Researchers Search Engine On The WEB (RSEWEB) and An Application

Yousef Abuzir[*], Ahmad Nada[**]

[*]Information and Communication Technology Center. ICTC, RamAllah, Palestine
yabuzirr@qou.edu

[**]Computer Information System, Al Quds Open University, Qalqilya Study Center, Palestine
a.nada1@(msn|yahoo|hotmail).com

ABSTRACT

Web is a large and growing collection of texts. This amount of text is becoming a valuable resource of information and knowledge. To find useful information in this source is not an easy and fast task. People, however, want to extract useful information from this largest data repository.

Researchers Search Engine on the WEB (RSEWEB) is a framework for automatic collection and processing of resource related to researchers' information in the World Wide Web. The current RSEWEB implementation searches, retrieves and extracts information about researchers from many servers in the Web and combines them into a single searchable database.

This paper discusses the background and objectives of RSEWEB and gives an overview of its functionality and implementation of RSEWEB system used to construct specialized database about researchers.

The intention is to develop the system to integrate it with other applications such as ThesWB for Advanced Document Management. The system can be utilized in the process of automating conference organization and its usage in real world applications.

Keywords: *Information Extraction, Knowledge disscovery, Web Mining, document Managment*

1. INTRODUCTION

The World Wide Web is a very exciting technology for deploying information on the Internet. In the last years the Web has grown both in terms of its use and the amount of published information.

As a side effect of this expansion it is becoming increasingly difficult to find things in the Web. This problem is known as resource discovery and occurs in any large information system. This paper investigates current methods for discovering resources in the Web, and presents Researchers Search Engine on the WEB (RSEWEB) a framework to aid resource discovery in the Web.

The Web is a rich source of information, but this information is scattered and hidden in the diversity of web pages. Search engines are windows to the web. However, the current search engines, designed to identify pages with specific phrases, have very limited power. For example, they cannot search for phrases related in a particular way such as (books and their authors), (researchers and their interests) and (user, e-mail address).

The Web poses itself as the largest data repository ever available in the history of humankind. Major efforts have been made in order to provide efficient access to relevant information within this huge repository of data. Although several techniques have been developed to the problem of Web data extraction, their use is still not spread, mostly because of the need for high human intervention and the low quality of the extraction results [1, 2, 3, 4, 5, 6,7].

At least two broad views of this problem have evolved recently. The first one, characterized by the unstructured view of data, has developed breakthrough technologies (such as Web search engines) based on information retrieval [8] methods, which have been used in many successful commercial products. The second one, characterized by the structured or semistructured view of data, borrows techniques from the database area to provide the means to effectively managing the data available on the Web [9]. Thus, several techniques have been adapted (or targeted specifically) to the problem of extracting data from the Web for further processing (querying, integration, mediation, etc.) [10]. However, these techniques are still not spread as the information retrieval based ones. This happens mostly because of two problems with these techniques:

- the need for high human intervention and
- the low quality of the extraction results.

Thus, the motivation to develop new methods and tools to allow the effective deployment of a more structured view of the data available on the Web still remains.

In this paper, we present a domain-oriented approach to Web data extraction and discuss its application to automatically extracting information about researchers and their interests from Web sites. Our approach is based on a highly efficient patterns structure analysis that produces very effective results.

The rest of this paper is organized as follows. Section 2 gives an overview of the theory information extraction from the web the basis of our work and related work. Section 3 presents the structure of the RSEWEB, while Section 4 shows the application of this system in the researchers profiles that comprise our approach. Experimental results demonstrating the effectiveness of our approach are in Section 5. Finally, conclusions and directions for future work can be found in Section 7.

