eBizSearch: An OAI-Compliant Digital Library for eBusiness

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

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

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

[3]School of Information Sciences and Technology
The Pennsylvania State University
001 Thomas Bldg.
University Park, PA 16802
{giles}
@ist.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 and when nontraditional search features are required. Niche search engines can take advantage of their domain of concentration to achieve higher relevance and offer enhanced features. We discuss a new digital library niche search engine, eBizSearch, dedicated to e-business and e-business documents. The ground technology for eBizSearch is CiteSeer, a special-purpose automatic indexing document digital library and search engine developed at NEC Research Institute. We present here 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 show 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 [13].

1. Introduction

E-business is concerned with e-zation (digitization) of business processes and encompasses areas as dissimilar as auctions, marketing and customer relationship management (CRM). Here we discuss eBizSearch, a digital library niche search engine for e-business based upon the technology of CiteSeer [5,17,22]. eBizSearch is an ongoing research project at the Pennsylvania State University and is supported by the Smeal School of Business through its eBusiness Research Center.

0-7695-1939-3/03 $17.00 © 2003 IEEE

eBizSearch is an experimental niche search engine that searches the web and catalogs academic articles as well as commercially produced articles and reports that address various business and technology aspects of e-Business. The search engine crawls websites of universities, commercial organizations, research institutes and government departments to retrieve academic articles, working papers, white papers, consulting reports, magazine articles, and published statistics and facts. It performs a citation analysis of all the articles collected, maintains an internal graph based on the citations these articles make and finally provides a web-interface allowing users to explore this graph through various ranking schemes, just as in CiteSeer [5,8,17,22,23]. Articles available through eBizSearch can be downloaded (for fair use) without any charge and in various electronic formats. To date more than 20000 documents are available from eBizSearch.

In section 2 we present the motivations that led to the creation of eBizSearch and what the intended audience for this search engine is. In section 3 we describe the architecture of the system and how it successfully integrates CiteSeer for information-extraction tasks. The issue of OAI compatibility is addressed in section 4. Section 5 is dedicated to our current efforts to extend the applicability of CiteSeer-like digital library niche search engines to various academic fields. Finally in section 6 we reference related projects and present future developments around eBizSearch.

2. Motivations for a Niche Digital Library for e-Business

Many disciplines find that their own focused resources are better sources than general-purpose resources. The current trend is hence in the development of specialized

digital libraries [1,10,11,26,29,30] and their aggregators [9]. As CiteSeer [8] would be a search engine for the computer science literature, eBizSearch would be to the e-business literature. Our goals for eBizSearch are:

1. To build a digital library of relevant academic publications in the field of e-business, and, in terms of the relevance of query results, to outperform general-purpose search engines such as Google, AltaVista, Lycos, etc.

2. To make it possible for users to browse through the digital library's papers database using the specificities of academic publications (e.g. citations between papers), as opposed to the traditional, HTML-based, hypertext navigation on which general-purpose search engines rely. This constitutes the navigation model introduced by CiteSeer.

Table 1: Availability of documents at their original URL (11608 URLs considered – HTTP Status of each URL established by requesting the resource header (HTTP HEAD))

HTTP Code	Semantics	Most probable cause	%
HTTP 200	OK	Document still available at original URL	88.23
HTTP 404	Not Found	Document no longer available at original URL	4.71
HTTP 500	Server Error	Server down or no longer exists	4.63
HTTP 400	Bad Request		1.2
HTTP 302	Moved Temporarily		0.59
HTTP 403	Forbidden		0.38
Other			0.26

3. To provide a resilient and durable source of publications. The ever changing topology of the web must be acknowledged: resource locations change from one day to another or simply disappear hence making the simple knowledge of an URL insufficient to guarantee the long term access to an electronic resource. In this perspective our system is independent from the documents authors/hosts and ensures long-term availability since documents are downloaded, processed, converted to multiple formats and hosted on our servers. We recently checked the availability, at their original source, of the documents available (i.e. referenced and downloadable) from eBizSearch (collection began in 1999); the results are listed in Table 1 and confirm the trend aforementioned.

4. To add features to document search that are appropriate to the e-business community such as automatic document filtering

5. To make eBizSearch compliant with the Open Archives Initiative [31].

CiteSeer [8] has been probably the most successful digital library niche search engine for Computer Science. The high popularity that it benefits from, together with the desire for a permanent archive, enables the documents referenced in its database to be highly ranked among the URLs listed by a general-purpose search engine such as Google. We expect eBizSearch, and the CiteSeer-like niche search engines that will follow, to perform as well, if not better. The intended audience of eBizSearch is researchers in the field of e-business as well as any individual having an interest in this field.

3. Anatomy of eBizSearch

3.1. System Overview

The internal organization of eBizSearch is presented in Figure 1. As can be seen the architecture of eBizSearch essentially exploits that of CiteSeer and uses much of that technology.

