CINDI System

Bipin C. DESAI
Concordia University
Montreal, Canada

Feb. 2008

The advent of the Web has highlighted the importance of information discovery and retrieval as it has become a daily task for most users of the Internet. Search engines have made information search tasks much easier, however they retrieve links to documents based on term frequency, location of terms, link analysis, popularity, date of publication, length of the document, and proximity of query terms. The CINDI System is a digital library(repository) for research papers in domain of computer science. The CINDI project is to improve discovery and search experience by targeting information to that required by academics and professionals in field of Computer Science. This paper describes the CINDI system and its components and our experience with both the push mechanism and the pull mechanism available in CINDI.

Components of the CINDI System

Concordia Indexing Discovery System (CINDI) is a digital library (repository) designed to become a repository for any academic domain but currently we are limiting it to that of Computer Science and Software Engineering. CINDI's mission is to provide a tool for members of the global academic communities to express and explore ideas openly and freely, to acquire and develop the skills of intellectual inquiry, and to examine critically the various beliefs, practices in an attempt to find new insights and resolve outstanding problems for the benefit of all. CINDI's collects the academic documents in two ways:

1. Push Paradigm
2. Pull Paradigm

The initial design of CINDI was based on the user initiated "push" mechanism. Herein, contributing users(authors) would register with the CINDI system and upload their documents. In addition to the user initiated push mechanism, current version of CINDI has opened up two additional channels. The documents in the repository comes from two sources: these are the papers submitted by authors to conferences managed by ConfSys (push paradigm) and documents download from the Web by the CNDRobot (pull paradigm).

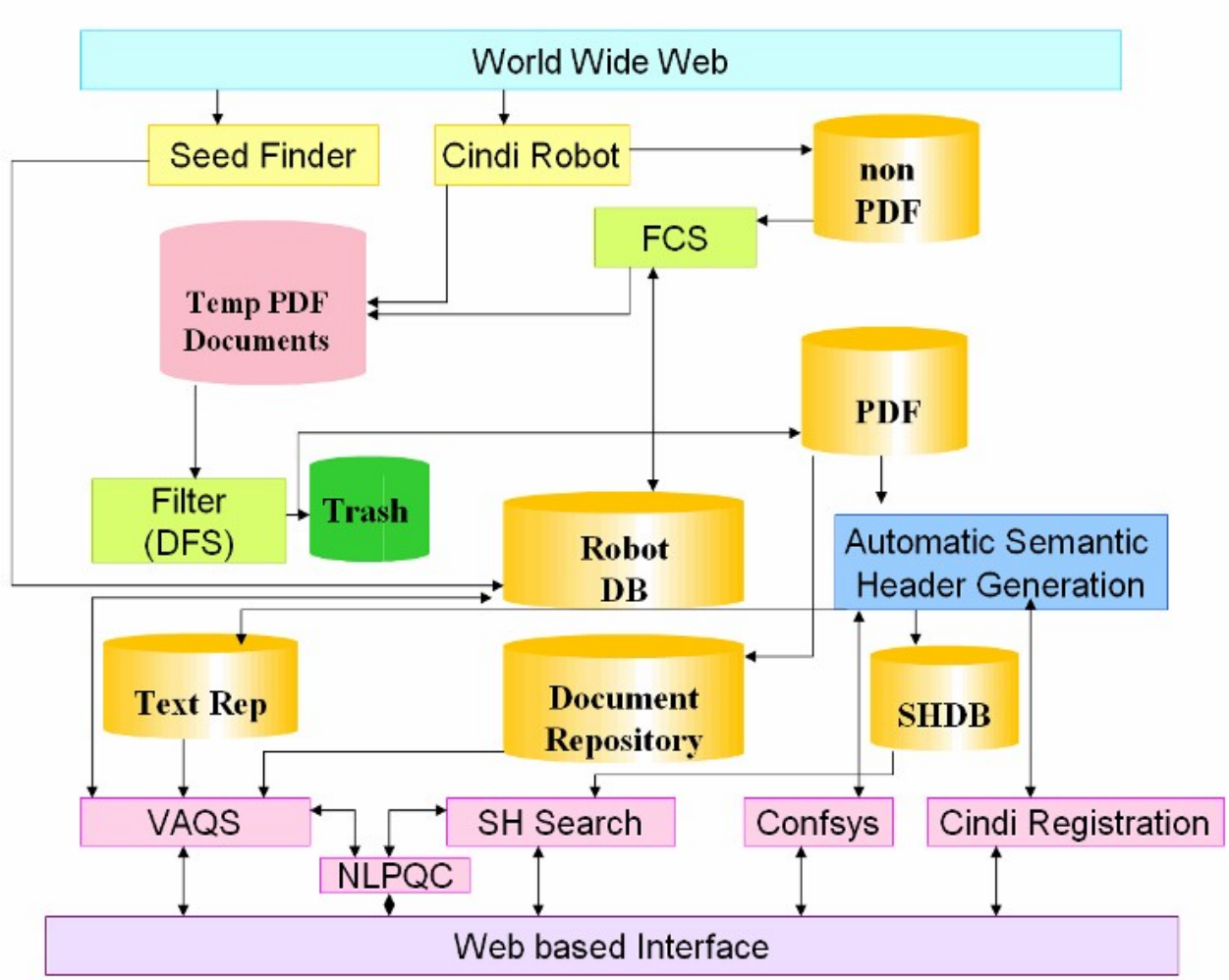
CINDI system includes following subsystems:

1. A registration subsystem to allow the provider of a resource to upload their documents and create and verify an index entry in the form of a Semantic Header(SH) for it using the

Automatic Semantic Header Generator(ASHG) system [Desai, Wang]. The SH is stored in the SHDB database and the document in the Document Collection. A SH search subsystem allows users to locate required information in the SHDB and thence access the source document. These components of CINDI are described in [Yan].

2. A Robot system to download research documents from Web[Zhou]
3. A document converter and gleaning system for converting documents into PDF format and filtering out the documents which are not desired documents[Tong].
4. A conference management system for users to register their resources into CINDI[Feng]

There are two additional components of CINDI consisting of a virtual question answering system for giving the exact answer to the question asked[Zhan] and a natural language interface[Stra] which are not described here.



Architecture of CINDI

1. CINDI Registering and Search System [Yan]

The registering sub-system provides a graphic interface to facilitate the provider of a resource to register the bibliographic information for the resource. The interface allows the provider to upload a resource to CINDI system. The files can be in any of the formats, such as HTML, TEXT, LATEX, RTF, and PDF, an automated mechanism called Automatic Semantic Header Generator (ASHG) [Haddad, Zhang], is used to generate a draft version of the semantic header for the new resource being uploaded to the system. Once this draft semantic header is verified by the resource provider, the information entry can be registered into the SHDB database

During a search request, the client process communicates with the SHDB database and retrieves one or more semantic headers. These results of the query are sent to the user's client workstation. The contents of these headers are displayed, on demand, to the user who may decide to access one or more of the actual resources. In these components the use of a graphical interface and an expert system to model a human reference librarian by providing controlled items for a bibliographic record, the interface guides the users during this operations. Since the scientific world depends on peer review of documents submitted for publication, the annotation sub-system encourages the users to make annotations on the existing resources; these serve the role of reviews and are stored along with the semantic header in the SHDB.

