1-1-2011

# Ontology based search mechanism in bilingual database resource

Norasykin Mohd Zaid
*University of Wollongong*, nmz056@uowmail.edu.au

Sim Kim Lau
*University of Wollongong*, simlau@uow.edu.au

# Ontology based search mechanism in bilingual database resource


## Abstract

The focus of our research is to investigate how ontology based search system can help a novice researcher in literature search in an online academic database. In this paper, we will describe ontology construction which is based on the existing relational database system. We will demonstrate three solutions using this ontology to resolve our problems in synonym relation, short-form-term and full-term search, and bilingual retrieval.



## Keywords

database, bilingual, mechanism, search, resource, ontology

## Disciplines

Physical Sciences and Mathematics

## Publication Details

Mohd Zaid, N. & Lau, S. K. (2011). Ontology based search mechanism in bilingual database resource. IDSI-APDSI 2011: Proceedings of the 11th International Conference Decision Sciences Institute and the 16th Annual Conference of Asia-Pacific Decision Sciences Institute (pp. 171-176). Taiwan: National Chengchi University.

# ONTOLOGY BASED SEARCH MECHANISM IN BILINGUAL DATABASE RESOURCE

**Norasykin Mohd Zaid, University of Wollongong, Australia, nmz056@uowmail.edu.au**
**Sim Kim Lau, University of Wollongong, Australia, simlau@uowmail.edu.au**

## ABSTRACT

The focus of our research is to investigate how ontology based search system can help a novice researcher in literature search in an online academic database. In this paper, we will describe ontology construction which is based on the existing relational database system. We will demonstrate three solutions using this ontology to resolve our problems in synonym relation, short-form-term and full-term search, and bilingual retrieval.

**Keywords:** Semantic search, ontology, bilingual search, and synonym.

## INTRODUCTION

Information and knowledge are increasingly becoming shareable and searchable resources, particularly in the current digitized world. Since 1996, the World Wide Web has become a primary source for information which offers online resources that are available 24/7. However, there is a need for a searcher to carry out query syntax for information retrieval process. Searcher has to form his own query and sometimes the query needs to be expanded. Thus searching for online resources becomes a challenging task especially for novice researcher. Studies have shown that outcome of search query is influenced by how queries are written [18, 22]. Queries conducted using advance search and Boolean techniques improves search outcomes compare to only using basic search. However using advance search and Boolean techniques require learned skills and is more complex than basic search that requires minimum skill. Basic search often uses keyword to enable users to generate a search term that best describes the query. However it is not an easy task to know what keyword to use; sometimes it is difficult to think of the correct term or word to use if the user does not have the domain knowledge. Thus it is not uncommon to find user struggles to obtain relevant search outcomes even after several attempts [16].

A number of techniques have been used to improve relevance of information retrieval. One of them is faceted search which enables categorization of documents based on more than one facet such as author, subject, document-type and so on. Another commonly used technique is Boolean search. More recently information retrieval has benefited from advancement in the Semantic Web [13]. Ontology, a set of representational primitives that include information about their meanings defined by classes, properties, relationships and constraints to model a domain of knowledge, plays a pivotal role to enable knowledge sharing and reuse in the Semantic Web [21].

This paper describes our research to develop an ontology-based search mechanism to assist university students in performing search query; in particular we focus on formulating query in a bilingual environment.

In the remainder of this paper, we organize the paper as follows. The second section presents literature review while the third section describes case background and motivation of this research. In section four, we illustrate sample outputs of three queries and we explain system development process in section five. Section six concludes the paper.

## LITERATURE REVIEW

There are many languages used around the world and the needs to search information in language other than English become important [11]. There are two common systems that focus on the search language and contents; they are Cross-Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR). CLIR is a search system where the target information which listed in the search result is different from the language uses in the query search box, while MLIR is to have one language in a query search box and the target information contain more than one language. Thus both the MLIR and CLIR applications require translation process of the language used [14]. Either the translation is to convert the current query language to the desired language of the retrieved documents or the translation of the documents themselves is based on the desired language the user requested for. Since the cost of translating documents is more expensive than a single translation of keyword search, many researchers of the CLIR and MLIR applications prefer to use query translation approach rather than documents translation [14].

Yilu *et al.* [27] identify three query translation

approaches which translate queries into each target language before matching and retrieving related documents. These three approaches that can be applied into CLIR and MLIR systems are: using machine translation based (MT-based), parallel corpus, and bilingual dictionary. Finin *et al.* [14] stated in their paper that by using multiple MT-based systems and the question sentence, translation correctness of the keywords term can be improved. There are some other researches also trying to improve the features of CLIR [19, 25] and MLIR [14, 27]. Kim *et al.* [20] have used semantic category tree and collocation technique to improve precision of the CLIR system.

