Detecting Academic Papers on the Web

Emi Ishita
Kyushu University
Hakozaki, Higashi-ku,
Fukuoka, Japan
ishita.emi.982@m.kyushu-u.ac.jp

Teru Agata
Asia University
Sakai, Musashino-shi,
Tokyo, Japan
agata@asia-u.ac.jp

Atsushi Ikeuchi
University of Tsukuba
1-2, Kasuga,Tsukuba,
Ibaraki, Japan
atsushi@slis.tsukuba.ac.jp

Miyata Yosuke
Keio University
2-15-45 Minatoku,
Tokyo, Japan
m@miyay.org

Shuichi Ueda
Keio University
2-15-45 Minatoku,
Tokyo, Japan
ueda@z5.keio.jp

ABSTRACT

Our research goal is to develop a search engine for open access to academic papers. English and Japanese test sets were built for detection of academic papers from 20,000 PDF files in each language using five annotators. Six classifiers were trained using similar features for each language. We report F1 of 0.74 for English and 0.54 for Japanese and argue that similar features could easily be generated for other languages as well.

Categories and Subject Descriptors

H.3.7 Digital Libraries—Systems issues

General Terms

Experimentation.

Keywords

Search engine, academic papers, PDF.

1. INTRODUCTION

Open access scientific papers available on the Web could be searched through several search engines. For example, Google scholar has higher coverage of literature [1], although it does not necessarily guarantee free access to full text. CiteSeer^X provides access to full text, but only for computer and information science papers [2]. We have developed and evaluated the "Aletheia" search engine for full text academic papers written in Japanese [3]. The system obtains PDF files on a broad range of topics and automatically detects Japanese academic papers using classifiers based on text and structure features [4, 5]. For Japanese queries, evaluation results indicate that Aletheia returns fewer zero-hit results queries and higher precision in the top 10 documents than Google Scholar or Scirus. However, Aletheia currently indexes only Japanese papers. In this paper, we have done extended our experiment to detecting both English and Japanese academic papers on the Web, using similar features to the extent possible.

2. TEST COLLECTION

We have built two test sets containing Japanese and English PDF files, respectively, on a broad range of topics. In July 2010, we collected 22,591,139 URLs for PDF files using the Yahoo! Search BOSS (Build your Own Search Service). We did this by individually posing 117,797 English nouns from WordNet 3.0 as queries, downloading the top 500 items in each result, and removing duplicates. From this set, 30,000 URLs were randomly selected and downloaded. English text was then extracted using Apache PDFBox1.2.1. After losses to bad links and extraction failures, the process resulted in text extracted from 27,848 files.

For Japanese, we used the same API as for English. A total of 27,383 nouns were selected from Japanese WordNet (v1.1) [6] and the Japanese IPAdic [7] noun dictionary and used individually as queries. We downloaded the top 1,000 items from each result set (because the total number of queries is smaller than English). After removing duplicates, this results in 6,602,504 URLs. From that set, 30,000 URLs were randomly selected and downloaded. We found that the result set contained several Chinese files, presumably because the presence of Chinese characters in our queries. We therefore removed any files with a URL ending in ".cn", ".tw", ".hk", ".kr" or ".sg". Japanese text was then extracted from the remaining files using Xpdf3.01p12. This process resulted in Japanese text extracted from 27,158 files.

To generate the test collections, 20,000 files were then randomly selected for each language and annotators were asked to mark academic papers that met all of the following criteria: 1) in the annotator's opinion, had a layout typical of an academic paper, 2) included a title, at least one author name, and an affiliation for at least one author, 3) included exactly one paper per file, 4) included at least one reference, and 5) included at least two pages. Annotators were instructed not to mark theses (Bachelors, Masters, or Ph.D) as academic papers. The annotators were the five authors of this paper. Each file was annotated by one annotator.

A total of 2,011 English files (10%) and 587 Japanese files (3%) were annotated as academic papers. The text extracted from Japanese academic papers is, on average, substantially longer than the text extracted from other Japanese PDF files (38kB vs. 25kB) The same is not true in English (50kB vs. 48kB). On the other hand, English academic papers are often more tightly presented, with the original PDF files averaging 13.1 pages for academic papers and 17.9 pages for other PDF files. A similar trend is not apparent in Japanese, however, (11.1 pages vs. 10.0 pages). The most common top-level domain in the URL for English academic papers (25%) is ".edu" (25%), with ".org" second. For other English PDF files, the most common top-level domain is ".com" (30%), with ".org" again second. An even more pronounced trend was evident for Japanese, with the corresponding ".ac.jp" domain being the most common for academic papers (54.5%) , followed by the government ".go.jp" domain (8%). For other PDF files, the ".co.jp" domain was most common (12.5%), with ".ac.jp" next (10%). These results suggest that domain and length might be useful features when seeking to detect academic papers.

