Intelligent Information Mining from Veterinary Clinical Records and Open Source Repository

Ploy Tangtulyangkul, Timothy S Hocking, and Chun Che Fung
School of Information Technology
Murdoch University
Perth, Australia
ploy@murdoch.edu.au | t.hocking@murdoch.edu.au | l.fung@murdoch.edu.au

***Abstract*—This paper reports an implementation of an intelligent mining approach from veterinary clinical records and an external source of information. The system retrieves information from a local veterinary clinical database and then complements this information with related records from an external source, OAIster. It utilizes text-mining, web service technologies and domain knowledge, in order to extract keywords, to retrieve related records from an external source, and to filter the extracted keywords list. This study meets a practical challenge encountered at the School of Veterinary and Biomedical Sciences at Murdoch University. The results indicate that the system can be used to increase the limited knowledge within a local source by complementing it with related records from an external source. Moreover, the system also reduces information overload by only retrieving a set of related information from an external source. Finally, domain knowledge can be used to filter the extracted keywords, in this case, selected medical keywords from the extracted keyword list.**

Keywords- ***clinical database; keyword extraction; query; text-mining; veterinary records; web services; information filtering; query***

I. INTRODUCTION

The School of Veterinary and Biomedical Sciences at Murdoch University has a teaching veterinary hospital, the Murdoch University Veterinary Hospital (MUVH). MUVH has operated for many years, with a large number of cases treated each year [1]. Subjects in the records range from small to large animals, and are dominantly domestic or farm animals. Most of the treated animals are dogs, cats, sheep, cattle, horses, and many other species. Some of these cases have also been operated at MUVH. The information has been recorded in a dedicated database system and it currently holds a large number of records, occupying several gigabytes of storage, and the system is growing continuously.

MUVH currently uses a Veterinary Practice Management Software package called RxWorks to store the clinical records of treatment of animals. This also provides management functionalities for appointments, work scheduling, accounting, and it also supports many types of data including treatments, diagnosis and patient history, as well as providing reporting and data query tools [2].

Researchers at the School of Veterinary and Biomedical Sciences regularly use this data for research and analysis. At present, when researchers wish to acquire information from the database, they need to acquire assistance from the IT Support to look for data or records based on particular criteria or keywords. Researchers will then be given the results according to their requested query, usually in a spreadsheet format.

In order to improve the process, the clinical records have been exported into a local database from which the data can then be retrieved via query and search systems. However, there might be instances where researchers may need to look for additional information based on other treatments, or descriptions of previous cases that relate to the present clinical records. Performing queries or searching directly from an external database may result in a large number of records. This is a shortcoming of the current system as it is not linked to any external sources.

This paper presents an intelligent information retrieval system based on a list of extracted keywords from the local veterinary clinical records. This keyword list can be used to enable the system to retrieve a comprehensive set of information related to the local clinical records. This paper is organized in the following manner. Section I outlines the problem and possible solution, Section II details the background of this work, Section III presents the proposed intelligent integrated query system, while Section IV concentrates on improving keyword extraction with medical keywords. Section V presents the results and discussion, and conclusions are given in Section VI.

II. BACKGROUND

Keywords are words or phrases that can assist querying and retrieving information [3], [4]. However, it is recognized that sometimes query systems can result in *information overload* [5], [6]. According to previous studies, there are many applications and systems that aim to assist users by reducing the amount of information presented to them, as well as increasing the relevancy of the information. Examples are search engines, personalization systems and content-filtering systems [5], [6], [7], [8]. However, it appears that previous systems and applications have focused on techniques to reduce the amount of retrieved information within the system. This proposed system aims to reduce information overload from *external* sources by using the existing information, and supplementing the local records with related information from

978-1-4244-4547-9/09/$26.00 ©2009 IEEE

external databases. This is achieved by using text-mining, keyword extraction and filtering based on specific domain knowledge.

