.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	4,600,000	1,380,000	5,980,000
２０１０年度	5,300,000	1,590,000	6,890,000
２０１１年度	3,900,000	1,170,000	5,070,000
年度			
年度			
総　計	13,800,000	4,140,000	17,940,00

研究分野：総合領域
科研費の分科・細目：情報学・図書館情報学・人文社会情報学

キーワード：学術論文，検索エンジン，ウェブ構造，情報検索，自動分類，機械学習

１．研究開始当初の背景

各種メディアのデジタル化の進展の中で，学術論文は，特に先行して電子化努力がなされてきた。これは，急速な電子ジャーナルの普及によるものであるが，学術雑誌に掲載された学術論文の電子版は，電子ジャーナルサイトばかりではなく，機関リポジトリ，セルフアーカイブなどオープンアクセス状況によるアクセスが可能になった。

従来の学術論文検索システムは，既存の雑誌論文データベースから出発し，その中でアクセス可能な論文ファイルにリンクする方法で，全文アクセスを実現してきた。

これに対し，ウェブをクローリングし，学術論文ファイル特有の形式である PDF ファイルを収集し，学術論文を自動判定して，学術論文データベースを構築し，検索エンジンを提供する方法が考えられる。この方法により，全文ファイルアクセスを提供するだけでなく，論文の全文検索が可能な高度な検索エンジンを提供できる。

２．研究の目的

研究の目的は，分野を限定せず，日本語及び英語の学術論文への直接的なアクセスを保証し，公開された検索アルゴリズムを用いた学術論文に特化した検索エンジンの構築と評価である。

３．研究の方法

(1)　ウェブ構造の解明

ウェブの規模が大きくなるに従い，検索エンジンからアクセスできない状態，すなわち深層ウェブも増えている。先行研究の手法を応用し，日本の機関リポジトリから収集した全文 PDF ファイルの URL を用いて，より大規模に深層ウェブの比率を計測した。

その結果，グーグル，ヤフー，ビングの三つの検索エンジンから検索できるウェブは 72.0%に過ぎず，28.0%が深層ウェブとなっていた。また，一つの検索エンジンでは，最高でもグーグルの 53.2%であった。また，PDF ファイルと URL の特徴の調査から，URL の動的性や長さが深層ウェブとなる要因であることが明らかとなった。

(2)　学術論文の自動判定を行うための判定ルールの構築

判定ルールは，「ページサイズ」「レイアウト」など PDF ファイルの特徴，ファイルの URL が ac.jp ドメインであるかなどの URL の特徴，論文の構造的特徴を示す語，論文に出現する特徴的な語をなどで構成されている。

学術論文は，専門的かつ論理的な記述であるという内容的特性だけでなく，文献の参照，抄録の存在，IMRAD のような構造化，といった形式的特性を有している。そこで，日本で出版された学術論文と海外で出版された学術論文 1,172 件を対象として，学術論文の構成要素と構造を調査した。学術論文の PDF ファイルは，ほとんどの場合，基本的な要素として，論題，著者，所属，抄録，引用文献を持っており，横書きで書かれていた。見出しに含まれる語に基づく学術論文の構造調査の結果，自然科学系を中心に IMRAD 形式に近い構造が多く採られていた。

自動判定に用いる機械学習に基づく分類器は 2 万件の学習用集合を用いて学習させた。分類器として，Weka に組み込まれている AdaBoost, 決定木, NaiveBayes, RandomForest, SVM を用いた。学習用集合を対象とした自動判定実験の判定性能は，RandomForest が F 値で 0.528 と最も優れていた。

(3)　日本語 PDF ファイルの収集

独自のクローラを用いた収集と，既存検索エンジンの利用の二つの方法で PDF ファイルを収集した。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。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 語を用いた。検索結果の上位 1,000 件までを取得した。重複を除き，ドメイン名から中国語等を除いた 6,602,504URL を得た。

この URL 集合から，30 秒以内にダウンロードが可能であり，PDF ファイルの情報やテキスト抽出可能であったファイル 2,947,898 件を取得した。これらのファイルを上記の判定ルールを用いて，5 分類器により自動判定した。個々のファイルについて，学術論文と判定した分類器の数を与え，これを検索結果のランキングに用いた。