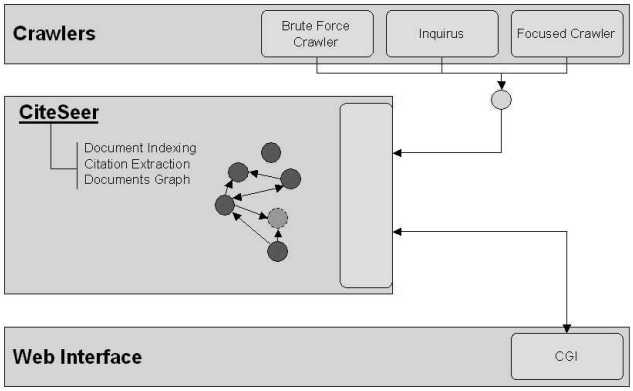

Figure 1: Internal organization of eBizSearch

A set of crawlers, independent from each other, provides the CiteSeer module with the URLs of sources of potential papers. The CiteSeer module takes care of the download phase, converting and parsing each document. If a document falls into the paper category according to CiteSeer (i.e. satisfies various requirements, among which the existence of a "Reference" section, the minimum paper length, etc.), then it is added to the database and made

available for user querying. Users can query the system through the dedicated web interface.

In the following sections we go into more details on the role of each component.

3.2. Crawlers

New documents can be submitted to the system in two ways: through manual submission of a given document (URL), or automatically as a result of a crawling phase. The web interface features a submission page allowing users to manually submit paper locations (humanly reviewed before actual addition to the system). We present here the crawling strategies experimented in eBizSearch.

The crawling phase consists into discovering new potential paper sources (URLs) by guided, focused or extensive exploration of the web or subsections of it. The input for a crawler is one or many seed URLs from which to start exploring the web. The crawler follows hypertext links from one page to another in a more or less biased fashion (focused as opposed to brute force). Source pages, that is, pages containing links to potential papers (e.g. links to file with PDF/PS extension), are logged. Periodically the collected URLs are submitted to CiteSeer for processing, and upon adequacy with paper features, the document/paper is added to the database (refer to overview of extraction process). Note also that known source pages of e-business papers are periodically revisited in order to collect new publications.

Three independent crawlers currently provide eBizSearch in potential publication sources:

- Brute force crawler: when a new repository of interest in the field of e-business is brought to our attention we explore the corresponding sub-network in an extensive fashion to locate most, if not all, of the relevant publications sources available on this site. Brute force crawl is the most efficient on sites that feature a publication section, in which case we can take advantage of this explicit organization to optimize the crawl time (e.g. eCommerce Research Forum at MIT [16]).

- Inquirus based crawler: Inquirus is a meta-search engine described in [23]. By querying Inquirus adequately (i.e. by including one or many keywords referring to publications (e.g. "publication" and/or "journal" and/or "preprint" and/or "ps", etc.) we take advantage of the wide coverage of the web of many general-purpose search engines such as Google or Lycos. The concentration domain of the niche search engine, in this case e-business, defines the queries submitted to Inquirus. Our system systematically generates all possible query strings out of a glossary of words relevant to e-business. The URLs returned by Inquirus are submitted to the CiteSeer module.

- Focused crawler: at an experimental level we work on the development of focused crawlers [6] that would follow only relevant links during their exploration of the web to maximize the eventual discovery of relevant documents. Various crawlers are being tried out for the suitability for this purpose; this includes rule-based crawlers and context-based focus crawlers [7].

As shown in Figure 1, URLs output by all crawlers are pushed in a common queue and batch-submitted to CiteSeer. Batch-submission is made necessary due to CiteSeer's internal organization in which document processing and querying are mutually exclusive operations: the document processing being quite time consuming (the average processing time, including download, is approximately 15 minutes), it is desirable to batch-submit documents to limit the amount of time the service is not reachable.

The crawlers described in this section strongly concentrate on localizing e-business publications originating from academic institutions (Business Schools essentially). We provide in Table 2 the list of US Business Schools for which the crawling of their web domain yielded the largest number of relevant publications in the field of e-Business (publications freely available from the web servers of these institutions). For comparison we mention the ranking of these institutions in the US News ranking of Business School (2003).

Table 2: US Business Schools accounting for the most documents in eBizSearch

School Name (US News Business School ranking)	Percentage of Total Documents.
Massachusetts Institute of Technology (4)	1.55 %
University of Pennsylvania (3)	1.53 %
Northwestern University (5)	0.95 %
University of Chicago (6)	0.53 %
Columbia University (8)	0.47 %
Duke University (6)	0.47 %
University of Virginia (10)	0.14 %
Cornell University (16)	0.13 %

For completeness we also list in Table 3 the general ranking of top sources for documents indexed by eBizSearch.

On completion of the crawling phase it is assumed that the collected URLs are indeed relevant in the domain of concentration of the niche search engine, i.e. e-business.

