Academic Researcher Information Extraction from the WEB (ARIEW)

Yousef Abuzir and Sondos kittane

Faculty of Technology and Applied Sciences/ Al-Quds Open University, Al-Bireh, RamAllah, Palestine
yabuzir@qou.edu ,sondos_kittane@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.**

Academic Researcher Information Extraction from the WEB (ARIEW) is a framework for automatic collection and processing of resource related to researchers' information in the World Wide Web. ARIEW 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 ARIEW and gives an overview of its functionality and implementation of ARIEW system used to construct specialized database about researchers.

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

Experimental results show that our approach to researcher profiling significantly gives accurate result and performance. The methods have been applied to find related researcher. Experiments show that the accuracy of researcher finding were significantly improved by using the proposed methods.

***Keywords*— Information Extraction, Knowledge discovery, Web Mining, document Management, Agent, Crawl**

1. Introduction

The enormous growth of the World Wide Web in recent years has made it important to perform resource discovery efficiently. It is often difficult to find useful information from thousands or hundreds of WebPages for a researcher or junior students. This procedure is time consuming even for sophisticate. Academic search engine, such as Google Scholar (Harzing, A. and R. van der Wal. 2008; Kloda, L. 2007), becomes an interesting and promising topic in recent years. However most search engines return the results to users by a list and users must scan each item/webpage one by one in order to collect and re-organize these information based on users' requirements.

Information on WEB creates difficulty for filtering relevant information for decision. A researcher may know the location of such information but has to periodically access the information using direct manipulation and navigation tools. Even if the access is automated, it is still very difficult for the researcher to select relevant information. We are motivated to develop an intelligent agent to retrieve Academic Researchers' profiles on the Internet. Extracting information about academic researcher which would interest a user is difficult. Many existing agents constructed a user profile to filter and extract the user interests. In this research, an intelligent agent is developed to extract user preference and build a database to store these information for further queries.

This paper presents an academic search engine, which is developed as an efficient tool to construct researcher's profile automatically. Moreover, some searching and indexing methods, text mining and computational linguistics for underlying this problem are exploited.

1. Related Work

WEB presents a huge resource of useful unstructured information and knowledge which makes it difficult to extract and retrieve a relevant data from those sources. Therefore, there is a great necessity for information extraction (IE) systems that extract information from the Web pages and transform into program-friendly structures such as a relational database. Many approaches for data extraction from Web pages have been developed. (Chang et al., 2006) present a survey of the major Web data extraction approaches and they suggest three criteria that provide qualitatively measures to evaluate various IE approaches. These three criteria are:

- **The task domain** explains why an IE system fails to handle some Web sites of particular structures
- **The techniques used** to classify IE systems based on the techniques used.
- The third criterion is **the automation degree** for IE systems.

Information extraction is an important task with many practical applications, and many research efforts have been made so far. Many text or/and web applications like Opinion mining from noisy text data (Dey, L. and Haque, Sk. M., 2009), Social Network Extraction of Academic Researchers (Tang J. et al 2007; Tang J. et al 2008), contact information search, question answering (Maiorano S., 2006)., integration of product information from websites (Li, L. et al 2007; Yang Z et al 2007; Yang Z et al 2010), biomedical text mining (Lourenço, 2009; Kheau, 2011), and removal of the noisy data benefit from information extraction are applications of information extraction. In these researches different methods and techniques were used. For example, rule learning based method, classification based method, and sequential labeling based method are the three state-of-the-art methods (Tang et al., 2007b). Fig 1 gives a summary of these different information extraction methods.