(4)　検索エンジンの構築

Apache Jakarta プロジェクト の下で開発が進められている Solr 3.5 を用いた。Solr は Java 言語で開発されている全文検索エンジンパッケージであり，標準では順位付け出力のためにベクトル空間モデルを採用している。日本語の形態素解析システムとしては，lucene-gosen 1.2.1 を組み込んだ。

検索システムの検索結果では，判定ルールを反映させて，論文らしさによる順位付けを行っている。

４．研究成果

構築した学術論文検索エンジンを「アレセ

イア」と名付けて，グーグル・スカラーとサイラスとの比較評価を行った。その結果，アレセイアは論文へのアクセスの点で優れており，高い確率で学術論文を自動判定できることが明らかになった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① [査読有り]深層ウェブの実態とその要因: 機関リポジトリに登録された文献を用いた調査
宮田洋輔，安形輝，池内淳，石田栄美，上田修一
日本図書館情報学会誌，Vol.58, No.2, (2012)

〔学会発表〕（計 10 件）

①学術論文の構成要素と構造
宮田洋輔，石田栄美，池内淳，安形輝，上田修一
2012 年度日本図書館情報学会春季研究集会発表要綱，三重大学，2012-05-12
http://web.keio.jp/~uedas/papers/webir121.pdf

②学術論文に特化した検索エンジンの構築と評価
石田栄美，安形輝，宮田洋輔，池内淳，上田修一
2012 年度日本図書館情報学会春季研究集会発表要綱，三重大学，2012-05-12
http://web.keio.jp/~uedas/papers/webir122.pdf

③Detecting Academic Papers on the Web
Emi Ishita, Teru Agata, Atsushi Ikeuchi, Yosuke Miyata, Shuichi Ueda
JCDL11, June 13-17, 2011, Ottawa, Ontario, Canada, (2011-06-13/17
http://web.keio.jp/~uedas/papers/webir112.pdf

④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，安形輝，宮田洋輔，池内淳，上田修一
2011 年度日本図書館情報学会春季研究集会発表要綱，東京学芸大学，2011-05-14
http://web.keio.jp/~uedas/papers/webir111.pdf

⑤ The Deep Web in Institutional Repositories in Japan
Teru Agata, Yosuke Miyata, Atsushi Ikeuchi, Shuichi Ueda
ASIST 2010, Pittsburgh, Pennsylvania, USA, 2010-10-22/27

⑥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，池内淳，石田栄美，宮田洋輔，上田修一
2009 年日本図書館情報学会研究大会発表要綱，2010，藤女子大学，(2010-10-9/10)

⑦A Search Engine for Japanese Academic Papers
Emi Ishita, Teru Agata, Atsushi Ikeuchi, Michiko Nozue, Yosuke Miyata, Shuichi Ueda
JCDL 2010, June 21-25, Gold Coast, Queensland, Australia, (2010-06-21/25)

⑧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，安形輝，池内淳，石田栄美，上田修一
2010 年度日本図書館情報学会春季研究集会発表要綱，同志社大学，2010-05-29，p.71-74

⑨学術情報流通における深層ウェブの実態 － 機関リポジトリに登録された文献を用いた調査
安形輝，宮田洋輔，池内淳，上田修一
2009 年度三田図書館・情報学会研究大会発表論文集. 2009. 慶應義塾大学(2009-09-26)

⑩Analyzing OPAC Use with Screen Views and Eye Tracking
Ishita, Emi, Mine, Shinji; Koizumi, Masanori; Miyata, Yosuke; Kunimoto, Chihiro; Shiozaki, Junko; Kurata, Keiko; Ueda, Shuichi.
ACM/IEEE Joint Conference on Digital libraries: Designing tomorrow, preserving the past - today (JCDL09). University of Texas, 2009-6-15/19. ACM/IEEE, 2009, p.405.

６．研究組織

(1)研究代表者
上田　修一（UEDA SHUICHI）
慶應義塾大学・文学部・教授
研究者番号：50134218

(2)研究分担者
安形　輝（AGATA TERU）
亜細亜大学・国際関係学部・准教授
研究者番号：80306505
池内　淳（EIKEUCHI ATSUSHI）
筑波大学・図書館情報メディア研究科・准教授
研究者番号：80338607

(3)連携研究者
石田　栄美（ISHIDA EMI）
九州大学・附属図書館・准教授
研究者番号：50364815
野末　道子（NOZUE MICHIKO）
(財)鉄道総合技術研究所・その他部局等・研究員
研究者番号：40426044