AN INTELLIGENT INTEGRATED QUERYING SYSTEM FOR FREE-FORM INFORMATION EXTRACTION FROM VETERINARY CLINICAL RECORDS

PLOY TANGTULYANGKUL, CHUN CHE FUNG

School of Information Technology, Faculty of Creative Technologies and Media, Murdoch University, Murdoch, Australia
E-MAIL: ploy@murdoch.edu.au, l.fung@murdoch.edu.au

Abstract:

The aim of this paper is to report an intelligent integrated query system that provides and uses information from local veterinary clinical records supplemented with information from external resources. The information from the local records is used to remotely retrieve related information from external sources, in order to supplement the existing records for diagnosis and decision purposes. In Murdoch University, a large number of historical clinical data are stored in a veterinary practice management system for several years. It is necessary to provide an efficient system for query and retrieval of data for research work. The proposed solution in this paper involves text-mining (keyword extraction) and web service technologies for enhancing the existing results and to provide a comprehensive set of information related to local clinical records from the internal database.

Keywords:
Clinical database; Keyword extraction; Query; Text-mining; Veterinary records; Web services

1. Introduction

The School of Veterinary and Biomedical Sciences at Murdoch University has a teaching veterinary hospital, the Murdoch University Veterinary Hospital (MUVH). MUVH has been operated for many years with a large number of cases treated each year [1]. The subjects in the records range from small to large animals and they are mostly domestic or farm animals. Most of the treated animals are dogs, cats, horses and many other species and the cases operated at Murdoch University Veterinary Hospital have been entered into a database. Currently, the database has a large number of records, which occupies several gigabytes of storage and it is growing continuously.

Murdoch University Veterinary Hospital uses a Veterinary Practice Management Software called RxWorks, to store their clinical records of treatment of animals. The system is quite complex and it provides functionalities for appointments, work scheduling, accounting and it also supports many types of data including treatments, diagnosis, patient history as well as provides reporting and data query tool [2].

Researchers at the School of Veterinary and Biomedical Sciences use this valuable data for research and analysis. At present, when researchers would like to obtain the data from the veterinary hospital’s database for their research, they have to go to MUVH IT Support and ask for data or records based on particular criteria or keywords. Researchers will then be given the results according to their requested query, for example, in a spreadsheet format.

Those clinical records have been exported and have been imported into a local database. The data can be retrieved via query and search systems. However, it may be necessary to look for additional information, such as, other treatments, description, previous cases that relates to these clinical records. There is a particular shortcoming performing query or search directly from an external database as it can result in a large number of results.

The aim of this paper is to propose an intelligent integrated query system that can demonstrate and provide information from internal clinical records and complement these records with related information from external sources. This will enable the system to provide a comprehensive set of information related to the internal database and the local clinical records.

This paper is organised in the following manner. Section 1 is the introduction and it is followed by Section 2 on background. Section 3 is the proposed intelligent integrated query system while Section 4 is the results. Finally, the conclusions are given in Section 5.

2. Background

Keywords are words or phrases that can assist querying and retrieving information [3, 4]. However, it is recognized that sometimes query system can result in
According to previous studies, there are many applications and systems that aim to assist users by reducing the number of information and increasing the relevancy of the information; such as search engines, personalized system and filtering system [5, 6, 7, 8]. However, it appears that those systems and applications 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 supplement the local records with related information from external database. The next Section discussed Text-mining and keyword extraction.

2.1 Text-Mining and Keyword Extraction

Text-mining is a technique used to discover information from a passage of text. For example, the passage can be obtained from a website of biomedical domain or paragraphs of text [9, 10]. Text-mining has been used heavily in acquiring knowledge from websites in both commercial and non-commercial applications as information from website is mostly text [10, 11]. Text-mining technique can be used in conjunction with Natural Language Processing (NLP), in order to increase the understanding of the given information, for example, handling a single word that may have a different meaning in different context [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 are discovery of suitable keywords for search engines, automatic keyword and article linking in wiki pages, and automated indexing contents for books [11, 13, 14].

This proposal utilizes text-mining and keyword extraction concept in order 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.

2.2 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].

Mashup is another popular technology which involves the creation of a new application that employs more than one web services [17]. Web service and mashup have been used in many areas, for example, providing location-based services, sharing data, real estate, weather, bioinformatics and research repository [17, 18].

This paper proposes to apply 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.

3 Proposed Intelligent Integrated Query System

An intelligent integrated query system has been proposed. The system architecture has been divided into three layers: data layer, querying layer and presentation layer as shown in Figure 1.

Firstly, data layer handles all the entered data into the system. Querying layer is a major part of the system. It handles three different types of querying, which are: clinical records, keywords and research output as shown in Figure 2. Lastly, the presentation layer deals with the presentation, layout and format of the output.

The proposed querying process is illustrated in Figure
3. When a user enters a keyword (or, base keyword) into the system, the clinical records in a local database will be queried and relevant records are retrieved through SQL. Then all the retrieved clinical records will be concatenated into one big free-form text. Text-mining process will be performed to the concatenated passage and keywords will be discovered from the clinical records as shown in Figure 4.

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

In order to complement the existing information from a local database, another level of querying can be executed on an external database. Querying the external database can be done by using a combination of user-defined keyword (or base keyword) and keywords from the list that has been extracted from the clinical records.

For example, the user may have one base keyword together with extra five keywords being suggested from the text-mining process. The user then may use the combination to retrieve external resources from predefined databases that are related to the keyword combination. Consider the case that the researcher searches with the base keyword, the user may end up with too many results causing information overload. There will be many external information on the same topic but they may not be related to any of the clinical records that user has. By combining keywords and base keyword, user may get a much reduced number of links or articles related to the local clinical records, as a query result.