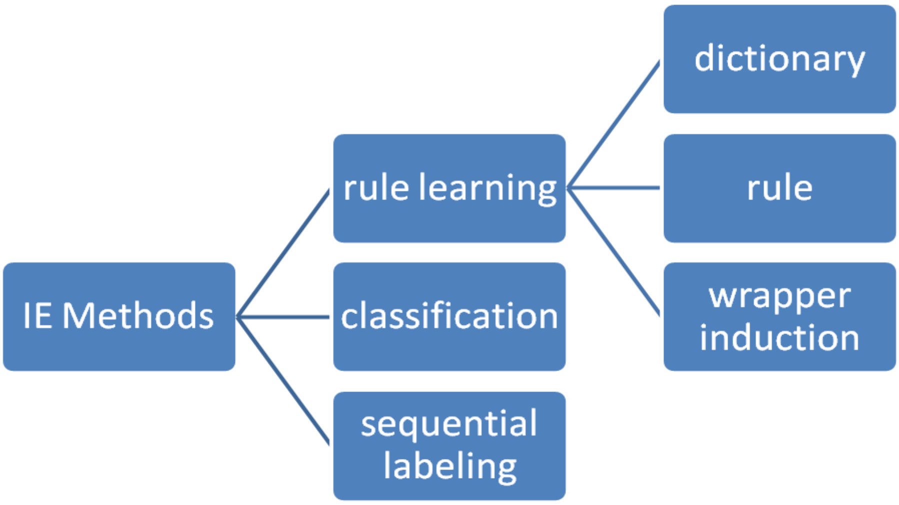


Fig 1 The Different Information Extraction Methods.

Downey (Downey D., et al 2004) introduced a Web information extraction system (KnowItAll) that uses the learned patterns as both extractors (to generate new information) and discriminators (to assess the truth of extracted information). Experimentally, by using learned patterns as extractors, they were able to boost recall by 50% to 80%; and by using such patterns as discriminators they were able to reduce classification errors by 28% to 35%.

Other work used Conditional Random Fields (CRF) model (Liu W. and Zeng J., 2011) to extract academic papers from web pages. The extensive experiments show the effectiveness of the approach. They improved the extraction performance by exploiting the neighboring relations among the academic paper properties. Multi-Theoretical Multi-Level (MTML) (Zarandi M. F., et al, 2011) used as a framework which investigates social drivers for network formation in the communities with diverse goals. This framework serves as the theoretical basis for mapping motivations to the appropriate domain data, heuristic, and objective functions for the personalized expert recommendation.

Another approach based on extracting contextualized user profiles in an enterprise resource sharing platform according to the users' different topics of interest was presented by Schirru (Schirru,R, et al, 2010). Each topic is represented as a weighted term vector.

Extraction Prioritization proposed as automatic technique for obtaining the most valuable extraction results as early as possible in the extraction process (Huang J. and Yu C., 2010). They formally defined a metric for measuring the quality of extraction results, which is suitable for the web retrieval context and developed statistical methods to efficiently estimate the page utilities without launching full-scale extractions.

2. Problem Analysis

Researcher Academic Profile is an important topic in research community. An academic can have different types of information: contact information (including address, email, telephone, and fax number), Academic profile (including homepage, position, portrait, affiliation, research interest, publications, and documents), and social network information (including person or professional relationships between persons, e.g. friend relationship). Figure 2 shows a sample presentation of these information. However, the information is usually hidden in heterogeneous and distributed web pages.