2. BACKGROUND OF AUTOMATIC INOFRMATION EXTRACTION FROM THE WEB

Web information extraction involves locating documents and identifying and extracting the data of interest within the documents [11,12]. Information extraction systems usually rely on extraction rules that are tailored to a particular information source. This system is defined as a program or a rule (wrapper) that understands information provided by a specific source and translates it into a regular form as, for instance, XML or relational tables. Information extraction systems are specific to a given Web site and are tightly linked to the mark-up and structure of provider pages. The most challenging aspect of these systems is they must be able to recognize the data of interest among many other uninteresting pieces of text (for example, mark-up tags, inline code, and navigation hints, among others [13]). The simplest information extraction systems utilize extraction rules that are constructed manually. These systems require a human developer to create a new rules (patterns) for each information source or for information sources that are structurally changed. This limits users to accessing information only from predefined information sources.

Currently there are two principal methods for identifying interesting data within Web pages: ontology- based extraction and position-based extraction.

Ontology-based Extraction. - Ontology-based information extraction tools feature many of the properties desired for an adaptive Web information extraction system. An ontology-based tool uses domain knowledge to describe data. This includes relationships, lexical appearance, and context keywords. Wrappers generated using domain ontology are inherently resilient (that is, they continue to work properly even if the formatting features of the source pages change) and general, (they work for pages from many distinct sources belonging to a specific application domain) [6,14].

However, ontology-based tools require that the data be fully described using page-independent features. This means the data must either have unique characteristics or be labeled using context keywords. Unfortunately, all interesting Web data does not necessarily meet these requirements. Some data is freeform and cannot be identified using a specific lexical pattern and also is not labeled. This type of data can only be extracted using its specific location in the HTML page.

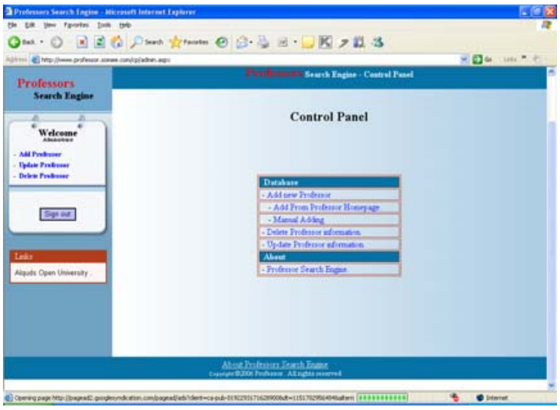
Position-based Extraction relies on inherent structural features of HTML documents to accomplish data extraction. Under a position-based extraction system, a HTML document is fed to a HTML parser that constructs a parsing tree that reflects its HTML tag hierarchy. Extraction rules are written to locate data based on the parse-tree hierarchy. If a collection of items is to be retrieved (as from a search results page), a regular expression is constructed to allow multiple items to be retrieved for a hierarchical pattern. Position-based extraction lacks the resilience of ontology-based extraction. When there are changes to the structure of the target Web pages, it frequently fails. However, it does guarantee a high accuracy of information extraction, with precision and recall being at least 98% [15]. In addition, it is possible to use wrapper induction to create position-based wrappers based on a sample of regularly formatted Web pages. This can greatly speed the development and update of position-based wrappers [16]. Thus, position-based extraction can be appropriate when the data to be extracted can only be identified based on its location within a Web page and not on domain information.

In this paper, we present a domain-oriented approach to Web data extraction and discuss its application to automatically extracting researchers profiles from Web sites. This approach is based on a highly efficient patterns structure analysis and allows not only the extraction of relevant text passages from the pages of a given Web site, but also the fetching of the entire Web site content, the identification of the pages of interest (the pages that actually present the researchers profiles) and the extraction of the relevant text passages discarding non-useful material.

3. THE RSEWEB PROTOTYPE

Devising generic methods for extracting Web data is a complex task, since the Web is very heterogeneous and there are no rigid guidelines on how to build HTML pages and how to declare the implicit structure of the Web pages. Thus, in order to develop effective methods for extracting Web data in a precise and completely automatic manner, it is usually required to take into account specific characteristics of the domain of interest. One of such domains is that of researchers profiles on the Web, which have become one of the most important sources of up-to-date information. Indeed, there are thousands of sites that provide information about researches in the universities and academic institutions in very distinct formats and there is a growing need for tools that will allow individuals to access and keep track of this information in a automatic manner.