A. Text-Mining and Keyword Extraction

Text-mining is a technique used to discover potentially unknown information from a passage of text. For example, a passage may be obtained from a website of biomedical domain or paragraphs of text [9], [10]. Text-mining has been used extensively in the process of acquiring knowledge from websites in both commercial and non-commercial applications in a variety of domains. This is particular useful as most websites contain a fair amount of text [10], [11]. Text-mining techniques can be used in conjunction with Natural Language Processing (NLP) to increase the understanding of the given information; for example, handling a single word that may have a different meaning in different contexts [9], [11], [12].

Previous studies have found that there are many applications that use text-mining techniques to extract keywords from paragraphs of text. Example applications include the discovery of suitable keywords for search engine optimisation of web content for major search engines, automatic keyword and article linking in Wikipedia pages, and automated content indexing for books [11], [13], [14].

The system presented in this paper utilizes text-mining and keyword extraction to discover keywords from existing local clinical records. In addition, keyword extraction technology for the local retrieved information can be done either locally or through services provided via web service technology.

B. Web service Technology

World Wide Web Consortium (W3C) defines web service as "...a software system designed to support interoperable machine-to-machine interaction over a network" [15]. Web service technology relies heavily on an eXtensible Markup Language (XML) [16]. It allows machine-to-machine interaction through an interface with a set of machine description format called Web Services Description Language (WSDL) [15], [16].

The system presented in this paper applies web service technology to create interaction between local system, external source and service, in order to allow information retrieval and knowledge acquisition via machine-to-machine interaction.

III. SYSTEM ARCHITECTURE

A. Overview

An intelligent information retrieval architecture is presented in this paper. The purpose is to allow users to retrieve information from local clinical records, and use information discovered from those records to remotely retrieve additional related information from external sources. This enables users to view local records with complementing records from external sources.

The user first enters a keyword or term to query the local database, for example, "cancer". This term of interest is called the "*root-word*". The architecture extracts all records related to the root-word from the database. All extracted records are submitted to a keyword extraction system for "keyword" extraction by utilizing text-mining and keyword extraction technologies. This results in a list of common keywords in the records. The user can further combine any keywords with the root-word. Such combinations can be used to further filter and extract relevant local records.

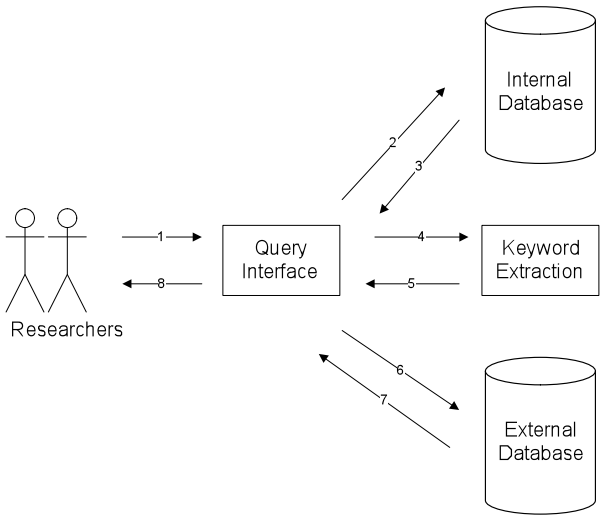

Figure 1. Processes of Intelligent Information Retrieval

In addition, users can also use the combination of the extracted keywords and the root-word to query external databases for further information or documents such as research papers, reports and/or other clinical records. This can be handled by combining keyword extraction processes and web service technologies to remotely retrieve external records. The purpose of this is to reduce information overload, to improve the relevancy of the retrieved information and to enhance the limited knowledge from local sources by supplementing it with information from external sources.

Fig. 1 illustrates and summarizes the process to retrieve the final results.

1) Users enter a root-word to the query interface.
2) Query interface queries records from internal database based on the given root-word.
3) Query interface receives the records from internal database query results.
4) The retrieved records will be concatenated and sent for the keyword extraction process.
5) The process then returns a list of keywords extracted from the local records.
6) A combination of root-word and keywords from the list will be used to query external databases.
7) Results return from external databases.
8) Users are presented with all the information.

