
IBIMA
Publishing
mobile

Journal of Software and Systems Development

Vol. 2011(2011), Article ID 611355, 220 minipages.
DOI:10.5171/2011.611355
www.ibimapublishing.com

Copyright © 2011 Norasykin Mohd Zaid and Sim Kim Lau. This is an open access article distributed under the Creative Commons Attribution License unported 3.0, which permits unrestricted use, distribution, and reproduction in any medium, provided that original work is properly cited.

Development of Ontology Information Retrieval System for Novice Researchers in Malaysia

Authors

Norasykin Mohd Zaid and Sim Kim Lau

University of Wollongong, Wollongong, New South Wales, Australia

Abstract

This research describes the development of an online thesis database system to assist novice researchers in

a Malaysian university to identify research topics in local language context. Two major problems have been identified in the current system of keyword search.

Firstly, student's lack of experience in querying often results in irrelevant search outcomes. Secondly, language barrier limits students' abilities to

conduct keyword search in bilingual language context. A semantic search approach that applies ontology-based search is proposed. This paper

presents system framework design, its ontological development and sample queries are also presented to demonstrate how the system works

Keywords: Semantic Search, Ontology, Information Retrieval.

Introduction

Increasingly, information and knowledge are becoming more shareable and searchable resources in

the current digitized world. Since 1996, the World Wide Web (WWW) has become a primary source for information offering online resources that are available

24/7. Traditionally, the library has become an important place for information especially academic resources for researchers. Library

classification system has migrated from Dewey Decimal Classification System (DDC) to a new digitized format such as Online Public Access

Catalog (OPAC) system that can be accessed through the web. Undoubtedly, digital library provides an improved source of information access that

include digital document creation and storage, documents classification and data indexing, documents searching and retrieving, distribution,

administration and access control (Garza-Salazar et al., 2003). However, human interpretation is still required when records matching the search

criteria (such as keywords) are returned to determine its relevance and usefulness. For example, in searching for a programming textbook

which we do not know its exact title, we tend to type the word *programming* in the search box. When search results are returned, we scroll down the list of

titles to look for the one that we search for. This is commonly encountered by students who are inexperienced in literature search.

As there are a lot of information available on the Internet, searching nowadays focus on getting right information. However human interpretation is

still required when records matching the search criteria (such as keywords) are returned to determine its relevance and usefulness as thousands of pages and

web documents are often returned when search is conducted. To date, a lot of research has been conducted and many techniques have been

proposed to find the best method in searching online digital resource. This includes the semantic search technology which provides semantic search

engine using web metadata. It is clear that description of things is very important and access to documents is hugely impacted by lack of information. Metadata are

becoming important in web 3.0 where more precise means of describing web documents are being developed. The World Wide Web Consortium, in its

official website http://www.w3.org/ has recommended that the Resource Description Framework (RDF) is a standard model of web

documents to allow data integration and data interchanging.

The motivation of this paper is to propose the

development of an ontology-based information retrieval system to assist inexperienced research students at a local university in Malaysia to

search for academic resources in the local language context (Bahasa Malaysia). We will discuss information retrieval in which ontology data is used

to improve online search of digital resource. The rest of the paper is organized as follows. Section 2 discusses the theoretical background of the proposed system.

Section 3 describes the case background of the previous system where the searching problems are identified. We also expressed how in this project the previous

database records are going to be used to transform the static search system into the dynamic search system. While in section 4, we will illustrated and explain the

system prototype, system framework, ontology development and the proposed system interface. This is followed by discussion on related work

in section 5 as various approaches and methodologies have been applied on the construction of ontology and semantic retrieval system. Finally, we

conclude the paper in Section 6.

Theoretical Background

The discipline area of information retrieval has

benefited from unique characteristics of the Semantic Web. According to Colomb (2002), we search for information when we know what we are

searching for. On the other hand, we only browse for information when we do not really know what kind of information we really want. As there are a wide

range of web documents in a variety of disciplines and topic areas available in the web, the importance of information retrieval to support or assist user to

find desired information is becoming increasingly important (Finin et al.). The capability of information retrieval technology is about finding and ranking

relevant documents to meet users' needs and the Semantic Web technology can enrich web documents and resources by annotations in machine-

readable format. It is therefore ideal to combine information retrieval technology with semantic documents.

The machine-processable capability refers to how machine can interact with one another through the use of software agent to generate search results. In

addition, the Semantic Web will also bring meaning to the documents and is capable of understanding the documents to fulfill users' needs (Casely-

Hayford, 2005, Java et al., 2007). Hendler (2005) suggests that not only these web documents have meanings in the Semantic Web; the semantic

technology will also be able to help users in searching for information in the best possible way to suit individual’s needs.

