eBizSearch: A Niche Search Engine for e-Business

C. Lee Giles[1,2,3], Yves Petinot[2], Pradeep B. Teregowda[3], Hui Han[2], Steve Lawrence[4], Arvind Rangaswamy[3] and Nirmal Pal[3]

[1]School of Information Sciences and Technology
The Pennsylvania State University
001 Thomas Bldg.
University Park, PA 16802
{giles}
@ist.psu.edu

[2]Department of Computer Science and Engineering
The Pennsylvania State University
213 Pond Lab.
University Park, PA 16802
{petinot, hhan}
@cse.psu.edu

[3]eBusiness Research Center
The Pennsylvania State University
401 Business Administration Building
University Park, PA 16802
{pbt105, arvindr}
@psu.edu

[4]Google Inc.
2400 Bayshore Parkway
Mountain View, CA 94043
{lawrence}
@google.com

ABSTRACT
Niche Search Engines offer an efficient alternative to traditional search engines when the results returned by general-purpose search engines do not provide a sufficient degree of relevance. By taking advantage of their domain of concentration they achieve higher relevance and offer enhanced features. We discuss a new niche search engine, eBizSearch, based on the technology of CiteSeer and dedicated to e-business and e-business documents. We present the integration of CiteSeer in the framework of eBizSearch and the process necessary to tune the whole system towards the specific area of e-business. We also discuss how using machine learning algorithms we generate metadata to make eBizSearch Open Archives compliant. eBizSearch is a publicly available service and can be reached at [3].

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]; H.5 [**Information Interface and Presentation**]; I.2 **[Artificial Intelligence];** I.7 [**Document and Text Processing**].

General Terms
Algorithms, Design, Experimentation, Standardization.

Keywords
CiteSeer, Digital Library, eBizSearch, e-Business, Machine Learning, Metadata, OAI, Search Engine, SVM, Tagging.

1 INTRODUCTION
E-business is concerned with e-zation of business processes and encompasses areas as dissimilar as auctions, marketing and customer relationship management (CRM). eBizSearch is an experimental niche search engine that searches the web and catalogs documents pertaining to e-Business. It performs a citation analysis of all the articles collected, maintains an internal graph based on citations between articles and finally provides a web-interface allowing users to explore this graph through various ranking schemes, just as in CiteSeer [2,5]. Articles available through eBizSearch can be downloaded (for fair use) without any charges and in various electronic formats. To date more than 20,000 documents are available from eBizSearch.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007…$5.00.

2 MOTIVATIONS
(1) To build a digital library of relevant academic publications in the field of e-business; **(2)** To enable the exploration of the papers database using the specificities of academic publications; **(3)** To provide a resilient and durable source of publications; **(4)** To add features to document search that are appropriate to the e-business community (e.g. automatic document filtering); **(5)** To comply with the Open Archives Initiative (OAI).

3 ANATOMY OF EBIZSEARCH
3.1 System Overview
The internal architecture of eBizSearch (Figure 1) is organized around a CiteSeer module, which is supplied with new document sources by several crawlers. Users can query the document and citation databases through the dedicated web-interface.

3.2 Crawlers
A set of crawlers discover new potential paper locations (URLs) by continuously exploring the web or subsections of it. Various crawl strategies are currently used: brute force, Inquirus [8] based, and, at the experimental level, focused [1]. The URLs collected by the crawlers are considered to be relevant to e-business.

3.3 CiteSeer
CiteSeer [5] maintains a database of documents and citations but has no intrinsic knowledge on the field of concentration of the documents. Starting from resource locations, it handles the download of the documents. These are then parsed, and their citations information extracted. All documents meeting paper criteria ("Reference" section, length, etc.), are indexed, added to the database and made available for user querying.

3.4 Web-Interface
The web-interface of eBizSearch is based on the same model as CiteSeer [2] and is covered in details in [5].

4 INTEROPERABILITY
As part of a larger scale effort, eBizSearch intends to integrate with the OAI project and to comply with its communication protocols [9]. OAI access to eBizSearch is available at [4].

4.1 OAI Metadata & Protocol Requirements
CiteSeer defines a proprietary set of metadata and communication with clients is HTML based. On the contrary the OAI protocol relies upon the XML-based Dublin Core metadata standard so that it is usable by a wider variety of clients. Our effort in this context

is to enable OAI-based access to eBizSearch by extending the supported set of metadata, and making it fully queriable.

4.2 Architectural & Organizational Issues

To provide OAI compliance with reasonable access time, we need to index on any metadata items that can be queried on (e.g. date, keywords). CiteSeer only indexes document and citation full texts, its metadata collection being available only for internal operations (e.g. ranking, linking). Finally some metadata items are simply not available. In this context, our current organization is presented in Figure 1. The generation of OAI XML records is performed on the fly. The metadata database of CiteSeer is mirrored (periodically sync-ed), and additional metadata items are extracted to address the different requirements induced by OAI compliance. Still our goal is to bring OAI support inside CiteSeer. Thus we progressively extend the set of metadata maintained by CiteSeer (section 4.3), increase the number of indexes to extend the queriable metadata (e.g. date-based queries) and finally upgrade the presentation layer to support the OAI protocol.