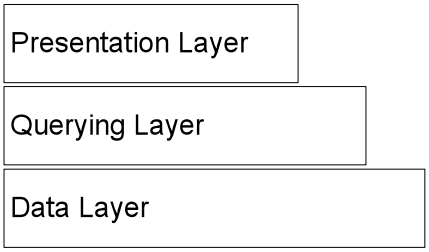

Figure 2. The three layer System Architecture

Researchers can also use the combination of extracted keywords and the root-word to further filter the clinical records (2) and repeat the rest of the processes.

B. Three Layer Architecture

The system architecture has been divided into three layers: data layer, querying layer and presentation layer as shown in Fig. 2.

The data layer handles all data entered into the system. Data for internal database querying is imported into the database of the system. In addition, suitable database indexing can be applied to improve and optimize the query performance.

The querying layer is a major part of the system. It handles three different types of querying such as clinical records, keywords and additional information from external sources, as illustrated in Fig. 3. When a user enters a keyword (or, root-word) into the system, the clinical records in a local database will be queried, and relevant records are retrieved through SQL. All the retrieved clinical records are then concatenated (joined) into one piece of free-form text. The text-mining process will be performed upon the concatenated passage and keywords will be discovered from the clinical records as shown in Fig. 4.

For example, when a researcher enters the keyword "cancer", 100 clinical records may be retrieved. The system then performs keyword extraction and returns five keywords to the user, based on the above 100 clinical records.

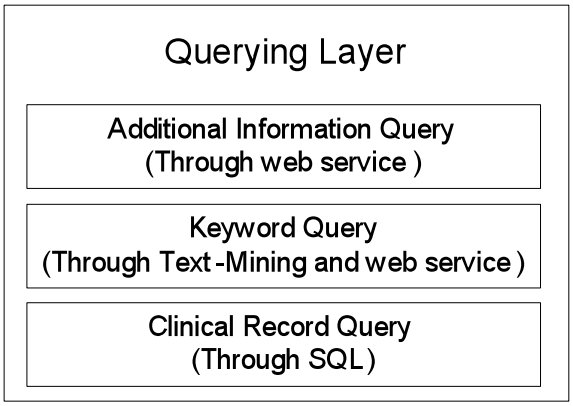

Figure 3. System's Querying Layer

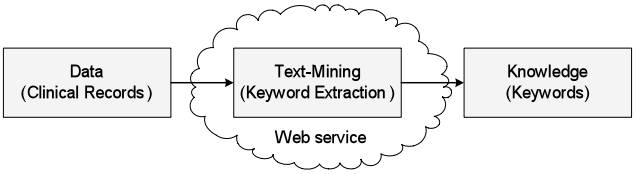

Figure 4. Keyword Extraction Process

A user has the choice to filter the existing result list. If the user decides to filter clinical records, the system will utilize SQL to query the local database based on the combination of keywords and the root-word.

In order to complement the existing information from a local database, another level of querying can be executed on an external database using a combination of the initial root-word and extracted keywords. For example, the user may have one root-word together with an additional five keywords suggested from the text-mining and keyword extraction processes. The user may then use the newly formed combination to retrieve information from external resources of predefined databases that are related to the keyword combination. Consider a case where the researcher searches with the root-word: the user may end up with too many results causing information overload. There will be too much external information on the same topic but they may be completely unrelated to any clinical records that user has.

By combining keywords and the root-word, a user may find a reduced number of links or articles related to the local clinical records, as a query result. The concept can be illustrated in the Venn diagram in Fig. 5.

The result given by this system is a subset of the results from the root-word and the extracted keywords from the given clinical records. This approach should reduce information overload by using information from the external sources to complement the existing information from the local database system.