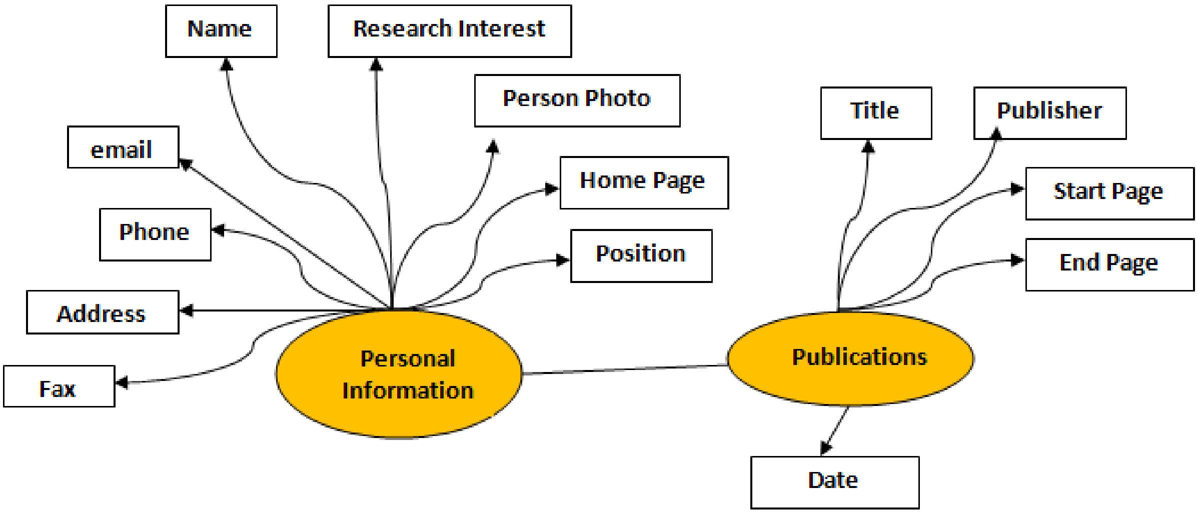


Fig 2 A Sample Presentation of the Researcher Profile’s Schema

Many previous studies have indicated that a researcher’s personal Web site can be considered as identity and self-presentation of the researcher on the Web, and it can be used to different and shows interesting information about the researcher (Doring, 2002). As we have discussed, a researcher’ s Web site usually has information about her/his research interest, publications, research projects, etc., which well represents this researcher.

We have investigated the problem of contact information extraction and academic profile extraction. We have found that the academic information is mainly hidden in person homepage, person introduction page (web page that introduces the person), person list (e.g. a faculty list), and email message (e.g. in signature). We employed the classification based method to extract the person information from the different types of web pages.

The usefulness of a research profile constructor system depends to a large extent on its ability to automatically determine one or more researchers profiles related to the work of interest. Various approaches exist to determine the degree of similarity of related in order to identify related work (Sabbah T. S., et al, 2009).

3. System Architecture

This section describes the architecture of our system which is an Internet agent that gathers information from the Web Pages in order to build a local database of researchers. As an application, we showed a case of an agent building a database with information about academic contacts (phone, email, Postal Address), photos, their interest's researches and publications.

Although, this system used to reduce the effort of the researcher in finding information about other researchers, the system may be used in other applications like indexing and classification, conference management. The database created by the agent was implemented using MySQL.

The system architecture in Figure 3 shows that there are five modules: the module of Concept Crawler, which is responsible for collecting data from WEB; Database module; the module of Researcher Agent; Query Processor and UI module. In the following paragraphs, these components and underlying methods are described one by one.