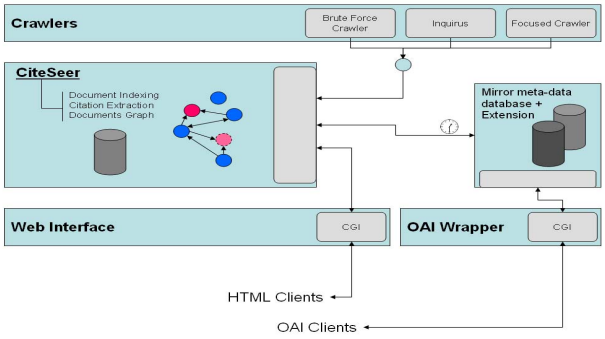

Figure 1: Internal Organization of eBizSearch

4.3 Enhancing Metadata Extraction

OAI compliance requires additional metadata items than currently available from CiteSeer. CiteSeer extracts metadata items using customized regular expressions. But the performance for some of them (esp. author(s) and date) turns out to be poor and often requires manual correction. To extend the set of metadata items, and improve the extraction quality, we propose a machine-learning oriented model where the metadata extraction algorithm results from training. The metadata extraction algorithm used is a Support Vector Machine (SVM) [7], a supervised learning and classification method. The algorithms extracts the 13 metadata items defined in [10] from the header of research papers. Table 1 provides a comparison of our latest experimental results to those reported by [10]. It supports the fact our SVM metadata extraction algorithm could achieve better performance than HMM for metadata extraction with less training data.

5 RELATED AND FUTURE WORK

To our knowledge eBizSearch is currently the only automated niche search engine in the field of e-business that focuses on academic publications indexing. Other free, yet manually maintained, similar services exist such as [6]. Furthermore, documents available from e-business portals usually do not follow the standards of academic publications. As such a major goal of our work is to expand the number of documents indexed by eBizSearch and to increase its functionality. As an example, we wish to automatically generate glossaries categories, such as documents which are primarily technical versus those that are primarily business.

Table 1: Performance of SVM vs. HMM (based on words)

Class	SVM Accuracy	SVM Precision	SVM Recall	HMM multi-state L+D Accuracy [10]
Title	98.9	94.1	99.1	98.3
Author	99.3	96.1	98.4	93.2
Affiliation	98.1	92.2	95.4	89.4
Address	99.1	94.9	94.5	84.1
Note	95.5	88.9	75.5	84.6
Email	99.6	90.8	92.7	86.9
Date	99.7	84.0	97.5	93.0
Abstract	97.5	91.1	96.6	98.4
Phone	99.9	93.8	91.0	94.9
Keyword	99.2	96.9	81.5	98.5
Web	99.9	79.5	96.9	41.7
Degree	99.5	80.5	62.2	81.2
PubNum	99.9	92.2	86.3	64.2

6 CONCLUSION

A new niche search engine, eBizSearch, was described. It has found and indexed over 20,000 documents in e-Business. Its organization around CiteSeer, was presented, along with the efforts to make it OAI-compliant. Finally the limitations of CiteSeer in terms of metadata availability, reliability and portability led us to propose a SVM machine learning approach for metadata extraction. Initial results show that this method accurately automatically extracts, tags untagged text and has the potential for extending the domain of tagged metadata.

7 REFERENCES

[1]: S. Chakrabarti, M. Van den Berg, B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery", In *Proceedings of the 8th International World Wide Web Conference,* pp 1623-1640, Amsterdam, Netherlands, 1999.

[2]: CiteSeer homepage, http://www.citeseer.com.

[3]: eBizSearch homepage, http://www.ebizsearch.org.

[4]: eBizSearch OAI base URL, http://www.ebizsearch.org/oai.

[5]: C.L. Giles, K. Bollacker, S. Lawrence, "CiteSeer: An Automatic Citation Indexing System", In *Proceedings of the 3rd ACM Conference on Digital Libraries (DL'98)*, pp. 89-98, 1998.

[6]: IDEAS homepage, http://ideas.repec.org/.

[7]: T. Joachims, "Text categorization with support vector machines: learning with many relevant features", In *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, pp 137-142, 1998.

[8]: S. Lawrence, C.L. Giles, "The Inquirus Meta Search Engine", In *Proceedings of the 7th International World Wide Web Conference*, pp 95-105, 1998.

[9]: "The OAI Protocol for Metadata Harvesting", http://www.openarchives.org/OAI/openarchivesprotocol.htm.

[10]: K. Seymore, A. McCallum, R. Rosenfeld, "Learning hidden Markov model structure for information extraction", In *Proceedings of the AAAI 99 Workshop on Machine Learning for Information Extraction*, pp 37-42, 1999.