Lastly, the presentation layer deals with the presentation, layout and format of the output. This presentation layer is important for users as it defines the look and feel of the system,

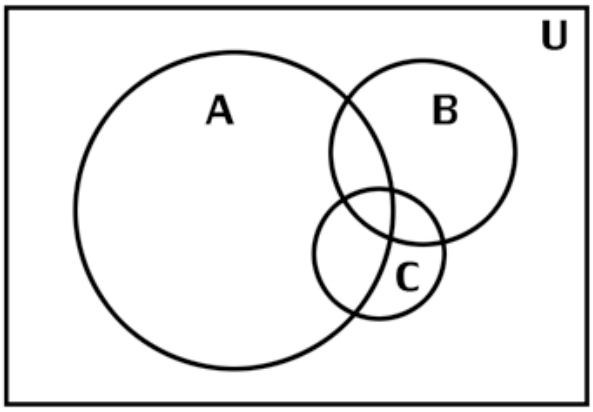

Figure 5. A Venn diagram representing the result set. U represents all possible results. A represents all possible results from root-word and B and C represent all possible results from each keyword, respectively.

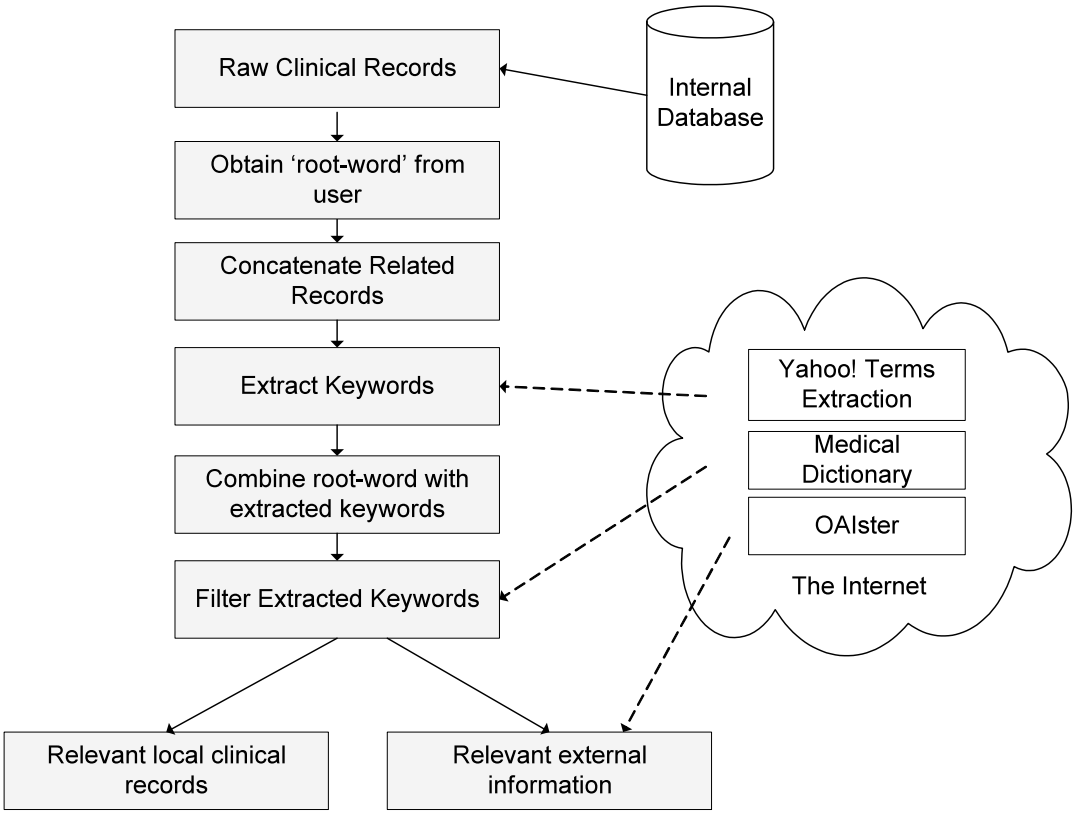

Figure 6. Overall system process

as well as the visualization of data for analysis. The proposed basic presentation layer involves clustering the results based on keywords. This permits researchers to visualize the given information as a group of results. For example, assuming that "cow" is one of the groups, users can view the subset of results based on "cancer" and "cow", rather than a large amount of results based on "cancer".