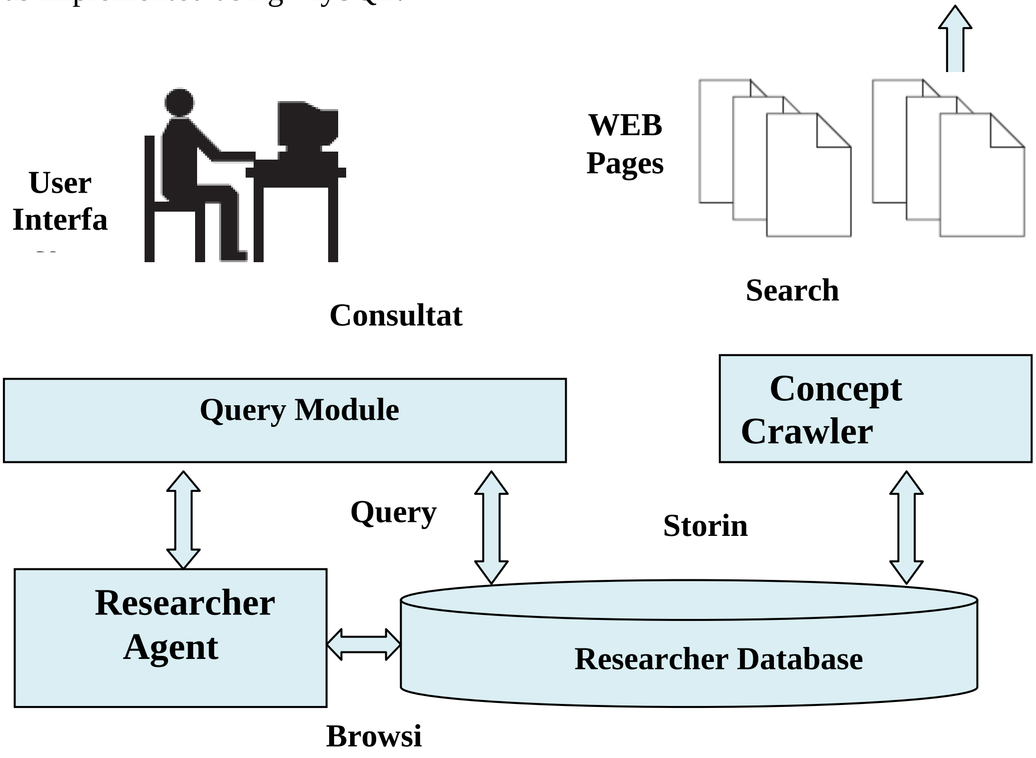
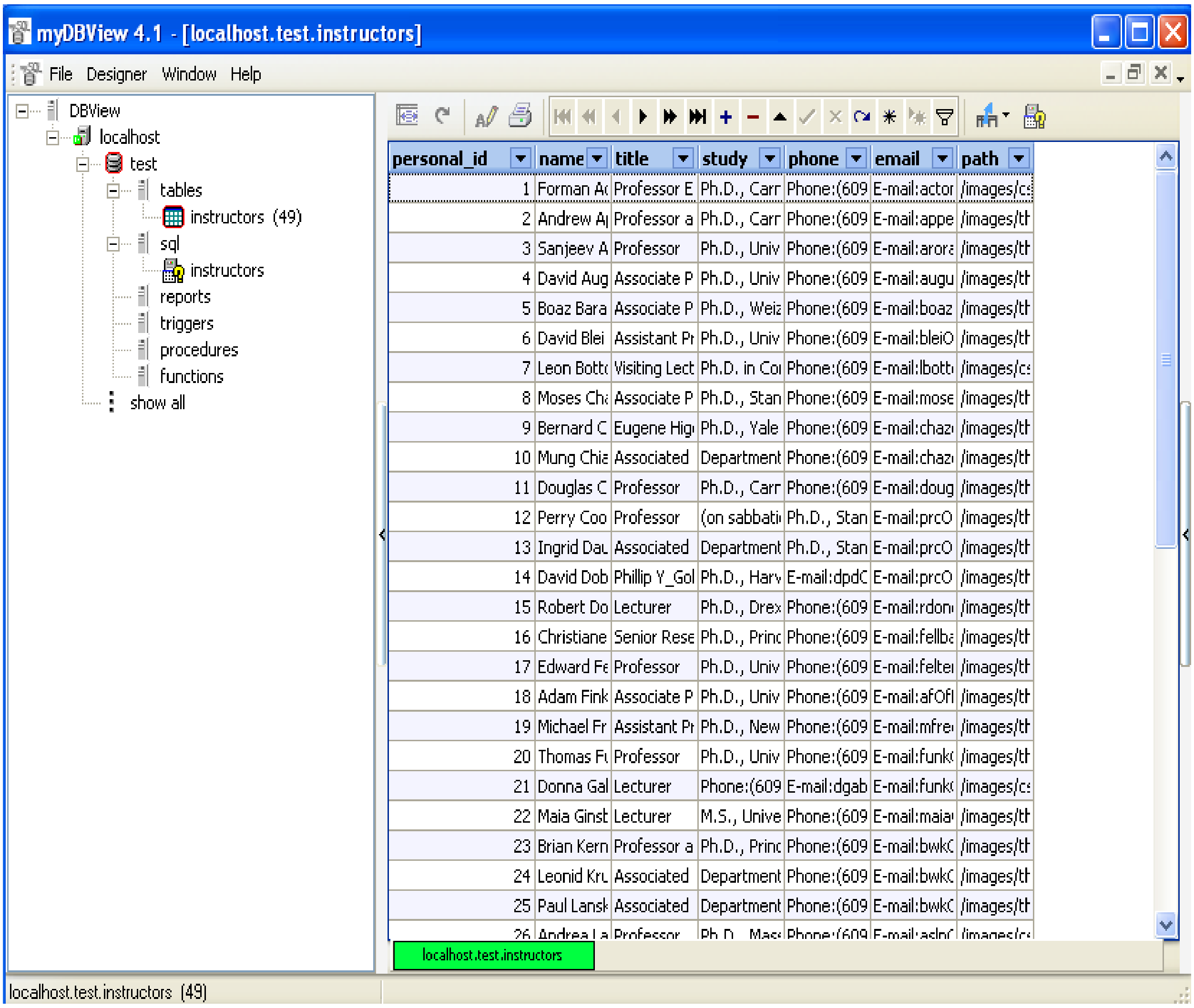