Consider the scenario where we use free text to search for *Dora*, which may result in outputs that relate to a person/ blog/ website/ forum named Dora. Assume

that what we actually intended to search is related to the children show ‘Dora the Explorer’, the characters in the ‘Dora the Explorer’ and the toy

‘Dora the Explorer’. In this case, the returned results of blog by Dora or person with the name Dora is not what we had in mind when we conducted the search.

This simple example illustrates the limitation of free text search in the current WWW environment. The problem may be compounded in

situations where the search is conducted in non-English language context. For example, the English word computer is written as *komputer* in the Malaysian

language (Bahasa Malaysia). For native speakers of Malaysian language, the word *komputer* will be entered as free word search which

may not return any searched results related to computer. Furthermore the use of acronyms such as IT for information technology, KL for Kuala Lumpur or

misspelling of words such as *labtop* for laptop can result in undesired and irrelevant results when query is issued.

Secondly lack of information accuracy is also a common problem encountered in free text search. When the search result is returned, the

results may match the search query. However it is difficult to differentiate between documents that are relevant and desired as compared to documents

that are relevant but not desired, particularly in cases when initial general search criteria is used for general topic area search (Mohd Zaid and Lau, 2008).

Semantic search is yet to be extensively replaced with web-based Boolean search. Relational database is commonly used prior to semantic web technologies.

Common problems encountered in search results are lack of precision, relevance and ranking of returned documents. It still requires human

interpretation when links/urls are imprecise, loosely classified and when there is a lack of machine interpretation capabilities. Human judgment is thus

required in determining which documents are relevant, and which are not. Therefore, using database merely as a medium to search for information will

not return the necessary result that the users seek. In particular, these problems are not only encountered by novice users, the same scenarios

are commonly faced by experienced users such as experienced researchers.

Ontology provides a means in which semantic search

can be implemented. Using ontology to keep data in a dynamic database is considered as an appropriate approach to understand contextual

relationships of term. This term is what we call vocabulary where the data is given a well-defined meaning that is consistent across context.

Furthermore, with the contextual relationships defined in the ontology, more information could be linked without the user realizing the information

primarily subsists. This happens without mapping, transforming or dumping all the records between two tables or ontologies. It is all about semantics. In

addition, when there are two different database server systems which are not cross-compatible, sharing of data between these two domain

knowledge cannot be done unless being defined by ontology (i.e. by adding semantic meaning to the existing data).

Ontology database is also believed to improve search accuracy as search results are the categorization of classes and subclasses of data which are structured

in hierarchy. It is simpler to describe and more dynamic in terms of relationships between classes and subclasses. Semantic searching is capable of

understanding the search intent thus, can overcome the failing of web-based Boolean search.

Currently, implementation of ontologies is found in a variety of development tools and methodologies. Ontology can be built from scratch or reused using

existing ontologies. The preference generally lies with the developer as well as how the ontologies will be applied. Casely-Hayford (2005) has reviewed

extensively on methodologies, languages and tools for building ontologies. To facilitate reuse and sharing of ontology, developers can

refer to existing ontologies in libraries such as DAML Ontologies http://www.daml.org/ontologies/, Protégé Ontology Library

http://protege.stanford.edu, NCBO BioPortal Ontology http://bioportal.bioontology.org/ontologies, and Swoogle http://swoogle.umbc.edu/.

Case Background

This paper describes a proposed semantic search system for a Malaysian university. The proposed

system is developed as a resource to search thesis reports (Honour, Master and PhD) for the Education Faculty. The thesis resource aims to assist students with

quick and easy reference point in which students can search for completed thesis. Currently, hardcopy of the thesis are available for full access in the main library

and softcopy is accessible online with limited information that consists of abstract and first ten pages of the thesis. Occasionally, the faculty department has

a second copy that the students can borrow for reference. The current system uses keywords search to match the title of thesis in a general category

such as Teaching and Learning, Science and Mathematic, Sport Science, Module, Education Studies etc. Common problems

encountered by students include:

(1) having difficulty in using the *right* keyword,

(2) the search results often do not match the expected search outcome, and

(3) search results do not match the desired discipline or topic area (Mohd Zaid and Lau, 2008).

The aim of the system is to assist students who are novice researchers in identifying research topics that have been researched in the past years. Currently,

the faculty consists of five academic departments and each department has its own field of study, courses and research. Sometimes, thesis title reflects

discipline areas that span across different departments (Figure 1).