C. *Proof-of-concept Prototype*

A prototype of this proposal has been designed and developed, for the investigation and evaluating the functions of the purposed intelligent integrated query system. The implementation is based on a practical case study at the School of Veterinary and Biomedical Sciences at Murdoch University. The input data is sourced from Murdoch University Veterinary Hospital and has been extracted and de-identified. This proof-of-concept-prototype (Fig. 7) is implemented as a web-based portal, using XHTML, CSS, JavaScript/AJAX, PHP and MySQL on a Linux based server. The prototype utilizes text-mining (keyword extraction) and web service technologies for extracting keywords and retrieving records from external databases.

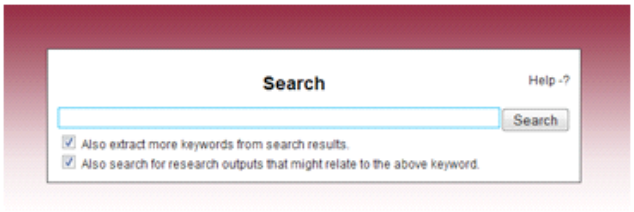

Figure 7. The prototype's query interface

In order to demonstrate this prototype, *Yahoo! Term Extraction* [18] and OAIster have been selected as keyword extraction tools and the external database, respectively. Such tool or resource could be changed to other alternative choices. Yahoo! Term Extraction is a part of Yahoo! Content Analysis Web Services [18]. It provides a web service interface for extracting keywords from free-form or large texts. Data retrieved from Yahoo! Term Extraction can be exported in various formats, which are: XML, JSON and Serialized PHP format [18].

OAIster has been selected to represent external resources in this prototype. OAIster is "*a union catalog of digital resources*" [19]. It allows users to access digital formats of research outputs from various research repositories. OAI-PHM refers to the Open Archives Initiative Protocol for Metadata Harvesting [17], [19]. OAIster uses OAI-PMH protocol to harvest those records [19].

OAIster also allows its data to be used outside the web interface [20]. OAIster implements SRU (Search/Retrieval via URL) protocol, in order to allow users to query its data via web service. Data retrieved from OAIster SRU web service is in DLXS BibClass metadata format [21].

TABLE I. NUMBER OF RELATED RESULTS

Keywords	Number of results
cancer	299,903
cancer, inj	33
cancer, cartilage	63
cancer, scab	3
cancer, bone cancer	85
cancer, inj, scab	36

IV. IMPROVING KEYWORD EXTRACTION WITH DOMAIN KNOWLEDGE

At present, the keyword extraction process returns a number of keywords extracted from the retrieved clinical records. However, not all keywords may be relevant to the medical or clinical domain. Therefore, this points to the need to implement a filtering system by incorporating *domain knowledge*, in order to improve the relevancy of the extracted keywords.

Accordingly, a medical dictionary has been integrated into the prototype to improve the keyword filtering process as indicated in Fig. 6. This filtering process is performed after the keyword extraction process. It extends the initial keyword extraction by adding a filtering system to further improve the relevancy of the extracted keywords.

V. RESULTS AND DISCUSSION

A. Prototype Results

A field test was conducted on the local *Murdoch University Veterinary School* data using a root-word of "cancer". The first queried result of the local database returned 231 clinical records based on the keyword *cancer*. Twenty keywords were extracted from those clinical records, such as: anti biotic, bone cancer, cartilage, inj (injection), raw wound, scab and serum. Those keywords, in combination with the root-word, were used to retrieve related information; in this case, a list of related research outputs from an external database (OAIster).