2. CndRobot [Zhou]

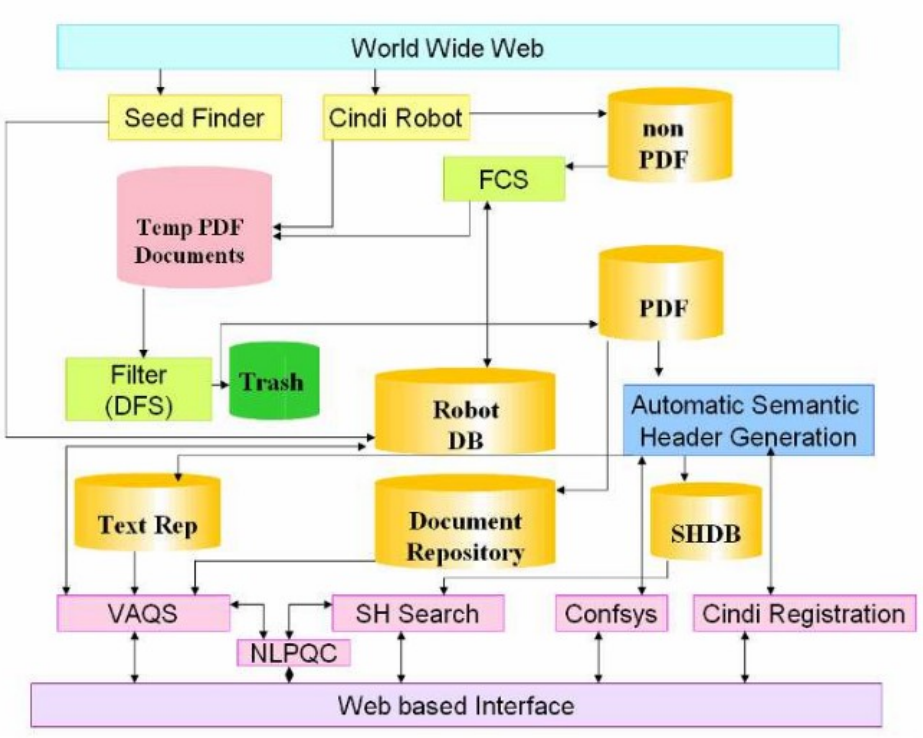


Figure 1. Architecture of CINDI

CINDI's web robot is a focused crawler, which starts with a set of seed URLs generated by the Seed Finder which uses a number of manually generated initial web sites from which other trusted web sites are generated. The sites visited by the robot would thus be trusted sites with high hub and authority scores. CndRobot extracts and follows the hyper links from the Web pages, filters unwanted documents (such as email archives, discussion group, video and audio files etc.) downloads and indexes pertinent documents in different file formats to local repository and revisits the Web pages to maintain current versions of documents and discover new resource.

Robot consists of the following five components: Seed Finder, Web Crawler, File Fetcher, Statistics Analyzer, and Link Analyzer. The Web Crawler works with other components in an interactive way. It first takes the seeds generated by the Seed Finder and then crawl the seed sites to discover potential documents**.** In the crawling process, the Web Crawler performs various tasks e.g. page download, content parsing, and links extraction, etc. It also stores relevant information into various tables in the CINDI database. Some of this information is used by File Fetcher to download documents to local storage. File Fetcher stores information regarding the downloaded documents, their URLs, types, sizes, and location. DFS retrieves the downloaded documents from the local storage; determines their quality; put them into a permanent location or trash; and writes back the decision to the database. The Statistics analyzer and Link Analyzer take into account the feedbacks provided by the DFS while analyzing the previous crawling data so that it can provide more practical suggestions as to improve the performance of the subsequent crawl.

3. Gleaning System [Tong]

Filtering Subsystem

Web search engines have become popular in providing search facility based on a few key words. They return a large number of irrelevant web pages. Document Filtering Subsystem (DFS) as a component of the Gleaning System for CINDI that filters out the garbage picked up by robot. A filter [Chun, Sobo, Kons] can be implemented in three ways.

1. Correlating users' ratings: In this approach, a document is recommended to a user because it is highly rated by other users with whom they tend to agree.
2. Content-based collaborative filtering technique
3. Structure-based filtering Technique: Autonomous Citation Indexing (ACI) system is an example of this technique [Lawr]. CiteSeer is an example of such a system; it collects research documents from the Web and filters them for reference or bibliography sections.

The purpose of DFS in CINDI Robot Subsystem is to filter out irrelevant documents from a set of downloaded files. Unlike content-based filtering agent which needs to know the content of documents and make multiple decisions for categorizing them, the decision process of DFS is binary: accept or reject. Therefore, DFS was developed using a structure-based filtering technique.

When we tested FCS on a set 1003 source documents downloaded by CNDRobot and documents were also manually inspected. The proportion of the accepted documents was 94% and 6% were

irrelevant documents. For the documents rejected, 96% were irrelevant and 4% were relevant. Rejected documents are academic papers from journals whose structure did not meet the ones we programmed into the DFS. Our future tasks is to decrease the acceptance of irrelevant documents while decreasing the percent of rejection of relevant documents by adding heuristics and other rule of thumb that we have uncovered.