Fig. 3 ARIEW System Architecture

The module of Concept Crawler - Our system highly depends on the topic related data using Concept hierarchy, hence we use topic focused crawler to collect data. This structure is presented to evaluate WEB Pages about the specific topic. Only the WEB Pages whose score is greater than a given threshold is fetched. Two factors contribute to the score. The first one is the content of given web pages, including title, keyword, text and description. The second one is the predefined patterns in the web page. For those satisfied Web pages, we access them, analyze the content inside, organize them with XML format, and store them into data storage. Figure 2 shows the hierarchical structure of the data collecting module and the procedure for data parsing.

Database Module - This module store the result of web pages collected from the Concept Crawler and the researcher information extracted from collected web pages using the Agent to index and queries the databases. The database created by the agent was implemented using MySQL. The key contact information in the database consists of researcher name (Figure 4). The database also stores general information about the researchers includes title, photos, address, phone, email, interests, information about activities and publications.


myDBView 4.1 - [localhost.test.instructors]
File Designer Window Help
DBView
localhost
test
tables
instructors (49)
sql
instructors
reports
triggers
procedures
functions
show all
localhost.test.instructors
localhost.test.instructors (49)

personal_id	name	title	study	phone	email	path
1	Forman Ac	Professor E	Ph.D., Carr	Phone:(609	E-mail:actor	/images/cs
2	Andrew A	Professor a	Ph.D., Carr	Phone:(609	E-mail:appe	/images/th
3	Sanjeev A	Professor	Ph.D., Univ	Phone:(609	E-mail:arora	/images/th
4	David Aug	Associate P	Ph.D., Univ	Phone:(609	E-mail:augu	/images/th
5	Boaz Bara	Associate P	Ph.D., Weiz	Phone:(609	E-mail:boaz	/images/th
6	David Blei	Assistant Pr	Ph.D., Univ	Phone:(609	E-mail:bleiO	/images/th
7	Leon Bottc	Visiting Lect	Ph.D. in Co	Phone:(609	E-mail:lbott	/images/cs
8	Moses Cha	Associate P	Ph.D., Stan	Phone:(609	E-mail:mose	/images/th
9	Bernard C	Eugene Higi	Ph.D., Yale	Phone:(609	E-mail:chaz	/images/th
10	Mung Chia	Associated	Department	Phone:(609	E-mail:chaz	/images/th
11	Douglas C	Professor	Ph.D., Carr	Phone:(609	E-mail:doug	/images/th
12	Perry Coo	Professor	(on sabbati	Ph.D., Stan	E-mail:prcO	/images/th
13	Ingrid Dau	Associated	Department	Ph.D., Stan	E-mail:prcO	/images/th
14	David Dob	Phillip Y_Gol	Ph.D., Harv	E-mail:dpdC	E-mail:prcO	/images/th
15	Robert Do	Lecturer	Ph.D., Drex	Phone:(609	E-mail:rdon	/images/th
16	Christiane	Senior Rese	Ph.D., Princ	Phone:(609	E-mail:fellba	/images/th
17	Edward Fe	Professor	Ph.D., Univ	Phone:(609	E-mail:felter	/images/th
18	Adam Fink	Associate P	Ph.D., Univ	Phone:(609	E-mail:afOfl	/images/th
19	Michael Fr	Assistant Pr	Ph.D., New	Phone:(609	E-mail:mfre	/images/th
20	Thomas Fu	Professor	Ph.D., Univ	Phone:(609	E-mail:funk	/images/th
21	Donna Gal	Lecturer	Phone:(609	E-mail:dgab	E-mail:funk	/images/cs
22	Maia Ginst	Lecturer	M.S., Unive	Phone:(609	E-mail:maia	/images/th
23	Brian Kern	Professor a	Ph.D., Princ	Phone:(609	E-mail:bwkC	/images/th
24	Leonid Kru	Associated	Department	Phone:(609	E-mail:bwkC	/images/th
25	Paul Lansk	Associated	Department	Phone:(609	E-mail:bwkC	/images/th
26	Andrea La	Professor	Ph.D., Mass	Phone:(609	E-mail:aslnC	/images/cs