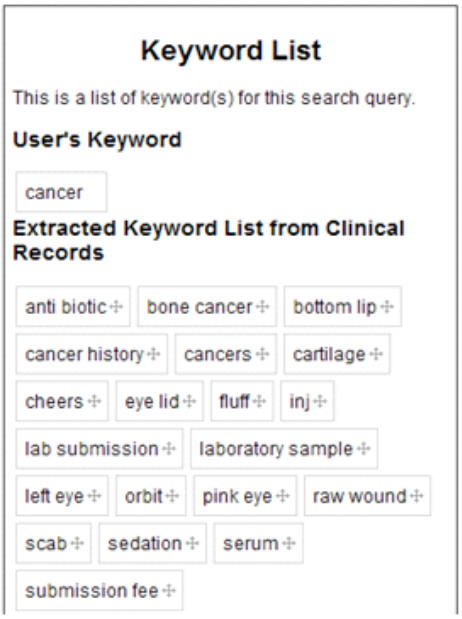

Figure 8. Keyword list based on cancer

The result showed that by querying for cancer, via the SRU protocol, 299,903 records were returned. In comparison, the combination of keywords from extracted keyword list returns a smaller subset (Table 1).

This clearly illustrates that the amount of related information retrieved from an external resource can be significantly reduced by filtering with extracted keywords to retrieve related records from an external source.

B. Improving Keyword Extraction Result

Initially, there were 20 extracted keywords retrieved from the root-word of *cancer* as shown in Fig. 8. The initial extracted keywords were: bone cancer, bottom lip, cancer history, cancers, cartilage, cheers, eye lid, anti biotic, fluff, inj, lab submission, laboratory sample, left eye, orbit, pink eye, raw wound, scab, sedation, serum and submission fee.

This result shows that some of these keywords are not directly relevant, as they are not medical or clinical terms.

For demonstration purposes, an online medical dictionary was applied as an additional filter for the extracted keywords. This reduced the keyword list to just two results, and both keywords are directly relevant to medical or clinical domain (Fig. 9).

This proof-of-concept prototype demonstrates that it is possible to use keyword extraction in order to aid research in the medical domain, by both attempting to reduce information overload, and also to provide links to relevant external research articles. In addition, the prototype has shown that a keyword filtering system can be used to remove keywords that may not be directly relevant; in this case, non-medical keywords. The relevance of the extracted keyword list is thus increased.

This demonstrates that the implementation of *domain knowledge,* for keyword filtering, can be used to improve relevancy of the results presented to the user by the system.

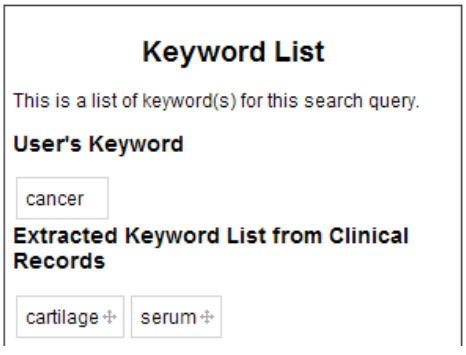

Figure 9. Filtered keyword list

VI. Conclusion

This paper investigated and reported a prototype intelligent keyword mining system, demonstrated by filtering a list of extracted keywords from veterinary clinical records. The overall system can be used to retrieve related information from both internal and external sources by using a combination of the search terms, obtained through a keyword extraction process performed on local clinical records. However, not all extracted keywords are directly relevant or related to a medical domain. Therefore, domain knowledge is incorporated into the system to improve the relevancy of results. The results demonstrate that the system can be used to increase the knowledge within a local source by complementing it with related records from an external source. The system also reduces information overload by retrieving a set of related information from external sources. The applicability of this system is not restricted to the veterinary discipline. It is planned to investigate the potential of the system for other domains such as energy and sustainable environment etc.

Acknowledgment

The authors would like to acknowledge Professor John Edwards and Associate Professor Ian Robertson of the School of Veterinary and Biomedical Sciences, Murdoch University for their kind support given to this project.