JCDL'11, June 13-17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

3. EXPERIMENT

Six types of Weka classifiers (AdaBoost, Decision Tree(C4.5), Naïve Bayes, Random Forest, Support Vector Machine, and Vote) [8] were separately trained for each language test collection using 10-fold cross-validation to automatically detect academic papers. The features were generated using hand-built rules and are similar to the features used in our previous work [3,4]. Table 1 shows the three types of features: structure, URL, and content. Because American and Japanese URL's use somewhat different structures for domains, we constructed domain features in ways appropriate to each language. We use string matching to detect predefined term categories such as Research, Article, Reference, Media, and Approach. For example, if "investigation", "survey", "experiment" or "analysis" occurs in some file, that file will receive an "Approach" feature. For Japanese, translations of the words in Table 1 were used.

Table 1. Feature set for detecting academic papers.

Category		Feature
Structure		File size
		Number of pages
		Un-coded/coded
		Layout (portrait/landscape)
URL	Domain	.edu, .com, .gov, [English]
		.ac.jp, .com, .go.jp, [Japanese]
	Word	paper, article, research
Content		Research = { research }, Article = { article, papers }, Reference = { reference, bibliography}, Abstract={ abstract }, Media = { journal, bulletin }, Approach = { investigation, survey, experiment, analysis }, Subjects = { subject }, Affiliation = { university, laboratory, research institute }, Figures = { figure, table }, This = { this (article | paper | research) }, Result = { finding, result }, Discussion = { discussion, conclusion, consideration }, Code = { doi, issn}, Reviewer = { reviewer, referee }, Greeting = { hello, good (morning | evening | afternoon) }

Tables 2 and 3 show precision, recall and F1 for five classifiers. The SVM (not shown) classified all Japanese files as non-academic papers. The Vote classifier yielded the best F1 for this two-class classification task for both English and Japanese. Notably, the recall (and thus the F1) is markedly higher in each classifier for English than for Japanese. One reason may be that the ratio of academic papers in English set is about three times as high as for Japanese. Another explanation may be that English academic papers are often in a well standardized defined format specified by a publisher, and our content features were initially designed with those formats in mind (e.g., our inclusion of Digital Object Identifiers (DOI) as a "Code" feature).

Table 2. English classification accuracy.

Classifier	Precision	Recall	F1
AdaBoost	0.72	0.65	0.68
Decision tree (C4.5)	**0.73**	0.69	0.71
Naïve Bayes	0.45	**0.90**	0.60
Random Forest	0.71	0.71	0.71
Vote	0.71	0.76	**0.74**

Table 3. Japanese classification accuracy.

Classifier	Precision	Recall	F1
AdaBoost	0.53	0.29	0.38
Decision tree (C4.5)	**0.65**	0.38	0.48
Naïve Bayes	0.26	**0.80**	0.39
Random Forest	0.53	0.45	0.48
Vote	0.63	0.47	**0.54**

4. CONCLUSION

We have demonstrated the ability to detect English and Japanese academic papers on the Web using similar feature sets. The most interesting aspect of our approach is that it can easily be extended to any language. There are three fundamental ways in which we might improve our overall classification accuracy. First, we plan to consider an unbalanced training set with many more non-academic papers. Second, the content features might be adapted to better reflect the characteristics of Japanese academic papers. Finally, we are interested in developing a statistical analysis process to partially automate the process of generating language-specific features.

5. ACKNOWLEDGMENTS

This research was supported by Grant-in-Aid for Scientific Research B (No. 21300095) provided by the Ministry of Education, Science, and Culture, Japan.

6. REFERENCES

[1] 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.

[2] About CiteSeerX, http://citeseer.ist.psu.edu/about/site

[3] Ishita, E., et al. A Search Engine for Japanese Academic Papers. *JCDL 2010*, p.379.

[4] Agata, T., et al. Automatic identification of academic articles in Japanese PDF files. *Library and Information Science*. 2006, No.56, pp.43-63 (in Japanese).

[5] Ikeuchi, A. et al. Automatic Detection for Academic Articles Using Pooling Method. IPSJ SIG Technical Report, 2007, Vol. 2007, No. 34, FI-86, pp. 33-40 (in Japanese).

[6] Japanese WordNet, http://nlpwww.nict.go.jp/wn-ja/index.en.html

[7] ipadic-2.7.0, http://en.sourceforge.jp/projects/ipadic/

[8] Weka 3, http://www.cs.waikato.ac.nz/ml/weka/