One key feature of the RSEWEB system is that both the extraction rules and the output data are represented by XML documents and Database. This approach increases modularity and flexibility by allowing the extraction rules to be easily updated (manually or automatically), and by allowing the retrieved data to either be converted to HTML for consumption by a human or returned in as part of a Web Service. The current RSEWEB prototype represents a cost-effective approach to developing large-scale adaptable information extraction systems for a variety of domains. Figure 1 shows the main interface of the RSEWEB system. The RSEWEB system structure , shown in Figure 2, consists of several modules:



Professors Search Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.professor.somee.com/cp/admin.aspx
Go Links

Professors Search Engine - Control Panel

Professors
Search Engine

Welcome
Administrator

- Add Professor
- Update Professor
- Delete Professor

Sign out

Links
Alquds Open University .

Control Panel

Database
- Add new Professor
- Add From Professor Homepage
- Manual Adding
- Delete Professor information
- Update Professor information

About
- Professor Search Engine

About Professors Search Engine
Copyright ©2006 Professor . All rights reserved

Opening page http://pagead2.googlesyndication.com/pagead/ads?client=ca-pub-0192293171628900&dt=1151702956494&altern

Figure 1 Main interface for RSEWEB

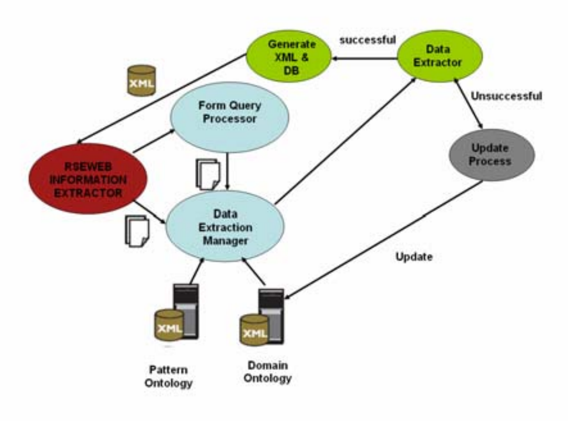
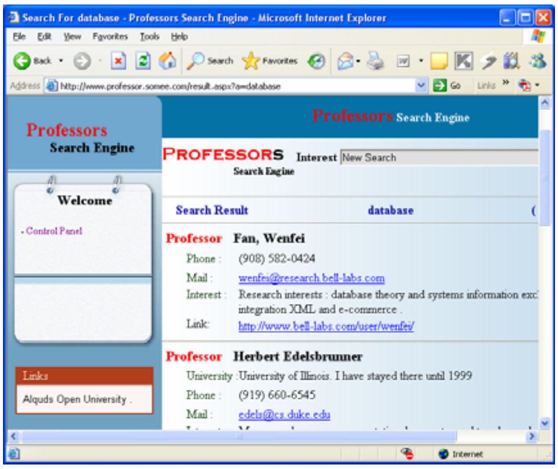


Figure 2: RSEWEB Structure.

The ***query engine*** creates a user query by parsing the site's search form, combining the user query with the site's form elements, and sending the resulting search parameters to obtain the HTML search result pages Figure 3.



Search For database - Professors Search Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.professor.somee.com/result.aspx?a=database
Go Links

Professors Search Engine

Professors
Search Engine

PROFESSORS
Search Engine

Interest New Search

Welcome

- Control Panel

Links
Alquds Open University .

Search Result database (

Professor Fan, Wenfei
Phone : (908) 582-0424
Mail : wenfei@research.bell-labs.com
Interest : Research interests : database theory and systems information exc integration XML and e-commerce .
Link: http://www.bell-labs.com/user/wenfei/

Professor Herbert Edelsbrunner
University : University of Illinois. I have stayed there until 1999
Phone : (919) 660-6545
Mail : edels@cs.duke.edu

Figure 3 RSEWEB Query Engine

The ***data extraction manager*** examines the page structure and determines how best to parse the site. This module analyzes the content of the HTML page, and constructs extraction rules using the domain knowledge. The extraction rules are used to locate data of interest (tokens) within the HTML page.