References

[1] "Murdoch University Veterinary Hospital," Mar 15, 2005. [Online]. Available: http://www.vetbiomed.murdoch.edu.au/hospital. [Accessed: Mar 18, 2009].

[2] RxWorks Inc., "Practice Management Software for Veterinary Clinics," 2008. [Online]. Available: http://www.rxworks.com. [Accessed: Mar 18, 2009].

[3] Montgomery College, "Library Vocabulary," 2009. [Online]. Available: http://www.montgomerycollege.edu/library/libtp/instructions/daudu/articlevoc.htm. [Accessed: Mar 18, 2009].

[4] Pay equity Commission, "Pay Equity – Glossary of Internet Terms," Dec 4, 2007. [Online]. Available: http://www.payequity.gov.on.ca/peo/english/pubs/glossarynet.html. [Accessed: Mar 18, 2009].

[5] Y. D. Wang and G. Forgionne, "Testing a decision-theoretic approach to the evaluation of information retrieval systems," *Journal of Information Science*, vol. 34, no. 6, p. 861, Dec 2008.

[6] N. Hochstotter and M. Koch, "Standard parameters for searching behaviour in search engines and their empirical evaluation," *Journal of Information Science*, vol. 35, no. 1, p. 45, Feb 2009.

[7] H. Levkowitz, "Personalized information retrieval and access: concepts, methods, and practices," *Choice*, vol. 46, no. 5, p. 942-943, Jan 2009.

[8] "Machine Learning; Scientists at Shanghai Jiao Tong University report research in machine learning," *Journal of Technology & Science*, p. 1494, Sep 2008.

[9] Y. Mivao, K. Sagae, P. Sætre, T. Matsuzaki and J. Tsujii, "Evaluating contributions of natural language parsers to protein-protein interaction extraction," *Bioinformatics*, vol. 25, no. 3 p. 394-401, Feb 2009.

[10] J. Y. Ahn, S. K. Kim and K. S. Han, "On the design concepts for CRM system," *Industrial Management & Data Systems*, vol. 103, no. 5, p. 324-331, 2003.

[11] TextDigger Inc., "TextDigger Announces $4.3 Million A-1 Funding," *Science Letter*, p. 4168, Feb 2009.

[12] D. Jiao and D. J. Wild, "Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods," *Journal of Chemical Information and Modeling*, vol. 49, no. 2, p. 263, Feb 2009.

[13] A. Csomai and R. Mihalcea, "Linking Document to Encyclopedic Knowledge" *IEEE Intelligent Systems*, vol. 23, no. 5, p. 34, Sep/Oct 2008.

[14] C. Andras, "Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing," Ph.D. thesis, University of North Texas, 2008.

[15] World Wide Web Consortium, "Web services glossary," Feb 11, 2004. [Online]. Available: http://www.w3.org/TR/ws-gloss. [Accessed: Mar 19, 2009].

[16] A. Carlos, "Web services will maintain operability during crunch time," *Business World*, p. 1, Jan 2004.

[17] "CrossRef Extends Web Services with OAI-PMH, Adds Partnerships," *Information Today*, vol. 24, no. 1, p. 40, Jan 2007.

[18] Yahoo! Inc., "Term extraction document for Yahoo! search web services," 2009. [Online]. Available: http://developer.yahoo.com/search/content/V1/termExtraction.html. [Accessed: Mar 20, 2009].

[19] OAIster, "OAIster about," 2009. [Online]. Available: http://oaister.org/about.html. [Accessed: Mar 20, 2009].

[20] OAIster, "OAIster using OAIster data outside this interface," 2009. [Online]. Available: http://www.oaister.org/sru.html. [Accessed: Mar 20, 2009].

"DLXS Bibliographic Class Documentation," 2009. [Online]. Available: http://www.dlxs.org/docs/12a/class/bib/index.html. [Accessed: Mar 20, 2009].