Table 3: Sources accounting for the most documents in eBizSearch

Source	Percentage of Total Documents.
International Institute for Applied System Analysis	3.32 %
Santa Fe Institute	2.89 %
AT&T	1.56 %
Massachusetts Institute of Technology	1.55 %
University of New Castle upon Tyne	1.54 %
University of Pennsylvania	1.53 %
University of Maryland (CS department)	1.48 %
Federal Reserve Bank of Boston	1.35 %

3.3. CiteSeer

CiteSeer maintains the 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, their citations information extracted, and if the documents follow the pattern of academic publications, they are eventually indexed and added to the database. The internal organization of CiteSeer is beyond the scope of this paper and can be found in [5] and [17]. We give a brief overview for each of the tasks carried on by CiteSeer.

3.3.1. Document Retrieval. Documents are submitted to the system by their location on the web (URL). For efficiency CiteSeer supports concurrent download of multiple documents. The system is resilient to unavoidable availability and connection issues.

3.3.2. Information Extraction. The information extraction (IE) tasks consist into the parsing of citation information following the typical patterns of academic publications. The document is first converted to plain text, the IE tasks being performed independently from the original electronic format. Among other criteria, CiteSeer will reject a document that cannot be converted to plain text, that is too short, or that is not referring to other documents. The specific problem of citation information extraction is addressed in [22].

3.3.3. Document / Citation Querying. CiteSeer provides a support for full-text querying of both documents and citations (documents and citations are indexed into two independent indexes). After a first query, citation-oriented exploration of the document graph is available to the user. Note that for efficiency in real-time query handling CiteSeer maintains various caches (index cache and query response cache).

3.3.4. Distributed Error Correction. CiteSeer provides functionalities allowing authors to provide correction regarding those of their publications that are available from CiteSeer. The correction functionalities are available from the back-end interface only. Correction requests can be made from the web-interface. Support of distributed error correction by CiteSeer is extensively discussed in [24].

3.4. Web-Interface

The last essential component to eBizSearch is its web-interface. A screenshot of the main form of the web-application is shown in Figure 2. The main form allows full text querying of both documents and citations.

Figure 2: Main form of eBizSearch

The presentation of search results is based on the same model as CiteSeer [8] and is covered in details in [17]. Figure 3 shows typical results presentation when querying the document database for "Malone".

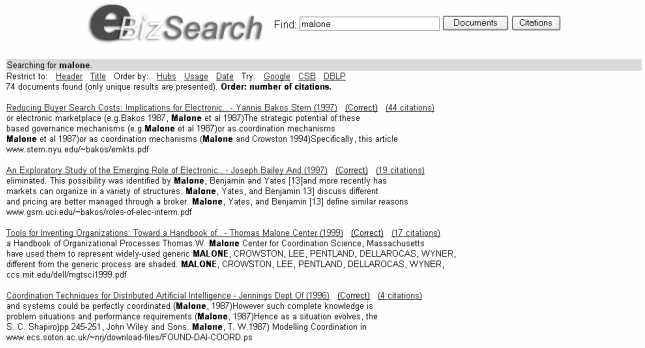

Figure 3: Results for document search

Figure 4 shows a document page for a paper hosted by eBizSearch.

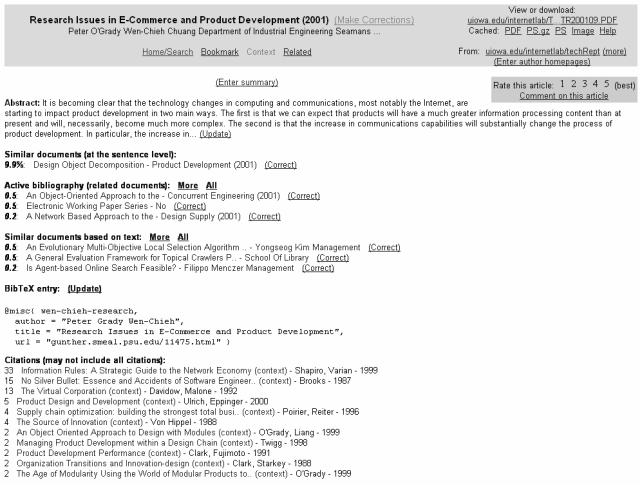

Figure 4: Document page.

Finally, Figure 5 shows typical result display when querying the citation database.

Figure 5: Results for citation search.

Originally the main interface of CiteSeer was to be adopted for eBizSearch. However due to the fact that simple interfaces encourage people to be at ease about using the search engine, a simpler interface was developed, that strongly draws from the search engine Google. The navigation by expected citations (along with the listing of the expected citations) was disabled, as it appears to confuse the user. Note also that the main form page, originally dynamic, was replaced by a static web page in order to reduce loads and ensure place holding in case the server (CiteSeer module) becomes temporarily unavailable. Finally the help pages were updated to make them simpler to understand.

The web-interface is currently CGI-based (Perl) and consists of a single script that provides the presentation layer to our system. CiteSeer runs as a server process accessible locally only and the communication between the presentation layer and CiteSeer is TCP/IP based.

4. Interoperability and the Open Archives Initiative (OAI)

As part of a larger scale effort, eBizSearch, as well as CiteSeer, intend to integrate with the Open Archives Initiative project and to comply with the associated communication protocol defined in [31]. In this section we discuss the issues arising from OAI metadata and protocol requirements and draw the roadmap towards a fully integrated OAI support of our system.