File Conversion System

The purpose of FCS is to provide a single document format to facilitate document processing in the subsequent subsystems of CINDI. Since PDF is emerging as the current favorite format for electronic documents, it was chosen as the single format for CINDI. Therefore, non-PDF documents such as TXT, PS, WPD, HTML, DOC, and LaTeX files located and downloaded by CINDI Robot need to be converted into PDF format.

Based on the proposed solution, FCS was developed as an automatic file conversion system. FCS checks the CINDI Robot database on the Linux platform every 30 minutes.

4. ASHG [Haddad, Zhang]

The main steps of using ASHG for Semantic Header generation are: Document Type Recognition, Generation of Key Components and Semantic Header Validation by the author of the document in the push paradigm. Authorship of documents is important information contained in textual data. Document's layout makes it very easy for people to identify the corresponding author(s). However, automatically extracting the authors from machine-readable document is still an open problem. Significant terms play a critical role in ASHG's classification strategy. A set St of significant terms for a document is defined as:

St = {(ti , wi) | ti is a term, wi is a weight}.

The term ti is a word appearing in the title, abstract, and other fields after word stemming, stop word removal operations, and/or one of those which are explicitly labeled as "Keywords" by the author(s). As for the weight associated with term ti , ASHG takes into account its position of occurrence and the frequency of occurrence ASHG's stemming process implements the removal of both suffixes and prefixes of a given word in order to get the word's root[Port].

One of the most important steps in generating a document's Semantic Header is to automatically assign to that document a number of relevant subject headings from CINDI's subject hierarchy. The CINDI's thesaurus is composed of a subject hierarchy and a set of controlled terms associated with the subject headings found in the subject hierarchy. The subject hierarchy is organized as a tree rooted at node root representing the domain of academic subjects.

We have conducted experiments on two document sets to evaluate the ASHG system. The results show close to 100% accuracies on Title, Abstract and Keyword extraction, and 78% for Author. However, the results generated for Subject Headings are somewhat unsatisfactory due to the drawbacks in ASHG's classification method, and decrease in textual quality when documents are converted in to a standard PDF format.

Carefully investigating the issues unsolved above in our future research can significantly improve the accuracy of the subject classification results.

5. ConfSys [Feng]

ConfSys is currently used in “production“ and can be classified as a push component of CINDI. It allows authors to upload papers which would be used in the CINDI digital library's collection. The features of ConfSys are the following: collect papers from authors of an academic conference; allow reviewers to record their preference for papers during an auction process; allocate papers to reviewers to meet the preference of the reviewers while matching the papers’ topics with the reviewers’ expertise and their preference at the same time avoiding conflict of interests; let reviewers download assigned papers, submit review results, and provide features to discuss with other reviewers the review for controversial papers on-line and anonymously; allow authors to get feedback of their papers; collect camera-ready copies of accepted papers from authors for publication; arrange sessions of conference based on topics of papers; manage conference registration process.; upload presentation slides to be used during the meeting and provide support during the actual meeting.

ConfSys has been used for managing the administrative tasks for IDEAS'03, IDEAS'04, IDEAS'04DH, IDEAS05, IDEAS06 and we have provided support also for RTSCA04. The integration of ConfSys with CINDI is under way. We plan to release the system for academic use in the Concordia community soon and thence to the community at large. A demo verson can be accessed at the URL https://confsys.encs.concordia.ca/demo

6. Conclusions

One of the problem being the lack of "marketing" of CINDI. For example, the user initiated 'push'' component of of CINDI has been very minimally used. We have been too busy developing and solving technical problems and left the user to find us. It hasn't worked! CINDI Confsys is another push component of CINDI. It allows authors to upload papers which would be used in the CINDI digital library's collection. Again, this system, in our opinion is a robust system having features still not implemented in other such system. It is available for free hosting of non-profit academic meetings. Unfortunately it has not been noticed by the academic community. This goes to show that academics, in general are not good marketers!