The ***data extractor*** pulls the specific data from the HTML pages. Following completion of the data preprocessing steps, a data extraction process is initiated to locate the tokens within the parsed HTML document. The RSEWEB system uses a three-step location process to correctly identify and extract the tokens. The system searches the set of location key/ content-text pairs generated by the HTML parser (and if appropriate grouped in the record separation process) for any of the keywords defined. If a keyword represents a path expression, this indicates that position- based extraction is being used and the location key is used for the search. Otherwise, the data is being extracted using ontology-based extraction and the content-text is searched to locate the keyword. Once a keyword is located, the RSEWEB data extraction module searches the content-text before and after the keyword location to find data that matches a pattern defined in the domain ontology. When Web data containing an appropriate pattern is found, the data type is used to extract the desired token from the content text. The extracted data is enclosed in relational database and XML tags and returned as a single XML record.

The ***Update Processor*** system is invoked when the RSEWEB system cannot locate tokens within the Web pages. When the data extraction module cannot locate a required token using the standard extraction procedures, the basic recovery strategy is to locate the token using a new set of keywords. The Update Processor system uses a thesaurus to generate additional keywords to locate the data of interest. In the RSEWEB system a thesaurus entry represents a special pattern set that can be defined for any word or group of words. These word patterns may then be used to replace a single word or set of words found in a keyword list for a domain. After new keywords are generated, a location process is used to identify candidate Web data that could be the tokens of interest. The set of location-key/content-text pairs generated by the HTML parser is scanned to identify content-text that contains one of the thesaurus-generated keywords. When a keyword is located, candidate tokens are identified by searching the content-text before and after the keyword to find the first occurrence of a pattern defined in the domain ontology.

A prototype RSEWEB information extraction system has been implemented using Microsoft Visual C#.NET.

The RSEWEB data extraction process and update processor was tested in the online Computer Science domain. In the proof of concept demonstrations, an XML ontology was developed for the online Computer Science domain. Since the domain ontology allows more than one set of keywords and more than one pattern to be specified, it can be used to extract data from several different Web sites in the computer science domain. Once appropriate domain ontology were created, the RSEWEB information extraction system was used to extract data from the online Internet computer science or universities sites. The prototype RSEWEB system showed excellent performance for different Web sites tested.

4. AN APPLICATION

Web information services and Web business intelligence applications would have access to structured information that could be easily extracted and incorporated into their value-added services, but currently this is not the case. The Web provides access to an enormous volume of semi-structured HTML data in a variety of ever-changing formats. This presents several major challenges to developers interested in using Web data in their applications: First, HTML documents containing interesting data must be located. Second, data of interest must be located within the Web page and rules that can be used to reliably extract the data must be created. Third, the mechanism used to create data extraction rules must either be sufficiently general or be easy to implement so that data can be extracted from the wide variety of page formats available on the Web. Finally, the information extraction system must be able to cope with changes to Web page structure since Web content providers frequently change the configuration and content of their pages. Figure 2 illustrates the adaptive information extraction process envisioned in this research.

To see who we can use our RSEWEB in a real application, we integrated this system with ThesWB [6]. The application is related to automatic conference organization.

In organizing and delivering electronic articles by their similarity and classifying them undergoes restrictions. The job has to be done fast, for instance managing the flow of the articles coming in to the organizer of conference or chairman of a session. A thesaurus-based classification system can simplify this task. This system provides the functions to index and retrieve a collection of electronic documents based on thesaurus to classify them. The thesaurus is used not only for indexing and retrieving messages, but also for classifying. By automatically indexing the electronic articles using a thesaurus, conference organizer can easily locate the related articles and find out the topic. By automatically indexing the articles based on a thesaurus, the system can easily selects relevant articles according to user profiles and send an email message to the reviewer containing the articles.

4.1 USER PROFILING

User profile is a collection of information that describes a user [17]. User profile may be defined as a set of keywords which describe the information in which the user he is interested in. Similarly, some approaches base the user profile on the user's likes and dislikes [18].

