A Search Engine for Japanese Academic Papers

Emi Ishita (Surugadai University), Teru Agata (Asia University), Atsushi Ikeuchi (Tsukuba University),
Michiko Nozue (Railway Technical Research Institute), Yosuke Miyata and Shuichi Ueda (Keio University)

αλήθεια Academic Search Engine

Our Research Purpose

- Development of the “Aletheia” search engine for academic papers written in Japanese using open software

Our Approach

- This system obtains PDF files on a broad range of topics using a commercial search engine, automatically detects Japanese academic papers using classifiers based on text structure, and supports end-user search using Lucene.

Training Set

IPAdic noun dictionary

Select 20,000 individual Japanese words as queries

↓

YAHOO! JAPAN

Download the top 100 PDF files in each results, and remove duplicates

↓

599,673 PDF files

Randomly select

↓

20,000 PDF files

Hand-tag academic papers

↓

Training set

Including 326 academic files (1.63%)

The Criteria for Academic Papers

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
5. Included at least two pages.

Our Research Process

Building a Training Set
- Collecting PDF files
- Marking academic papers

Implementing “Aletheia”
- Using 16 classifiers
- Words / Rules

Evaluation
- Compare with another academic search engine

Approach, Classifiers, and Token

Approach	classifier	(round)	Line break	Token
Word based	SVM		None	Words
	SVM		None	Bigram
	SVM		Delete	Words
	SVM		Delete	Bigram
	AdaBoost	(10)	None	Words
	AdaBoost	(100)	None	Words
	AdaBoost	(1000)	None	Words
	AdaBoost	(10)	Delete	Words
	AdaBoost	(100)	Delete	Words
	AdaBoost	(1000)	Delete	Words
Rules	AdaBoost	(10)	-	-
	AdaBoost	(100)	-	-
	AdaBoost	(1000)	-	-
	Decision tree（C4.5）		-	-
	NaïveBayes		-	-
	Vote		-	-

Implementation

- When the user enters a query, Lucene is used to identify on-topic PDF files, which are then re-ranked in decreasing order of number of classifiers that classify each file as an academic paper.

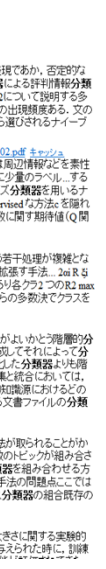
Rules + Hand-Built Features

Layout
- File size
- Second level domain of the URL (ac.jp / go.jp)

Style
- Literacy style
- Using Kanji more

Words
- Research, reference, participant, survey, analysis, experiment, research report, research note, figure, table, paper, consideration, conclusion, research institute, research center, University

Example of a Search Result



www-lab.slis.tsukuba.ac.jp/Aletheia

Evaluation

- A measure of topic coverage; How many queries resulted in at least one result.
- A measure of classifier accuracy; On average, how many of the top 10 documents were (manually determined to be) academic papers by our definition.

Test Queries

- 180 technical terms, randomly selected from titles of doctoral dissertations completed in 2005 at Tokyo, Kyoto and Keio Universities.

	αλήθεια Academic Search Engine	Google Scholar BETA	SCIRUS for scientific information only
Number of results	1,110	1,327	201
No-result rate	22.2%	6.7%	79.4%
Number of academic papers	451	386	39
Academic papers rate	40.6%	29.1%	19.4%
Access failures	0	636	36

Conclusion

- Although commercial services such as Google Scholar index a far larger number of papers than Aletheia, we have shown that with suitable attention to language-specific processing and classifier design it is possible to create systems that are somewhat more robust (i.e., that have fewer zero-result sets) and somewhat better focused (i.e., that return fewer non-academic papers.
- In future work we plan to work on focused crawling for academic papers to address the coverage issue.

JCDL 2010, June 21-25, Gold Coast, Queensland, Australia