Figure 4 MySQL Database shows a sample of the extract information from select WEB Page

The Module of the Agent parses the web pages in the database to get related information from the storage component Figure 4. We use predefined patterns to find the related information about researchers. Each keyword (except for the stop words) extracted by the module using these patterns will be an attribute of the researcher, and stored in the database as a result of indexing. After analyzing all the related web pages, agent module returns the required information about the academic researchers. Our agent module use pattern based extraction mechanisms to extract information on researcher contacts.

Based on the home pages for a researcher in the database, the module of the agent starts to extract information from these web pages and referenced pages or URL Links. Searching WEB pages is done in two different approaches. The first approach based on Keyword and called keyword-based and the second one is called pattern-based search (Figure 5). In the first approach, keyword-based search, the agent searches for keywords as specified in the extraction profile. For each keyword, a set of options is specified which tells the agent what information may be found in proximity to the keyword. Although such keyword searching is relatively simple, it has proved effective and is used in our system to find general information about the researchers and publication lists or project descriptions.

Our agent model use pattern based extraction mechanisms to extract information on researcher contacts. However, the agent itself generates these patterns based on the structure of individual items found in repeating items such as HTML lists and tables.

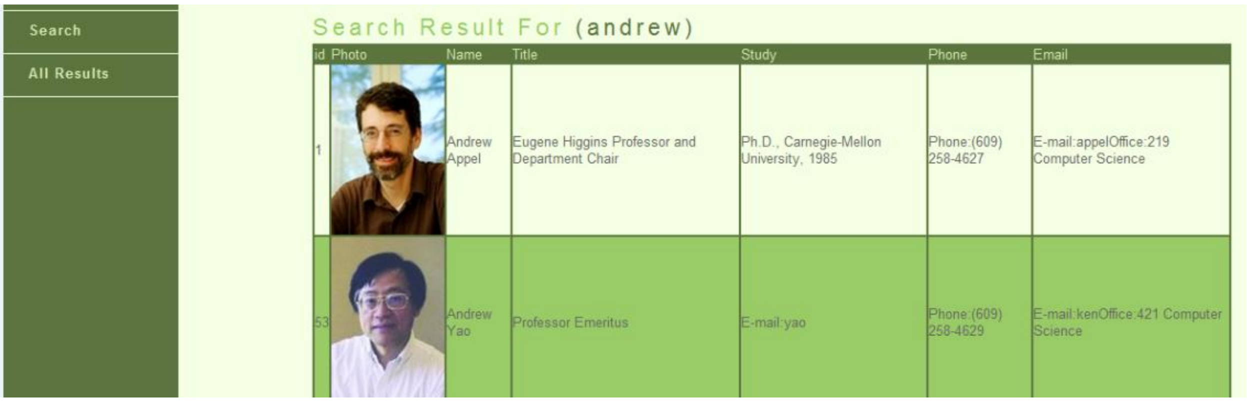
```
<--parse the Researcher objects--!>


<"var-def name="instruct_objects>
xpath expression="data(//td[@class='people >
<"(['people_center
<html-to-xml>
http >
</"url="http://www.cs.princeton.edu/people/faculty
<html-to-xml/>
<xpath/>
<var-def/>
```

Figure 5 A sample XPath code used by the crawl to select WEB Page

```
<-- parse the Researcher photo--!>
<"var-def name="photourl_objects>
<"xpath expression="//td[@class='people']//img[1]/@src>
<html-to-xml>
</"http url="http://www.cs.princeton.edu/people/faculty>
<html-to-xml/>
<xpath/>
<var-def/>
```

Figure 6 A sample XPath code used by the crawl to extract Photos from the WEB Page

Query Processor Module - The main task of query processor is to execute query and provide the information related to the query as a results to the user.