With user profile the user can set certain criteria of preference and ask for articles of specific. The profiling can be done by user-defined criteria in our case here by collecting keywords (interests) from researcher's web pages that extracted by RSEWEB. With the use of classification techniques based on thesaurus, the articles adapts to reviewers' (researchers') needs and interests according to his/her profile.

The system seeks to map the indexed articles with the reviewer interests to choose a subset of articles that best reflects reviewer interests. User's interests are represented as a profile as described above.

4.2 USER PROFILE CREATION

User Profile of the researchers constructed as follows Fig 4. At first email address, interests, address of the researchers, and the other fields are extracted from the WEB using RSEWEB. The body part of the selected web pages will be extracted and parsed by ThesWB system. As a result of indexing process all keywords from the web pages, email, address and other information are stored in a database reflects user interests.

Thesaurus is used to create the user-profile from web pages. These web pages include many common terms appeared in the thesaurus. Moreover, these terms reflect the interest of researchers.

User profiles store the interests of a given researcher. The user profiles generated automatically from the web pages. The profile is hierarchically structured, and they are not just a flat list of keywords.

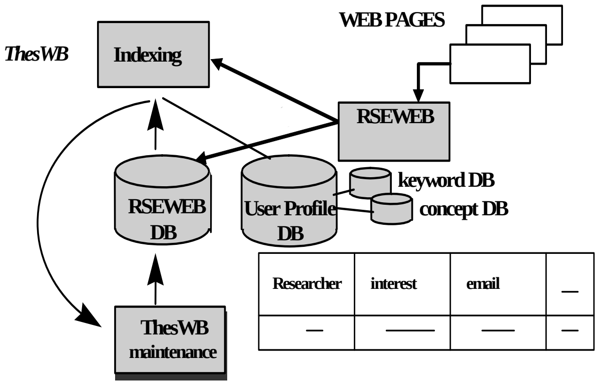


Figure 4: User-Profile creation using RSEWEB and ThesWB.

4.3 COSTRUCTING THE CONFERENCE THESAURUS

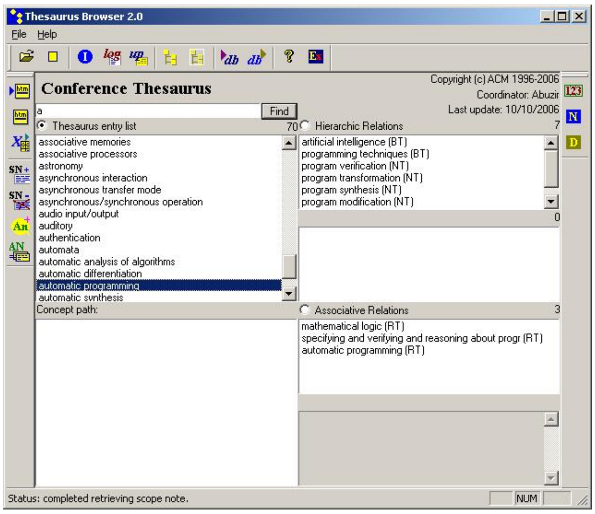
In order to turn Information Retrieval systems into more useful tools for both the professional and general user, one usually tends to enrich them with more intelligence by integrating information structure, such as thesauri. Since it is difficult and expensive to build thesauri manually, many researchers attempted to construct thesauri automatically.

There are two approaches to construct a thesaurus. The first approach, on designing a thesaurus from document collection, is a standard one [19,20]. By applying statistical [21] or linguistic [22], [23] procedures we can identify important terms as well as their significant relationships. The second approach is merging existing thesauri [24], [25]

Our experiment to construct the Conference Thesaurus is based on selecting web pages that are well represented of the domain. The web pages we selected were sample of pages relate to call for papers. We start by parsing those web pages using ThesWB. The parsing process will generate list of terms represented in hierarchical structure. ThesWB search for the string “will include, but are not limited to” in these pages to find the appropriate structure to construct the hierarchical relationships. The main tags were used in this process were <DIV> and <P> We used tab to represent this structure in flat file. The second step is to eliminate and remove noisy terms and relationships between terms from the list. These terms would not appear in the thesaurus. The list in Fig. 5 shows a sample of the new list after removing the noisy terms. Later, we convert this list to the thesaurus using ThesWB Converter Tool.