Table 4: OAI compatibility (*: automated metadata extraction)

Metadata item	Available in CiteSeer database	Required from OAI compliant library (Dublin Core metadata standard)
Title	Yes*	Yes
Creator	Yes*	Yes
Subject – keywords	No	Yes
Abstract	Yes*	Yes
Contributor	No	Yes
Publisher	No	Yes
Date (archived)	Yes	Yes
Type	Yes	Yes
Format	Yes* – multiple formats available via conversion	Yes
Identifier	Yes*	Yes
Source	Yes*	Yes
References	Yes*	Yes
Referenced by	Yes*	Yes
Language	No – English only	Yes
Full-text document querying	Yes	Not required
Full-text citation querying	Yes	Not required

4.1. OAI Metadata and Protocol Requirements

OAI defines an XML-based access and exploration protocol aiming at standardizing the access to digital libraries on the web [2,25]. The National Science Digital Library [28] program promotes this usage and Citidel is an example of metadata information aggregator [9]. Historically the metadata set of CiteSeer has been

proprietary as well as the protocol to access it: CiteSeer defines its own metadata requirements and communication with clients is HTML based (i.e. web browsers). On the contrary the OAI project is based upon the XML-based Dublin Core metadata standard [12] and the OAI protocol can potentially be used by a wider variety of clients. Our effort in this context consists into enabling OAI-based access to CiteSeer/eBizSearch and, as suggested in Table 4, to extend the information extraction capabilities of CiteSeer as its original set of metadata for each document (and their citations) does not fully cover the set of metadata required by the Dublin Core metadata standard.

4.2. Architectural and Organizational Issues

Enabling OAI access to CiteSeer intrinsically means upgrading the presentation layer (CGI implementation) such that, provided certain flags are passed in the HTTP query string, the output is OAI-Compliant instead of HTML. This represents an easy aspect of the migration to OAI compliancy and will not be further detailed in this document.

A more delicate and contrived task toward compliance is to provide support for additional metadata items and the possibility to query on them. Some metadata items required by the Dublin Core standard are simply not available from the database of CiteSeer, for instance keywords are not currently extracted. Beyond the issue of the availability of a single metadata item, we must also enable queries based on this item (e.g. keyword-based or date-based queries). Here we need to address an organizational limitation of CiteSeer, which only enables full text querying of both the document texts and citation texts, its metadata collection being available only for internal operations by the system (e.g. ranking, linking) but not directly queriable. The two query modes are enabled through two dedicated indices (inverted files). In order to provide good performance, OAI compliance implies that we index on all the metadata items that can potentially be queried on, that is, all the metadata items required by the OAI specification: we need to provide an index linking keywords to documents, dates to documents and so forth. An item such as publication date is originally available from CiteSeer, but the database organization, even internally to the system, does not provide the capabilities of a standard DBMS, and thus does not allow direct querying on the date. Currently the date information can only be used on the web-interface to sort the set of documents/citations retrieved in response to a query.

Three approaches have been considered to provide OAI support to eBizSearch.

- The first approach is static, in the sense that the Dublin Core XML records are generated (periodically or on demand) and persistently stored (files, DBMS, etc.). This approach is mentioned for completeness only, but was not given further attention due to the strongly dynamic and interrelated nature of metadata in CiteSeer: for efficiency, CiteSeer maintains for a given document D, not only the set of documents X_i (i in $1...N_i$) cited by D, but also the set of documents Y_j (j in $1…N_j$) citing D. This is how the document graph is internally maintained. The citation information is contained in the Dublin Core record for a document, and therefore a static approach seems inappropriate since citation information is prone to automated modifications. Finally the presentation layer is inherently dynamic and therefore handling generation of XML records at this level, on-the-fly, provides much more flexibility and consistency without affecting performance: the two remaining approaches derive from these observations.

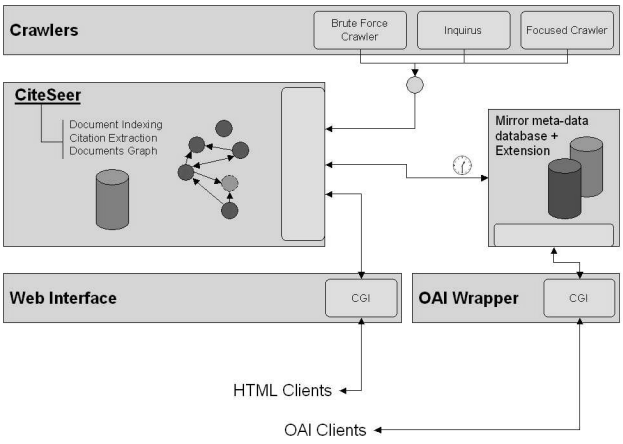

Figure 6: Organization of eBizSearch with non-integrated OAI support

- The second approach, the one currently adopted by eBizSearch, is represented in Figure 6. The level of dynamicity is higher in that XML records generation is performed on the fly. As can be seen, the integration remains poor since the metadata database of CiteSeer is mirrored (external database) and extended in order to address the different requirements induced by OAI compliance. The external database is periodically synchronized with the master database; further metadata extraction is then performed using this mirror. Physically, mostly for isolation purposes during the test phase, we set up an adjunct server to hold the mirror database, this server also supporting an HTTP server dedicated to servicing OAI requests. The URIs that are provided in the XML records are nonetheless pointers to the main server, where documents can be downloaded from. We chose to implement this solution first in order to provide support for OAI as early as possible; this serves as a transient solution which will

be superceded by the third approach upon completion of its implementation.