Figure 1: Possibilities of Thesis-Department Relationships

Please see Figure 1 in full PDF version

For instance, a thesis that investigates computer-based teaching methodology in chemistry subject can be undertaken by a student enrolled in a

Science department studying for a Teaching diploma course in the Education faculty under the supervision of an academic in the multimedia

department. This example shows a one-to-many relationships between title and department. Students who are newly enrolled in the Honour, Master and

PhD programs are often inexperienced researchers. They tend to search based on the academic department that they enrolled in. This way if the

thesis is not categorized under their department then limited results are returned. They are not able to search for thesis that is conducted under cross-

discipline or cross-department. One would think that this problem can be easily resolved by categorizing the thesis under individual

department. However this approach is not recommended and inefficient as the librarian is required to use some classification mechanisms

such as categorizing based on the faculty or department the supervisor is associated with. This is an ineffective method when the number of thesis

records is large. Consequently, it creates inconsistency and difficulty to maintain. We suggest developing an ontology-based information retrieval

system that enables the database to be described based on mind map as shown in figures 2 and 3. Figure 2 shows an example of mind map that was

designed for the proposed thesis ontology in English, and Figure 3 shows the same mind map in the native language (Bahasa Malaysia). The proposed

mind map was designed based on the fields of study in the academic departments of the Education Faculty. Two main classes are identified

as a 1st level ontology classes: teaching (pengajaran) and learning (pembelajaran). In each class of teaching and learning, subclasses are

created to include another keyword of the thesis's title. Under the teaching class, we include approach (pendekatan), motivation (motivasi), resource

(sumber) and perception (persepsi). For each of the sub-classes (for instance, the sub-class of *teaching>approach*), there exists another sub-class

that can be extended from the 2nd level of ontology class (such as CBL, CBI and PBL) by which these sub-classes are related to the 1st

and 2nd level of ontology class of the thesis's title.

Same goes as the *learning* class where it consists of other sub-classes which

are: strategy (strategi), effectiveness (keberkesanan), factor (faktor), performance (prestasi) and skill (kemahiran), which are

expanded from the 1st level of *learning* ontology class. Additionally, for each of these sub-classes (for instance the sub-class of *learning>strategy*), there

are another level of sub-class that can be expanded (such as CAL, self-learning and simulation) to form the 2nd level of ontology class.

Figure 2: Mind Map of Thesis Ontology in English Language

Please see Figure 2 in full PDF version

Figure 3: Mind Map of Thesis Ontology in "Bahasa Malaysia" Language

Please see Figure 3 in full PDF version

The mind map presented here is an improved categorization compared to the current database system. Currently there are twelve categories that

include English, Social Problems, Islamic Studies, Teaching & Learning, Education Management, Sport Sciences, Module, Math and Sciences,

Teaching with Computer, Psychology Studies, Technique & Engineering and General. The major problem associated with current categorization

method is that it is not easy for the librarian to categorize all the theses in these categories manually. Through the mapping process we hope to apply

some form of rules to the system.

Although most of the available theses are written in the “Bahasa Malaysia”

language, however there is one course in the Education Faculty in which it is taught in the English language. This means that the students need to write the

thesis in English. Therefore, there exists a situation where some theses are written in English and some are in the “Bahasa Malaysia” language. It is

important that the students can search all theses, regardless of the language it is submitted, to enable students to identify

previous research topics that have been conducted.

Furthermore, the rising number of international students taking a course in

this University has increased in recent years and we aim to develop a search system that can cater for both languages. For example, when the

keyword “computer” is entered, the search system is supposed to search also the synonym of the keyword “computer” in “Bahasa Malaysia” which is

“computer”
(computer=komputer).

System Development

In the proposed project, user can access the system

through a web server where the appropriate user interface will be created to let the query process easy to be used. Ontology is created using an ontology

editor to build a knowledge repository that could keep all the data for the query. When the datastore is completed and ready to be queried; then it will be

converted into RDF document, which is then parsed and all the data are dumped into a relational database. Apache Tomcat is used as a web services

which will handle all the data records as well as the queries. Further process can be seen in Figure 4 where the system framework is presented.

Figure 4: System Framework

Please see Figure 4 in full PDF version

We are going to create an ontology repository as our dynamic database which is a datastore that can be updated easily through the process of restructuring the

ontology hierarchy. This process of restructuring is required to build the ontology hierarchy that can be referenced as hierarchical categories and

sub-categories. The ontology hierarchy will be presented as drop-down list of the search items to enable users to search for relevant thesis records

based on the concept of mind map as described above.