Using ThesWB, terms and relationships are extracted automatically and stored in ASCII file. However, this file is general and need manual constraints. The system did not prevent the extraction of pairs of terms that are not linked by the target relation.
The Conference Thesaurus has been designed primarily to used for indexing and classification of articles sent by emails for reviewers of a conference for evaluation. This thesaurus provides a core terminology in the field of Computer Science. In addition to indexation and classification, it can be used for terminological guide for the standardization of descriptors, and search aid with other subject-related vocabularies in the databases of the electronic articles. By proving translation of the terms in the thesaurus into other languages, it can be used for linguistic equivalencies for translation purposes.

Fig. 5 shows the Conference Thesaurus. The thesaurus may be viewed here in a hierarchical or alphabetical display. The alphabetical version presents all preferred and non-preferred terms in a single alphabetical sequence. Next to the alphabetical list with relations, there is a graphical tree viewer. This viewer is mainly intended to show terms’ path to the root (it shows the position of a term in relation to the root of the thesaurus, including the path of all parent terms).



Conference Thesaurus

Copyright (c) ACM 1996-2006
Coordinator: Abuzir
Last update: 10/10/2006

a

● Thesaurus entry list 70

associative memories
associative processors
astronomy
asynchronous interaction
asynchronous transfer mode
asynchronous/synchronous operation
audio input/output
auditory
authentication
automata
automatic analysis of algorithms
automatic differentiation
automatic programming
automatic synthesis

Concept path:

○ Hierarchic Relations 7

artificial intelligence (BT)
programming techniques (BT)
program verification (NT)
program transformation (NT)
program synthesis (NT)
program modification (NT)

0

○ Associative Relations 3

mathematical logic (RT)
specifying and verifying and reasoning about progr (RT)
automatic programming (RT)

Status: completed retrieving scope note.

NUM

Figure5. Conference Thesaurus. The draft version contains 141 terms.

In this paper we used the thesaurus to classify the articles into concepts "subject hierarchy". In our application, all articles are classified into concepts. Our classification approach uses Conference Thesaurus as a reference thesaurus. Each article is automatically classified into the best matching concept in the Conference Thesaurus. ThesWB system gives weight for each concept. This weight can be used as selection criteria for the best matching concepts for the articles.

In the creation of users profile that shows reviewer’s interest we used ThesWB to parse and index the email of the author. Later, the system determines which articles in the input sets are relevant and which are not. Comparing the articles to a list of key words that describe a reviewer to classify the articles as relevant or irrelevant, Fig. 6. The system uses the user profile to nominate an appropriate reviewer for each article.

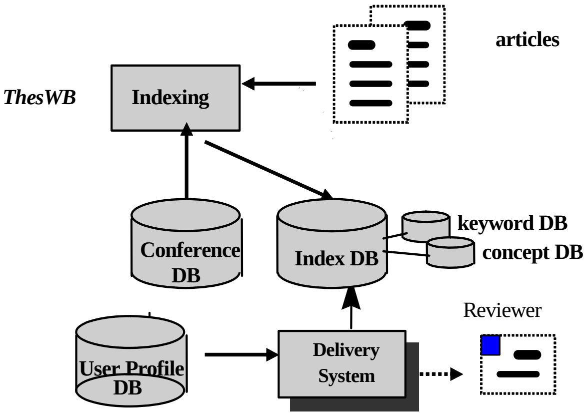


Figure. 6. An overview of the Conference articles Delivering-Service.

5. AN EXPERIMNET AND EVALUATION

While there are many uses for thesauri, this work is aimed at exploring their application to articles classification. Electronic thesauri have been identified as a strategic instrument for indexing electronic documents [19].