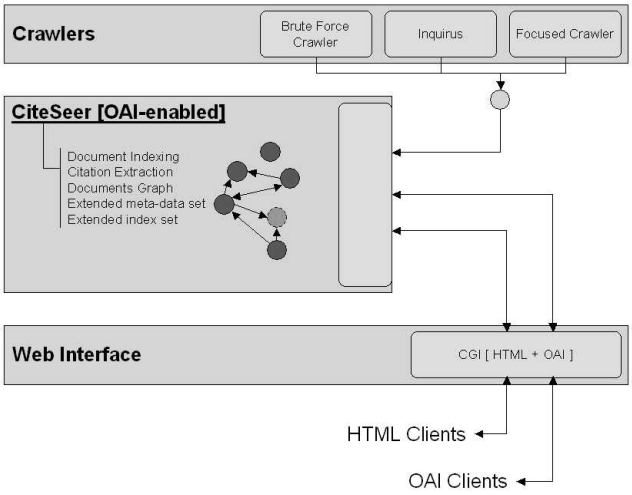

Figure 7: Organization of eBizSearch with integrated OAI support

- In a third approach we intend to extend the set of metadata maintained by CiteSeer itself (see section 5.2), increase the number of indexes in order to extend the queriability of metadata (i.e. date-based queries, keywords-based queries, abstract-restricted queries) and finally upgrade the presentation layer to support the OAI protocol. Ultimately the CiteSeer software package is intended to be shipped in this OAI-enabled form. As can be seen from Figure 7, this approach is an improvement over the second approach wherein all redundancies would have been removed. We are currently improving the CiteSeer to follow this approach.

4.3. OAI Access to eBizSearch

The eBizSearch repository is reachable using the OAI protocols at [14].

5. Extending CiteSeer Capabilities

CiteSeer has been originally developed to meet the needs of the Computer Science research community. As a consequence it makes various assumptions based on the peculiarities of this field and takes advantage of these assumptions to collect, parse and extract information. In this section we present enhancements aiming at improving the reliability of eBizSearch, improving the quality of the metadata extraction and ultimately making CiteSeer-like search engines more portable to other research fields. We provide our most recent experimental results to support the validity of our strategy.

5.1. Extending Document Conversion Capabilities

Document conversion is the corner stone to CiteSeer-like niche digital library: the system must provide support for the conversion of various electronic formats to address the needs of various research communities. As mentioned earlier, CiteSeer performs information extraction using the plain text version of documents. For any electronic format to be supported by CiteSeer, there must exist a converter for documents in that format to plain text. Conversion to plain text of a given electronic format is generally a complex task and is therefore handled by third-parties software libraries. This raises several issues, first the availability of such a software library/service for a given format (this may include platform concerns); second, conversion libraries might fail in converting valid documents (for us, not an uncommon experience). This section presents our setup to combine multiple conversion software libraries to improve the conversion reliability for a single format. We also briefly discuss the adjunction of new conversion software libraries to support additional electronic formats.

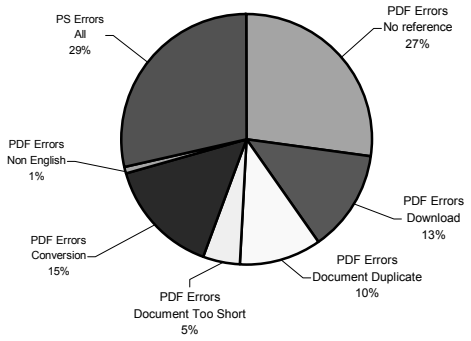

Figure 8: Original breakdown of processing errors

5.1.1. Increasing Conversion Reliability. CiteSeer relies on a single application, *pstotext*, to perform all the conversions needed. Yet, the test phase of eBizSearch revealed that some perfectly valid PDF documents would fail to be converted. Figure 8 shows a breakdown of processing errors of eBizSearch in its original configuration: the failure rate in the conversion of PDF documents is unacceptably high. To improve on the success rate, we conducted an experiment using three freely available tools [27] for conversion of PDF files to text/ASCII files: *pstotext*, *ps2ascii* and *pdftotext*. The applications were called from within a Perl script simulating the working of the applications in the actual eBizSearch/CiteSeer conversion environment. The same set

of valid documents in PDF format was submitted to each application for conversion. The converted documents were then examined as to whether information could be extracted from them after conversion.