In online database experience, Xie [26] investigate ways to support both user control and ease of use for novice and skilled users in online information retrieval system. Lau and Goh [12] propose three methods to improve quality of the Online Public Access Catalog (OPAC) system: (i) using interactive query reformulation by recommending alternative query terms, (ii) browsing selected items in the search result list through hyperlink records, and (iii) provide content-sensitive assistance which can identify user's searching skill.

Other studies have been conducted to address general web search problem and online database experience [6, 10, 15, 26]. Full text search has been reported as incompetent method for database information search [6]. This type of search has been unable to deal with problems of synonym, homonym, aboutness and incognito that mostly happen when dealing with the academic libraries. Chen *et al.* [10] report that sense of control of doing activities on the web generally is influenced by how the web environment is such as users finding it easy to use the web search if they are familiar with the environment and have the sense of being able to control the web environment. For example, users go for deep search when their first search leads them to more information and they know how to go to next step. Xie [26] describes problems of online database and web search engines when categories of presentation are not well structured and irrelevant results are obtained. In addition, Griffiths and Brophy [15] explain how successful information search often relied on having some check point item to support results.

Beall [6] identifies synonym relation problem. For example, the words 'mum', 'mother' and 'mummy' all refer to the same meaning of 'mother'. It is important to consider a range of words when using open keywords. Variant spelling is also another problem; for instance, spelling variant between American and British words; color/colour. This kind of problem can be overcome by adding both words when querying. However this approach can become complex when more than one words are involved. Short form term has made it easier to refer as in business name, centre names, familiar terms and others such as WWW for World Wide Web, USA for United States of America. As the short form term is often represented by the first letters in the sentence names, the possibilities of having the same short form term cannot be discounted. For example, COM may be a shortened form for "communication" or "commercial" or even can be a short form term for "Center of Momentum".

Another problem of full text search in information retrieval area involves multi language. Previous traditional database systems are unable to process any query that used a keyword term in another language and request to retrieve the document in a different language. For instance, network in English language is referred as "rangkaian" in the Malay language. Common databases are unable to search documents which are written in the Malay language if the keyword used is in English language since the system used keyword matched approach to search for related results. In addition, the search system is also unable to deal with homonym problem where the word has more than one meaning. For examples, the word "sort" can refer to categories or a kind of things or it may refer to "putting elements in a certain order" as in computer science and mathematics. The various types of problems described here are compounded by machines and computers unable to understand meaning of words as that understood by human.

Relational database is commonly used to store all database records. Records are kept in the database to enable search by the query engine. However, common problems encountered in the full text search results include its lack of precision, relevance and sorting of returned documents result in most related documents cannot be organized into specific published dates, authors name, titles or even subjects [6]. Very often the query engine can retrieve a list of search results but it still requires human interpretation when links/urls are imprecise, loosely classified and lack of machine interpretation capabilities. Human judgment is required in determining which documents are relevant and which are not. Ontology provides a means in which semantic search can be implemented. Ontology enables contextual relationships in the database to be defined and data is given a well-defined meaning that is consistent across context. Furthermore, with the contextual relationships defined in the ontology, more information can be provided to describe ontology domain.

Semantic search is yet to be extensively replacing web-based Boolean search. The number of information that is machines-understandable is not high enough to be used through the semantic search

engine for general purposes. However, there is an increasing number of semantic web projects developed using semantic approaches and technologies such as Semantic web portal (Swoogle, Hakia, GoPubMed, etc), corporate semantic web such as Corese (http://www-sop.inria.fr/acacia/soft/corese/) and social semantic web such as SIOC (http://sioc-project.org/) and Twine (http://www.evri.com/). In fact most of the industry groups and individuals who are pertaining to the dynamic search environment are considering building ontology-based database records.

As explained, ontology provides a means in which semantic search can be implemented. Ontology database can improve search result accuracy using structured categorization of classes and subclasses which enable contextual relationships to be defined and data be given a well-defined meaning consistent across context. Thus semantic search that is capable of understanding the search intent can overcome failing of traditional web-based full text search.

Currently implementation of ontology is found in a variety of development tools and methodologies. Ontology can be built from scratch or reused from existing ontologies. The preference generally lies with the developer as well as how the ontologies are applied. Casely-Hayford [9] 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 NCBO BioPortal Ontology [2], Protégé Ontology Library [4], Swoogle [3] and DAML Ontologies [1].

There are several works that address the ontology construction approach. Some of the approaches include (i) starting ontology development from scratch, (ii) transforming or migrating database schema from relational database into existing ontology using a mapping process, or (iii) joining or merging two different existing ontologies together [5, 17]. Stojanovic *et al.* [7] 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 the ontology, which then maps the content of the database into ontology. User intervention is required to choose the most suitable mapping rules to apply. Other researchers propose reverse engineering techniques of mapping relational databases into ontology [8, 23].