After the Conference Thesaurus was constructed, its performance is tested. Our articles collection was indexed by ThesWB thesaurus-based indexing engine. This automatically created keywords and concepts.

In order to classify articles a thesaurus is used, the thesaurus can be used to reflect the interests of a reviewer as well as the main topic of the article. The thesaurus is used not only for indexing and retrieving messages, but also for classifying articles.

We evaluated this classification method using ThesWB Toolkit and a collection of articles. The collection includes a list of abstract and keywords relate to the Database and expert system domain in general

The classification mechanism in our approach is based on Information Retrieval Thesaurus. The ThesWB Toolkit parses the articles and indexes the articles using a thesaurus. Thereafter the user/organizer can use Document Search environment to retrieve the related article

The task of the delivery system is to get a collection of articles to be delivered to a user. The ultimate goal the system is to select articles that best reflect reviewer's interest. The articles undergo pre-processing. We used ThesWB as tool to index these articles. The indexing process being used by ThesWB are based on the hierarchical structure of the thesaurus. The thesaurus hierarchy is used in order to create association between the article and the concepts in user profile. A user profile consists of one or more topics. Topics represent reviewer's information and interests.

We can use the result of indexing to classify the articles according to the main root terms or other concepts that reflect the different topics. Reviewer interest will be match to these database results to select the articles that reflect his/her interest. We used VC++ API application to map the interest of each user to the indexing result and get the articles the best reflect his/her interest. Then, the system will delivery these articles by electronic mail to that reviewer.

The proposed approach has already been put to practice. A sample of 20 articles was automatically indexed using Conference Thesaurus. The results were manually evaluated. The test results showed that a good indexing quality has been achieved.

We get articles from a cache directory. The batch process to index all the articles takes about 15 second. It takes about 0.75 seconds to classify each article compared to 1-2 minutes human indexer needs.

ThesWB, thesaurus based indexing provides the fully automatic creation of structured user profiles. It explores the ways of incorporating users' interests into the parsing process to improve the results. The user profiles are structured as a concept hierarchy. The profiles are shown to converge and to reflect the actual interests of the reviewer.

In this application, we describe the approach of using thesaurus for articles classification and distribution. Experimental results analysis of our approach shows excellent performance and good managing of article. The use of thesaurus is effective for classification problem.

6. CONCLUSION

The use of external information for business decision making is not new. What is new is the abundance of information freely available via the Internet. However, this information is not being systematically included in current decision-making applications. This research demonstrates that it is possible to reliably extract Web information for use in Web base intelligence applications. It will be possible for organizations to use a system like RSEWEB to extract information of interest from Web pages for a wide variety of domains. These potential intelligence applications will allow a deep and detailed look at small portions of the Web relevant to specific domains.

The RSEWEB system has been used to extract data relevant to the study of online computer science. In the future, similar systems can be used to extract data related to financial markets, online travel, or benefits, just to name a few. Use of an information extraction system, like RSEWEB, has the potential to provide businesses with access to up-to-date, comprehensive, and ever-expanding information sources that can in turn help them in their applications..

To support this approach, we have developed a RWEWEB prototype. We have tested our approach with several important international universities and academic or research institutions sites and achieved very precise results, correctly extracting 87.71% of the researchers profiles in a set of web pages.

The increasing number of article sent to the conference presents a rich area, which can benefit immensely from our system RSEWEB and automatic classification approach. We present an approach of automated articles classification through thesaurus for the purpose of developing an article delivering-service system.

Experimental result of our approach shows that the use of thesaurus contributes to improve accuracy and the improvements offered by classification method.

To summarize, automatic articles classification is an important problem nowadays. This paper proposes an approach base on thesaurus to classify and distribute the articles. The experimental results indicate accurate result.

REFERENCES

[1] S. Brin. Extracting patterns and relations from the world wide web. In WebDB, Valencia, Spain, 1998.

[2] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-speci_c web resource discovery. In Proc. of The 8th International World Wide Web Conference, Toronto, Canada, May 1999.

[3] A. Arasu, H. Garcia-Molina, and S. University. Extracting structured data from web pages. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pages 337.348. ACM Press, 2003.