Table 5: Distribution of documents across the years and failure count for each application

Publication Year (Number of documents)	Failure Count		
	pstotext	*pdftotext*	*ps2acii*
1999 (10)	2	8	1
2000 (15)	2	2	4
2001 (12)	0	0	2
2002 (21)	3	2	7
Total Failure Count (out of 58)	7	12	14

We were primarily interested in documents from a business source and for this we considered all 58 PDF files of the research documents available from the web site of PennState University's eBusiness Research Center [15] which originated from various institutions and authors. We used the following versions of the applications: *pstotext* (DEC - modified version), *ps2ascii* (Ghostscript, version 7.05), *pdftotext* (xpdf, version 1.01). To emulate the real conversion environment, we limited the time taken for the conversion of each document (a maximum of 4 minutes, otherwise the conversion is considered to have failed): even so, all conversions completed before the deadline and therefore the experiment was not affected by this constraint. The conversion was considered successful when the output text file was human-readable and the conversion of the document was complete (entire text content of the original document excluding any figure or table). Note finally that the option q was used with *pdftotext* to ensure suppression of error messages. The counts of conversion failures for each application are listed in Table 5.

Figure 9 allows a more intuitive visualization of the problem: while it is clear that none of the applications converts all the documents successfully, we can nevertheless increase the number of documents successfully converted by combining several conversion applications.

The justification of conversion failures is not quite obvious, and apparently the causes are diverse. Our experience shows that *pdftotext* has problems with version 1.3 of Adobe PDF files. This is reflected in the high failure rate for documents created in 1999 (mostly created using version 1.3, while only two were created using version 1.4). It is also to be noted that some of the documents were image scans and contributed significantly to the failure of all applications. Finally some failures from *ps2ascii* were due to the fact that the words in the resulting text documents were joined together, making information extraction almost impossible. For further discussion on the suitability of each of these applications for extraction, the reader is referred to [32].

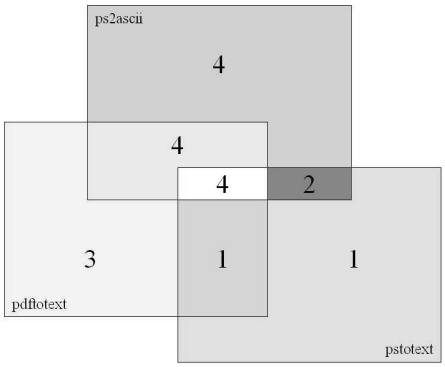

Figure 9: Superposition of conversion failures

Considering these results, a conversion system was adopted for eBizSearch: on completion of the download phase, a document is passed on to a default application for conversion to text, upon failure of the default converter it is passed to an alternative application (if any) and so on until successful conversion, eventually reducing the failure rate conversions (Figure 10). Note also that, since our document conversion system is being constantly upgraded, there is a need to track the efficiency of conversion after each upgrade.

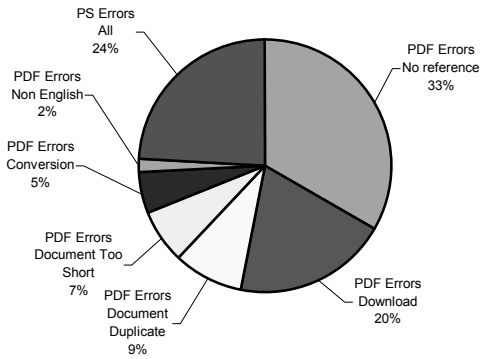

Figure 10: Breakdown of processing errors with conversion enhancement

5.1.2. Supporting Additional Electronic Formats. The large majority of research publications in Computer Science are made available in either PostScript or Portable Document Format formats (or their variants, e.g. compressed PostScript). As a consequence CiteSeer has always been shipped with built-in support for PDF and PS documents only. However the publication of documents using alternative formats, for instance Microsoft Word, has

become non negligible in some fields and it is desirable that a support for a wide variety of electronic formats be added to the CiteSeer package in order to facilitate portability to other domains. In this perspective, we extended CiteSeer with a Word to text converter. Accordingly our crawlers now seek candidate URLs of publications in Microsoft Word format.

5.2. Extending Metadata Extraction Capabilities

We saw in section 4 that OAI requires more metadata items than those currently available from CiteSeer. Moreover, due to the strong focus of CiteSeer on Computer Science publications, metadata extraction has been efficiently achieved by using customized regular expressions. Still the performance of CiteSeer in extracting some metadata items (esp. author(s) and date) turns out to be poor and often requires manual correction. To extend the set of metadata items extracted, 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), a supervised learning and classification method. This algorithm is extensively covered in [18]. For more details about support vector machines, please see [3,21].

This metadata extraction algorithm extracts the 13 metadata items defined in [34] (Title, Authors, Authors' address, Authors' affiliation, Authors' email, Authors' URL, Authors' phone number, Publication Date, Degree of the thesis, Keywords, Abstract, Publication Number and Note) from the header of the research papers. (Title, Author, Keywords, Abstract and Publication Date) are mapped to their Dublin Core metadata equivalent, i.e. (Title, Creator, Subject, Description and Publication Date respectively).