Search Result For (andrew)

id	Photo	Name	Title	Study	Phone	Email
1		Andrew Appel	Eugene Higgins Professor and Department Chair	Ph.D., Carnegie-Mellon University, 1985	Phone:(609) 258-4627	E-mail:appelOffice:219 Computer Science
53		Andrew Yao	Professor Emeritus	E-mail:yao	Phone:(609) 258-4629	E-mail:kenOffice:421 Computer Science

Figure 7 Results show information about the researchers in query

User Interface - A friendly browser-based user interface is presented to the end users. After submitting query keywords, the user will get a comprehensive result shown in Figure 7 which is composed of the information about the researchers in query. And by clicking each of the labels of clustering result, users can get some analysis for each sub-topic respectively, the topics are clustered hierarchically.

In addition, if a user is interested in a particular author, the system provides different information related to the author, likes name, address, email, phone, scientific interests, etc. And the user can also get the answers for the most related information too. The user accesses the database directly or retrieves and process information on researcher contact.

4. Summary

Intelligent Agent aims to facilitate the construction of researchers' profiles by decreasing the amount of effort required to construct researcher databases in special domain. However, there are few studies that attempt to automate the entire construction process from the collection of domain-specific literature, to text mining to build new database or enrich existing ones. In this paper, we present a complete framework for an intelligent Agent that enables us to retrieve documents from the Web using Concept Hierarchy crawling that identify domain-specific documents, and then perform text mining in order to extract useful information form university web pages. We have carried out several experiments on components of this framework in a computer science domain. Other domain can be easily used by adding the concept hierarchy for that domain. This paper reports on the overall system architecture and our initial experiments on information extraction using text mining techniques to enrich the domain researcher database.

This paper presents an academic search engine, which is developed as an efficient tool to construct researcher's profile automatically. Moreover, some searching and indexing methods for underlying XML data are exploited. The paper describes the architecture and main features of the system. It also briefly presents the experimental results of the proposed methods.

Table 1 shows result of performance of the system for the stage of information extraction from different web pages. It is clear that the overall result of the pattern approach is more accurate than the keyword approach. The overall precision of the system was 84.90, which is a good indication about the performance of the system. The result differences refer to the web page and the structure of the web page.

TABLE 1

PERFORMANCE OF THE SYSTEM BASED ON PATTERN AND KEY WORD APPROACH.

Researcher Profile Information	**Patterns Approach**		**Keyword Approach**	
	Precision	**Recall**	**Precision**	**Recall**
Name	91.09	89.32	87.98	89.99
Photo	90.32	88.41	73.12	58.87
Phone	89.75	91.89	76.95	83.25
email	83.21	84.28	81.77	78.32
fax	92.54	89.78	73.12	75.45
address	87.90	84.89	77.98	80.24
Interesting topics	66.78	64.45	59.23	63.47
Position	77.57	65.01	73.99	57.67
Result	**84.90**	**82.26**	**75.52**	**73.40**

5. Conclusion and Future Work

This paper focuses on approaches to extract valuable information from large quantities of unstructured textual information, combining methods from several research areas, including information retrieval, text mining, computational linguistics, and machine learning.

This agent is tested with different experiments. These experiments aim at examining the effect of User Profile and Concept Hierarchy used by the module of Concept Crawler, The overall precision was 91.23 for selecting a related web pages and finally, Precision of the retrieval performance (table 1).

The result shows that agent collects general interests of users when extracting the user profile of the academic Researcher's profile from the WEB. This research indicates agent with using both concept hierarchy and user profile approaches can achieve good result in information retrieval performance.

As future work, such an agent can provide a value added service by using information extracted from Web documents to maintain the database and ensure its currency. The agent may either update the database directly or consult with the user as to whether or not it should perform the updates.

References

[1] Harzing, A. and R. van der Wal. (2008). "Google Scholar as a new source for citation analysis." Ethics in Science and Environmental Politics (ESEP) 8(1):61–73. doi:10.3354/esep00076

[2] Kloda, L. (2007). "Use Google Scholar, Scopus and Web of Science for comprehensive citation tracking." Evidence Based Library and Information Practice2(3):87.