In this project, we use Protégé 3.4.4 to develop the

ontology classes and subclasses. This stage of ontology development is achieved by referring to the existing records in the MYSQL database. Figure 5

shows a part of the ontology classes/subclasses with the object properties that are created based on existing records from MYSQL tables.

Figure 5: Ontology Relationships

Please see Figure 5 in full PDF version

The thesis records which are already stored in MySQL database are being re-used where the process of mapping the database records into ontology

class/subclasses is required to transform the existing database record into ontology database. This process of dumping data from MySQL into Protégé is

carried out using Protégé Plug-in called *DataMaster* (Nyulas and Tu, 2007). The *DataMaster* plugin can be used with any relational database with JDBC/ODBC

drivers. The importing process is done through its user interface where a connection to relational database has to be login first. Then we can choose to

import either the entire database or some of the tables only. We can also select the desired destination of ontology class to dump the records

into. In our project, we have imported only some of the table content to map into the pre-defined ontology classes and subclasses. Figure 6 shows the user

interface of the DataMaster plugin in the Protégé file. In this step, the mapping process can be done very quickly.

Figure 6: Datamaster Plugin is Used to Map the Database Records into Ontology Classes/ Subclasses

Please see Figure 6 in full PDF version

The proposed system also focuses on the use of synonym in the ontology classes/subclasses. As explained in the previous section of case background,

synonym can be used to overcome the problem of conducting keyword search in foreign language either by local students or international students. For

example, the student may only know the keyword “motivasi” in “Bahasa Malaysia” but do not know the corresponding keyword in English. The Protégé (as

the ontology development tool) provides the synonym function to enable this capability of the search system. Thus, we are able to create all the appropriate

relationships as well as the synonyms to complete the thesis database.

We use the PHP code to store RDF documents into

relational database. RDF documents are parsed and stored as triple store in MySQL database. The retrieval process of the RDF documents is organized in

three columns in the MySQL table which are subject, predicate and object. The process of parsing is being done once only as long as there is no

change in the class/subclass hierarchy and no new data/records are added. Then we use RDQL language to query the

database and return the results in html.

User Interface Query System

Figures 7, 8 and 9 show interface design of the drop-down list. These drop-

down menus are created from the ontology which is defined based on the mind map. With guided keywords, it is easier for the students to select an

appropriate keyword based on ontology. However, this system will also offer open queries (such as open keyword search) which are

preferred by experienced users.

Figure 7 shows the first drop-down that listed the first level of ontology class

that include: *Pengajaran Berpandukan Komputer*, *Pengajaran & Pembelajaran*, *Pendidikan Sains & Matematik*, *Pendidikan Islam*, *Bahasa*

Inggeris, *Sains Sukan* and etc. When we select the second value from the list, another list of value (the name of supervisor) will be propagated that include

“Abdul Hafiz bin Abdullah, Ust”, “Abdul Wahid b. Mukhari, Dr. Hj”, “Abdul Rahim bin Hamdan, En” as shown in Figure 8. After we have selected the value

“Jamaludin bin Harun, En” in this second drop-down list a list of search results will return. The drop-down list is only encountered as two-level category-

subcategories search system. Figure 9 shows the result that matches with 5 records from the database.

Figure 7: System Interface Design: 1st Drop down Box

Please see Figure 7 in full PDF version

Figure 8: System Interface Design: 2nd drop down Box

Please see Figure 8 in full PDF version

Figure 9: System Interface Design: Search Results (Output) Based on the Input of 1st and 2nd Drop down Box

Please see Figure 9 in full PDF version

In another example, consider a user who wants to search “computer-based learning as a learning method among university students”. As discussed

previously, if the students are to enter the above phrase as the search string, it may not return any result. However we can create the following

hierarchy of classes in ontology which can be used to create the drop-down list as discussed above to help users to search for thesis records:

Learning: Method: University: Computer-Based Learning

Using the same approach, a user who wishes to search

for past research topics related to “effectiveness of computer-based learning as a learning method among university students”. The following hierarchy of

classes in ontology can be structured:

Learning: Method: University: Computer-Based Learning: Effectiveness

Therefore, the ontology development is a vital part

of this project. It helps students to guide their search based on how the ontology is categorized.

Related Work

In the literature, there are several works that address the ontology construction approach whether;

(i) to start the ontology development from scratch,

(ii) to transform or migrate database

schema from relational database into existing ontology (also known as a mapping process), or

(iii) to join (merge) two different existing ontologies (Taylor et al., 2005, Hitzler et al., 2005, Ali et al., 2005).