We use the following manually tagged header [34] to illustrate the working of the extraction algorithm.

```
1:<note> Computational Intelligence, Volume 12, Number 3, 1996 +L+ </note>
2:<title> LOCALIZED TEMPORAL REASONING USING SUBGOALS +L+
3:AND ABSTRACT EVENTS +L+ </title>
4:<author> Shieu-Hong Lin, Thomas Dean 1 +L+ </author>
5:<affiliation> Department of Computer Science, Brown University, </affiliation> <address> Providence, RI 02912 +L+ </address>
6:<abstract> We are concerned with temporal reasoning problems where there is uncertainty about the order +L+
7:in which events occur. The task of temporal reasoning is to derive an event sequence consistent with +L+ </abstract>

Note: +L+ is the new-line marker
```

The actual metadata output by our algorithm is given below.

```
1: chunk(1) - <note> - Computational Intelligence, Volume 12, Number 3, 1996
2: chunk(1) - <title> - LOCALIZED TEMPORAL REASONING USING SUBGOALS
3: chunk(1) - <title> - AND ABSTRACT EVENTS
4: chunk(1) - <author> - Shieu-Hong Lin
chunk(2) – <author> - Thomas Dean 1
5: chunk(1) - <affiliation> - Department of Computer Science
chunk(2) - <address> - Brown University, Providence, RI 02912
6: chunk(1) - <abstract> - We are concerned with temporal reasoning problems where there is uncertainty about the order
7: chunk(1) - <abstract> - in which events occur. The task of temporal reasoning is to derive an event sequence consistent with
```

Except for line 5, which is a multi-class line, all the lines are single-class lines. The job of the SVM wrapper is to classify each natural line of the research paper header into one or more classes and then seek the best chunk boundaries of the multi-class lines. The identified chunks are the metadata actually extracted. For example, line 5 contains chunks of affiliation and address that have been successfully identified. We also identify each of the authors in the lines with multiple authors (line 4).

We tested our SVM wrapper on the dataset provided by [34]. This dataset contains 935 headers of computer science papers, with 500 training header and 435 test headers. Our method using the SVM yields a slightly better overall accuracy (92.9) than the Hidden Markov Model (HMM) of [34] (90.1). It is also to be noticed that [34] used additional data, namely 5,000 unlabeled headers (287,770 word tokens) and 176 BibTeX files (2,463,834 word tokens) for training.

We list in Table 6 the class-specific classification accuracy, precision and recall achieved by SVM, and the class-specific classification accuracy achieved by HMM (as reported by [34]). The performance is evaluated based on words. These results supports the fact our SVM metadata extraction algorithm could achieve better performance than HMM for metadata extraction with less training data.

Table 6: Comparison of SVM and HMM performance

	SVM Accuracy	SVM Precision	SVM Recall	HMM multi-state L+D Accuracy
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. Related and Future Work

To our knowledge eBizSearch is currently the only automated digital library 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 IDEAS [20] in the closely related field of economy. A couple of paying or subscription-based portals exist, among which Bitpipe.com [4]. Bitpipe.com provides and sells a common interface to access papers from various analysts group in the field of IT/e-business. From a broader perspective there exist various general-purpose search engines and portals gathering resources on e-business. We can mention searchCIO.com [33] and IBM's e-business homepage [19]. Like most of the e-business resources on the web, they loosely compile materials from various sources. Most of these materials originate from corporations and do not follow the standards of academic publications. As such a major goal of our work is 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.

The portability of eBizSearch/CiteSeer-like search engines will be a key to their success. The limitation to the portability of eBizSearch is the portability of CiteSeer itself: CiteSeer was originally developed to take advantage of its underlying platform in order to optimize the runtime performance, thus the code is very specialized towards its running platform, i.e. Linux. CiteSeer makes extensive use of the UNIX semantics (e.g. system calls, file operations, POSIX) as well as of software packages that are traditionally shipped with Unix/Linux platforms (e.g. conversion software such as *pstotext*). If these considerations do not fundamentally prevent porting the eBizSearch/CiteSeer package to commercial platforms (essentially Windows platform) they nevertheless require an additional effort to unify access to OS-managed resources (conversion libraries, file system, network, etc.). Finally the web-interface (CGI engine) is not handled directly by our system but is instead delegated to an Apache HTTP server, which requires specific configuration of its mod-rewrite module. All in all the setting-up of the application is for now fastidious and requires a good level of expertise from the operator. We currently work toward improving the portability and setting-up of CiteSeer (and thus eBizSearch) while still preserving good runtime performance.

Beyond the platform portability we seek to extend the applicability of our model to other fields of academic research. As outlined in section 5.2, a learning-based approach for metadata extraction should allow us to reach this goal. Ultimately, providing configuration towards a specific area of concentration (in the form of sample publications) will be sufficient to deploy our niche search engine technology in various academic fields.