References

[Chun] Chun-sheng L., Cheng-qi Z., Zi-li Z., ``An agent-based Intelligent System for Information Gethering from World Wide Web Environment", IEEE Proceeding of the First International Conference on Machine Learning and Cybernetics, Beijing November, 2002. pp1852-1857.
[Desai] Bipin C. Desai, Rajabihan Shayan Nader, R. Shinghal, Youquan Zhou, ``CINDI: A System for Cataloging Searching and Annotating Documents in Digital Libraries", Library Trends, Summer 1999, 48(1), pp209-233.
[Feng]Yu Wei Feng ``ConfSys: Enhancements \& Integration", Masters Thesis, Webster Library,

Concordia University, TK 5105.884 F46 2004
[Hadd] Haddad, Sami, ``Automatic Semantic Header Generator", Masters Thesis, Webster Library, Concordia University, Z 695.92 H33 1998
[Hara] Sanda Harabagiu, Marius Pasca, Steven Maiorano, ``Experments with open-domain textual question answering", COLING-2000, pp292-298, Association for Computational Linguistics/Morgan Kaufmann, Aug 2000.
[Hara1] Sanda Harabagiu, Dan Moldovan, Razvan Bunescu, ``Answering Complex, List and Context Questions with LCC'¡Çs Question-Answering Server", Tenth Text Retrieval Conference (TREC-10), Gaithersberg, MD. November 13-16, 2001.
[Hovy] Eduard Hovy, Laurie Gerber, Chin-Yew Lin, ``Question Answering in Webclopedia", Ninth Text Retrieval Conference (TREC-9), Gaithersberg, MD. November 13-16, 2000.
[Katz] Boris Katz, ``From sentence Processing to Information Access on the World Wide Web", AAAI Spring Symposium on Natural Language Processing for the World Wide Web, Stanford, California. 1997.
[Kons] Joseph A. Konstan, Bradley N.Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl, ``GroupLens: Applying collaborative filtering to Usenet news", Communications of the ACM, 40(3):77-87, March 1997.
[Korn] Philip Korn, Nicholas Sidiropoulos,Christos Faloutsos, Eliot Siegel, and Zenon Protopapas, ``Fast and Effective Retrieval of Medical Tumor Shapes", IEEE Trans. On Knowledge and Data Engineering, Vol. 10, No. 6, 889, 1998
[Kwok] Cody C. T. Kwok, Oren Etzioni, Daniel S. Weld, ``Scaling Question Answering to Web", Tenth World Web Conference, pp150-161, Hong Kong, China. May1-5, 2001.
[Mill] Miller, A.G. et al. (1995), ``WordNet - A lexical database for the English language", Comm. of the ACM, 38 (1), No., pp. 39-41, ACM Press, New York, ISSN:0001-0782, 1995
[Port] M. F. Porter, ``An Algorithm For Suffix Stripping", Program 14 (1980) pp130-137.
[Rahm] M. M. Rahman, P. Bhattacharya and B. C. Desai, ``A Framework for Medical Image Retrieval using Machine Learning and Statistical Similarity Matching Techniques with Relevance Feedback", IEEE Trans. On Information Technology In Biomedicine (to appear).
[Stra] Niculae Stratica ``A Natural Language Processor for Querying", Masters Thesis, Webster Library, Concordia University, QA 76.9 N38S83 2002
[Wats] M. Watson. Nlbean(tm) version 4: a natural language interface to databases. www.markwatson.com.
[Tong] Tong Zhang, ``A Gleaning Subsystem for CINDI", Masters Thesis, Webster Library, Concordia University, TK 5105.884 Z438 2004
[Wang] Tao Wang, Bipin C. Desai, ``Extracting Document Semantics for Semantic Header", 2006 IEEE Canadian Conference on Electrical and Computer Engineering, Ottawa, May 2006
[Zhang] Zhang, Zhan, ``Cindi's ASHG", Department of Computer Science, Concordia University, 2002
[Zhan] Hong Bing Zhang ``Virtual Question Answering System for CINDI", Masters Thesis, Webster Library, Concordia University, QA 76.9 Q4Z43 2004
[Zhou] Cong Zhou ``CNDROBOT - A Robot for the CINDI Digital Library System", Masters Thesis, Webster Library, Concordia University, LE 3 C66C67M 2005 Z47