The result given by this proposed system is in fact a subset of the results from based keyword and extracted keywords from the given clinical records. This approach will reduce information overload by using information from the external source to complement the existing information from the local database system.

A prototype of this proposal has been designed and developed, for the investigation and evaluating the functions of the purposed intelligent integrated query system. In order to demonstrate this prototype, Yahoo! Term Extraction [19] and OAIster have been selected as keyword extraction tools and the external database, respectively.

Yahoo! Term Extraction is a part of Yahoo! Content Analysis Web Services [19]. It provides web service interface for extracting keywords from free-form or a large text. Data retrieved from Yahoo! Term Extraction can be in various formats, which are: XML, JSON and Serialized PHP format [19].

In addition, “OAIster is a union catalog of digital resources” [20]. OAIster allows user to access digital format of research outputs from various research repositories. OAI-PMH refers to the Open Archives Initiative Protocol for Metadata Harvesting [18, 20]. OAIster uses OAI-PMH protocol to harvest those records [20].
OAIster also allows its data to be used outside the web interface [21]. OAIster implements SRU (Search/Retrieval via URL) protocol, in order to allow users to query its data via web service (See Figure 5). Data retrieved from OAIster SRU web service is in DLXS BibClass metadata format (See Figure 6) [22]. Therefore, it has been used as a simulation of external source.

![Figure 7 Venn Diagram Representing Query Data Set](image)

Figure 7 Venn Diagram Representing Query Data Set

The querying logic and data set are represented using Venn Diagrams (Figure 7). Assuming that A is a base keyword result set and B, C, ...n are result sets from each keyword extracted from the local queried clinical record list. A set of results related to clinical records from base keyword can be found by using the following set operations (1):

$$A \cap (B \cup C \cup ...n)$$

(1)

However, OAIster has not supported parentheses in the query phrase. Therefore, the above formula has been expanded to (2):

$$A \cap B \cup A \cap C \cup ...A \cap n$$

(2)

Finally, the purposed basic presentation layer involves clustering the results based on keywords. This allowed researchers to visualise the given information as a group of results. Assuming that cow is one of the main groups, user can view the subset of results based on cow, rather than a large number of results based on cancer.

A proof-of-concept prototype has been designed and developed in order to demonstrate the purposed intelligent integrated query system. A screenshot of the prototype system is shown in Figure 8.

![Figure 8 A screenshot of the proof-of-concept prototype](image)

Figure 8 A screenshot of the proof-of-concept prototype

4 Results

The proof-of-concept prototype was examined by entering cancer as a base keyword. The first queried result of the local database and returned 231 clinical records based on the keyword cancer. 20 keywords were extracted from those clinical records; such as: anti biotic, bone cancer, cartilage, inj (injection), raw wound, scab and serum. Those keywords or a combination of cancer and any combination of keywords was used to retrieve related information, in this case is a list of related research output, from an external database (OAIster).

The result showed that by querying for cancer, via SRU protocol, 299903 records were returned. In addition, the combination of keywords from extracted keyword list returns a smaller subset (See Figure 9).

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Number of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>299903</td>
</tr>
<tr>
<td>cancer, inj</td>
<td>33</td>
</tr>
<tr>
<td>cancer, cartilage</td>
<td>63</td>
</tr>
<tr>
<td>cancer, scab</td>
<td>3</td>
</tr>
<tr>
<td>cancer, bone cancer</td>
<td>85</td>
</tr>
<tr>
<td>cancer, inj, scab</td>
<td>36</td>
</tr>
</tbody>
</table>

Figure 9 Number of related results

Figure 9 illustrates that the number of related information retrieved from an external resource can be significantly reduced by using extracted keywords based on the local clinical records. For example, by using a combination of cancer and keywords; such as, cartilage and bone cancer, both of the number of results has been reduced by 99.97% (See Figure 10).

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer, inj</td>
<td>99.99%</td>
</tr>
<tr>
<td>cancer, cartilage</td>
<td>99.97%</td>
</tr>
<tr>
<td>cancer, scab</td>
<td>99.99%</td>
</tr>
<tr>
<td>cancer, bone cancer</td>
<td>99.97%</td>
</tr>
<tr>
<td>cancer, inj, scab</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

Figure 10 Percentage of number of results reduced
The initial results have demonstrated that a local set of information can be used to reduce information overload, when getting a set of related information from an external source. The result also showed that this proposed solution can be used to significantly reduce the number of results. However, this paper only reported the initial phase of the proposed system. Ongoing work is carried out to examine the usefulness of the system by inviting experts to assess the quality of the returned records from the local and external resources. Other work required are optimising and improving the relevancy and accuracy of the results. Other areas that need to be investigated in are the text-mining, keyword extraction and querying algorithms and determination of the optimal number of returned keywords and the optimal number of records from external sources.

5 Conclusions

The aim of this paper is to report the development of a proposed intelligent integrated query system that can be used to provide information from internal clinical records and complement them by related information from external sources. This will help researchers for their research and analysis on any particular keyword. The system can provide a comprehensive related set of information based on internal database from an external sources that relates to local clinical records.

Results of the initial study have shown that this proposed system is able to provide a comprehensive set of related information from external source based on a set of local clinical records. It has demonstrated that the number of remotely retrieved information and information overload have been reduced significantly.

However, this paper only provided an overview of the system architecture and further research will be continued on the examination and improvement of individual technology in order to optimise and improve the relevancy of the results.

Acknowledgements

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

References


[16] A. Carlos, “Web services will maintain operability

Authorized licensed use limited to: Murdoch University. Downloaded on October 13, 2009 at 04:15 from IEEE Xplore. Restrictions apply.