[3] Chang, C.-H.; Kayed, M.; Girgis, R.; Shaalan, K.F.; (2006), IEEE Transactions on Knowledge and Data Engineering, (2006), 18(10), pp 1411 – 1428, DOI: 10.1109/TKDE.2006.152

[4] Dey, L. and Haque, Sk. M., (2009), Opinion mining from noisy text data, International Journal on Document Analysis and Recognition, Volume 12, Number 3, pp 205-226. Doi: 10.1007/s10032-009-0090-z.

[5] Tang J., Zhang D, and Yao L.,(2007a) Social network extraction of academic researchers. In ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp 292–301,

[6] Tang J., Zhang J., Yao L, and Li J., (2008). Extraction and mining of an academic social network. In WWW '08: Proceeding of the 17th

international conference on World Wide Web, pp 1193-1194, New York, NY, USA, 2008. ACM.

[7] Maiorano S., (2006), Question answering: Technology for intelligence analysis. In Tomek Strzalkowski and Sanda Harabagiu, editors, Advances in Open Domain Question Answering, volume 32 of Text, Speech and Language Technology, chapter 16, pages 477–504. Springer Netherlands, Dordrecht, 2006.

[8] Li L., Liu Y., Obregon A., Weatherston M., (2007), Visual Segmentation-Based Data Record Extraction from Web Documents, Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on In Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on (2007), pp. 502-507. doi:10.1109/IRI.2007. doi:10.1016/j.jbi.2009.04.002

[9] Yang Z., Li L., Wang B.and Kitsuregawa M., (2007) Towards Efficient Dominant Relationship Exploration of the Product Items on the Web. AAAI 2007, pp1483-1488

[10] Yang Z., Li L., Wang B.and Kitsuregawa M., (2010) Efficient Analyzing General Dominant Relationship Based on Partial Order Models. IEICE Transactions 93-D(6): 1394-1402 (2010)

[11] Lourenço A., Carreira R., Carneiro S., Maia P, Glez-Peña D., Fdez-Riverola F., Ferreira E. C., Rocha I., and Rocha M.(2009), @Note: a workbench for biomedical text mining. Journal of biomedical informatics, 42(4):710–720, August 2009.

[12] Kheau, C. S., Alfred, R. and Obit, J. H., (2011), BioDARA: Data Summarization Approach to Extracting Bio-Medical Structuring Information, Journal of Computer Science 7(12), pp1914-1920, doi: 10.3844/jcssp.2011.1914.1920

[13] Tang J., Hong M., Zhang D., Liang B., and Li J.,(2007b). Information Extraction: Methodologies and Applications. In the book of Emerging Technologies of Text Mining: Techniques and Applications, Hercules A. Prado and Edilson Ferneda (Ed.), Idea Group Inc., Hershey, USA, 2007. pp. 1-33

[14] Downey D., Etzioni O., Weld D. S., and Soderland S. (2004),. Learning Text Patterns for Web Information Extraction and Assessment. Proceedings of the AAAI-04 Workshop on Adaptive Text Extraction and Mining, 2004.

[15] Liu W. and Zeng J.,(2011) Automatically Extracting Academic Papers from Web Pages Using Conditional Random Fields Model, JOURNAL OF SOFTWARE, VOL. 6, NO. 8, AUGUST 2011, pp1409-1416

[16] Zarandi M. F., Devlin H. J., Huang Y., and Contractor N.,(2011),. Expert recommendation based on social drivers, social network analysis, and semantic data representation. In Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '11). ACM, New York, NY, USA, 41-48. DOI=10.1145/2039320.2039326.

[17] Schirru R., Baumann S, Memmel M and Dengel A (2010), Extraction of Contextualized User Interest Profiles in Social Sharing Platforms, Journal of Universal Computer Science, vol. 16, no. 16 (2010), pp 2196-2213.

[18] Huang J. and Yu C., (2010). Prioritization of Domain-Specific Web Information Extraction. In Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010), AI & the Web Special Track. Atlanta, GA, USA. July, 2010.

[19] Doring, N. (2002). Personal Home Pages on the Web: A Review of Research, Journal of Computer Mediated Communication, 7(3), pp 1-28.

[20] Sabbah T. S., Jayousi R. and Abuzir Y. (2009),"Schema Matching Using Thesaurus”, The 3rd Int. Conference on Software, Knowledge Management and Applications, SKIMA 2009, Fez, Moroco,2009.