7. Conclusion

We described a new digital library, eBizSearch, which is a digital library niche search engine based upon CiteSeer technology that has found and indexed over 20,000 documents in e-business. The internal organization of eBizSearch, organized around CiteSeer, was presented, along with a discussion on the migration to a fully-integrated OAI-compliant system. Finally the limitations of CiteSeer in terms of metadata availability, reliability and portability led us to propose a Support Vector Machine machine learning approach for the extraction of metadata. Our initial results show that this method accurately automatically extracts and tags untagged text and has the potential for extending the domain of tagged metadata.

8. Acknowledgements

We acknowledge partial support from NSF NSDL 0121679 and useful comments from Josef Behling, Lillian Cassel, Sandip Debnath, Ed Fox, Aaron Krowne, and Eren Manavoglu.

9. References

[1]: H. Anan, X. Liu, K. Maly, M.L. Nelson, M. Zubair, J.C. French, E.A. Fox, P. Shivakumar, "Preservation and transition of NCSTRL using an OAI-based architecture", In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, pp 181-182, 2002.

[2]: M. Baldonado, C.K. Chang, L. Gravano, A. Paepcke, "Metadata for Digital Libraries: Architecture and Design Rationale", In *Proceedings of the 4th Annual Conference on the Theory and Practice of Digital Libraries*, pp 47-56, 1997.

[3]: K.P. Bennett, C. Campbell, "Support vector machines: Hype or Hallelujah", *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, Explorations 2(2):1-13, 2000.

[4]: Bitpipe.com, http://www.bitpipe.com/.

[5]: K. Bollacker, S. Lawrence, C.L. Giles, "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications", In *Proceedings of the Second International ACM Conference on Autonomous Agents (Agents'98)*, pp 116-123, 1998.

[6]: 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.

[7]: M. Diligenti, F. Coetzee, S. Lawrence, C.L. Giles, M. Gori, "Focused Crawling Using Context Graphs", In *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000),* pp 527-534, 2000.

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

[9]: CITIDEL project homepage, http://www.citidel.org.

[10]: Crane G., "Building a Digital Library: The Perseus Project as a Case Study in the Humanities", In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pp 3-10, 1996.

[11]: Digital Libraries Initiative Phase 2, http://www.dli2.nsf.gov/projects.html.

[12]: "Dublin Core Metadata Element Set, Version 1.1: Reference Description", http://dublincore.org/documents/1999/07/02/dces/.

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

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

[15]: eBusiness Research Center (eBRC) homepage, the Pennsylvania State University, http://www.ebrc.org.

[16]: eCommerce Research Forum at MIT, http://ecommerce.mit.edu/.

[17]: 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.

[18]: H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang and E.A. Fox, "Automatic Document Metadata Extraction using Support Vector Machines", In *Proceeding of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*, In this proceeding, 2003.

[19]: e-Business homepage at IBM, http://www.ibm.com/ebusiness/.

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

[21]: 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.

[22]: S. Lawrence, C.L. Giles, K. Bollacker, "Autonomous Citation Matching", In *Proceedings of the 3rd ACM Annual Conference on Autonomous Agents*, pp 392-393, 1999.

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

[24]: S. Lawrence, K. Bollacker, C.L. Giles, "Distributed Error Correction", In *proceedings of the 4th ACM Conference on Digital Libraries*, p. 232, 1999.

[25]: C. Lagoze and H. Van de Sompel, "The open archives initiative: building a low-barrier interoperability framework", In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2001)*, pp 54-62, 2001.

[26]: C. Lagoze, W.Y. Arms, S. Gan, D. Hillmann, C. Ingram, D.B. Krafft, R.J. Marisa, J. Phipps, J. Saylor, C. Terrizzi, W. Hoehn, D. Millman, J. Allan, S. Guzman-Lara, T. Kalt, "Core services in the architecture of the national science digital library (NSDL)", In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, pp 201-209, 2002.

[27]: C.G. Nevill-Manning, Reed T. and Witten I.H., "Extracting Text from PostScript", In *Software Practice and Experience (SPE) 28(5)*, pp 481-491, 1998.

[28]: The National Science Digital Library community homepage, http://comm.nsdlib.org.

[29]: H. Suleman, A. Atkins, M.A. Gonçalves, R.K. France, E.A. Fox, V. Chachra, M. Crowder, J. Young, "Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress", *D-Lib Magazine 7(9)*, 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html.

[30]: H. Suleman, A. Atkins, M.A. Gonçalves, R.K. France, E.A. Fox, V. Chachra, M. Crowder, J. Young, "Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research", *D-Lib Magazine 7(9)*, 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html.

[31]: "The Open Archives Initiative Protocol for Metadata Harvesting",http://www.openarchives.org/OAI/openarchivesprotocol.htm.

[32]: N. Robinson, "A Comparison of Utilities for converting from Postscript or Portable Document Format to Text", *CERN-OPEN-2001-065*, 2001.

[33]: searchCIO.com, http://searchcio.techtarget.com/.

[34]: 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.