Stojanovic et al. (2002) propose an approach based on semi-automatic generation of ontology from a relational database model using a F-logic

inference engine. This ontology generation approach first transforms the relational database model into equivalent class structure in ontology which

then maps the content of the database into ontology. However, user intervention is required to choose the most suitable mapping rules to apply.

Other researchers propose reverse engineering techniques (Bussler et al., 2004, Meersman et al., 2005) of mapping

relational databases into ontology. The process also involves semi-automatic ontology extraction, which creates ontology classes/subclasses that

corresponds to the relational database content migrating from the HTML forms which need validation.

However, in our project, the creation of ontology is done manually, where the ontology classes/ subclasses construction is not directly extracted from

the database table. We have created the ontology based on mind map which is based on existing thesis titles in the database.

Conclusion

In this paper, we have presented a system to help novice researchers in searching for information.

We have proposed a framework that shows that ontology approach can help novice researchers to apply semantic search techniques to improve current search

capabilities. Preliminary user interface with simple ontology has been designed. The next phase of the project is to review alternate approach of

developing mind map and to evaluate the ontology design.

Acknowledgment

This work was conducted using the Protégé resource, which is supported by grant LM007885 from the United

States National Library of Medicine

References

Astrova, I. (2004). "Reverse Engineering of Relational Databases to Ontologies," *The Semantic Web:*

Research and Applications. Springer Berlin / Heidelberg.

Benslimane, S. M., Malki, M. & Bensaber, D. A. (2005).

“Automated Migration of Data-Intensive Web Pages into Ontology-Based Semantic Web: A Reverse Engineering Approach,” *On the Move to Meaningful*

Internet Systems 2005: Coopis, Doa, And Odbase. Springer Berlin / Heidelberg.

Casely-Hayford, L. (2005). "A Comparative Analysis of Methodologies, Tools and Languages Used for Building Ontologies," Swindon, *CCLRC*.

Colomb, R. M. (2002). Information Spaces: The Architecture of Cyberspace, New York, *Springer*.

Defense Advanced Research Projects Agency

(DARPA). "DAML Ontology Library," Retrived 25 February 2011. http://www.daml.org/ontologies/

Finin, T., Mayfield, J., Joshi, A., Cost, R. S. & Fink, C. (2005). "Information Retrieval and the Semantic Web," Proceedings of the 38th International

Conference on System Sciences.

Hendler, J. A. (2005). "Knowledge is Power: A

View from the Semantic Web," *AI Magazine*, 26, 76.

Hitzler, P., Krotzsch, M., Ehrig, M. & Sure, Y. (2005). "What Is Ontology

Merging?," *American Association For Artificial Intelligence*.

Java, A., Nirneburg, S., Mcshane, M., Finin, T.,

English, J. & Joshi, A. (2007). “Using a Natural Language Understanding System to Generate Semantic Web Content,” *International Journal on*

Semantic Web & Information Systems, 3, 50-74.

Kim, J., Jang, M., Ha, Y.-G., Sohn, J.-C. & Lee, S. J.

(2005). "MoA: OWL Ontology Merging and Alignment Tool for the Semantic Web," *Innovations in Applied Artificial*

Intelligence. Springer Berlin / Heidelberg.

Mohd Zaid, N. & Lau, S. K. (2008). 'Improving the Internet Search Capability

by Semantic Technology,' Proceedings of The 4th International Conference On Information Technology And Multimedia (Icimu' 2008). Uniten Malaysia.

NCBO BioPortal: Ontology Listing. Retrived 26 February 2011. http://bioportal.bioontology.org/ontologies

Nyulas, C., O'Conner, M. & Tu, S. (2007). “Datamaster – A Plug-In for Importing Schemas and Data from Relational Databases into Protégé,” In Proceedings of

10 the International Protégé Conference.

Stojanovic, L., Stojanovic, N. & Volz, R. (2002). “Migrating Data-Intensive

Web Sites into the Semantic Web," *Symposium on Applied Computing,* Madrid, Spain Acm New York, Ny, Usa.

Taylor, J. M., Poliakov, D. & Mazlack, L. J. (2005). “Domain-Specific Ontology Merging for the Semantic Web,” *Fuzzy Information Processing Society*, 2005.

Nafips 2005. Annual Meeting of the North American.

UMBC Ebiquity Research Group. "Swoogle," Retrived

25 February 2011. http://swoogle.umbc.edu/

World Wide Web Consortium (W3C). "W3C,"

Retrived 17 February 2011. http://www.w3.org/