[4] L. Arllota, V. Crescenzi, G. Mecca, and P. Merialdo. Automatic annotation of data extraction from large Web sites. In Proceedings of the International Workshop on the Web and Databases, pages 7.12, San Diego, USA, 2003.

[5] V. Crescenzi, G. Mecca, and P. Merialdo. Wrapping-oriented classi_cation of Web pages. In Proceedings of the 2002 ACM Symposium on Applied Computing, pages 1108.1112. ACM Press, 2002.

[6] Abuzir, Y. and Vandamme, F., "ThesWB: A Tool for Thesaurus Construction from HTML Documents", Accepted in Workshop on Text Mining Held in Conjunction with the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taipei, Taiwan, May 6, 2002

[7] Y. Abuzir, D. Vervenne, P. Kaczmarski and F. Vandamme, Extracting Semantic Relationships

between Terms using IKEM Tool, KIM/KIT NEWS, Vol. 15, nr.3, Nov.2000.

[8] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, Harlow, England, 1st edition, 1999.

[9] D. Florescu, A. Levy, and A. Mendelzon. Database techniques for the world-wide web: a survey. SIGMOD Rec., 27(3):59.74, 1998.

[10] A. Laender, B. Ribeiro-Neto, A. Silva, and J. S. Teixeira. *A brief survey of Web data extraction tools*. SIGMOD Record, 31(2):84.93, 2002.

[11] Gregg, D. and Walczak, S. Exploiting the Information Web. IEEE Trans. on System, Man and Cybernetics Part C 2006.

[12] Knoblock, C., Leramn, K., Minton, S., and Muslea, I. Accurately and reliably extracting data from the Web: A machine learning approach. Bulletin IEEE Computer Society Technical Committee on Data Engineering 23, 4 (2000), 33–41.

[13] Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S., and Teixeira, J.S. Surveys: A brief survey of web data extraction tools. ACM SIGMOD Record 31, 2 (June 2002), 84–93.

[14] Embley, D., Campbell, D., Smith, R., and Liddle, S. Ontology-based extraction and structuring of information from data-rich unstructured documents. In Proceedings of the Conf. on Info. and Knowledge Management (Nov. 1998), 52–59.

[15] Chidlovskii, B. Automatic repairing of Web wrappers by combining redundant views. In Proceedings of IEEE Conf. Tools with AI (Nov. 2002), 399–406.

[16] Muslea, I., Minton, S., and Knoblock, C. A hierarchical approach to wrapper induction. In Proceedings on Autonomous Agents (1999), 190–197.

[17] Jovanovic, D., A Survey of Internet Oriented Information Systems Based on Customer Profile and Customer Behavior, SSGRR 2001, L'Aquila,, Italy , Aug06 12 2001.

[18] Krulwich, B., ‘Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data,’ Al Magazine, summer, 1997.

[19] Salton, G., McGill, M. J., Introduction to modern information retrieval. McGraw Hill, New York. 1983.

[20] Crouch, C. J., An approach to the automatic construction of global thesauri, Information Processing & Management, 26(5): 629-40, 1990.

[21] Qiu, Y., Frei, H.P., Concept Based Query Expansion. Proc. of the 16th Int. ACM SIGIR Conf. on R&D in Information Retrieval, Pittsburgh, SIGIR Forum, ACM Press, June 1993.

[22] Grefenstette, G., Use of syntatic context to produce term association lists for text retrieval. In SIGIR'92, pp. 89-97, 1992.

[23] Ruge, G. Experiments on linguistically based term associations. In RIAO'91, pp. 528-545, 1991.

[24] Sintichakis, M. and Constantopoulos, P. A Method for Monolingual Thesauri Merging. In Proc. 20th International Conference on Research and Development in Information Retrieval, ACM SIGIR, Philadeplphia PA, USA, July 1997.

[25] Mili, H., Rada, R. "Merging Thesauri: Principles and Evaluation". IEEE Transactions On Pattern Analysis and Machine Intelligence,10(2):204-220, 1988.