## CASE BACKGROUND AND MOTIVATION

The focus of our research is to investigate how to help a novice researcher in conducting literature search as a first step of choosing a research project or research thesis. In the university under our investigation, students are required to complete a one semester research project to complete their degree requirement. However, students who embark on research project for the first time often find identifying a viable research topic to be a daunting process. Choosing a research topic for thesis and dissertation is a complex process. The students not only need to have some interests in the topic area, they also need to choose a topic area that will make contribution to the area of research. It is not uncommon for students to seek advice from academic staff from the department they enrolled in; they will also search existing literature to determine how much information is already available on the topic. Usually students also search previous dissertations or thesis available in the library to identify potential research topic. This approach not only functions as a good starting point, it also enables student to develop an initial focus on the research topic that s/he wishes to embark on.

Very often these students are novice inexperienced researchers who may not be skilful in searching for information. Although students have undergone at least three years of undergraduate study before embarking on the research programme and have acquired some degrees of literature search skills or using library services, very often their literature search experience is limited. For example students often are given a list of bibliographies or reference lists in the subject area by their lecturers, which provide a good starting point when literature search is conducted. This may not be the case when students need to search for a thesis or dissertation topic that is of interest to them as well as a doable research topic that can be completed within the specific time frame of their degree. Thus a search and retrieval system that can help students in this initial phase of identifying research topic is desired.

## SYSTEMS DEVELOPMENT

The ontology is considered as our main database resources for the proposed online thesis searching. We decide to develop an ontology-based search system based on the existing relational database. We aim to develop an ontology that can overcome problems encountered in keyword search such as synonym, short form term and bilingual query.

Figure 1 shows a partial snapshot of ontology used in this project. The ontology is designed based on the scope of keywords used in the Education Faculty as well as by conducting examination of keywords used in past thesis database. We have made teaching, learning and level as the three main classes. This is

because courses and research areas in the Faculty deal with teaching and learning and preparing students for future teaching career at five levels of education (primary, secondary, matriculation, diploma and university) in the country. Research on teaching and learning is conducted based on the level of education. The organization of ontology in this way enables us to move away from the department-course hierarchy. In the teaching and learning classes we have included eight sub-classes, which are approach, factor, learner, performance skill, strategy, subject and tool. The determination of these eight sub-classes is based on common themes in which research thesis have been conducted in the past.

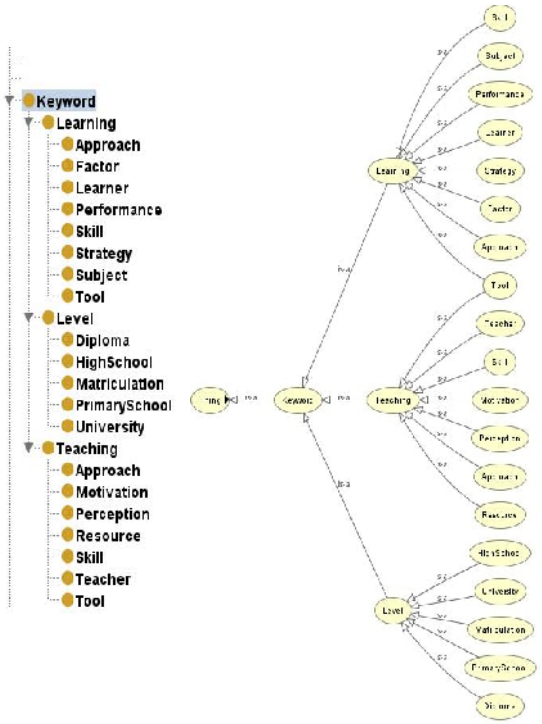


Figure 1: A partial snapshot of hierarchy of thesis ontology.

The advantages of classifying ontology this way is to enable students to have a unified structure when searching the thesis database. When a student conduct a query, what it has in his/her mind is "I am interested in investigating research issues on teaching of mathematics for secondary school children because that is what I am trained for and I like to investigate strategies or approach of teaching mathematics using computers". This query can be illustrated in the following hierarchical form:

*Teaching: Approach*

This way the students are not searching the database based on keywords that were included in the index of the system. On the other hand, we are using our own formulation of relationship in determining the flow on how to search for information. We call this dynamic search which can offer flexibility how query is made.

Consider an example when the user is searching for past research topics that are related to "teachers' perceptions of delivering lessons using computer assisted learning". Using the ontology as described in Figure 1, we can structure the search as follows:

*Teaching: Approach: Computer*

Consider another example where the user wants to search "compare learning styles second year education students compare with the teaching styles of their lecturer". Using the above ontology the search can be formulated as:

*Teaching: Approach: Style*
*Or*
*Learning: Approach: Style*

The next step after the ontology is created is to consider bringing in all records into appropriate class-subclasses properties. The process of dumping data from MySQL into Protégé is carried out using Protégé Plug-in called "Datamaster" [24]. The importing process is done through its user interface where a connection to relational database is established. We have an option to import either the entire database or only some of the tables. In our project, we used the option to bring the entire database.

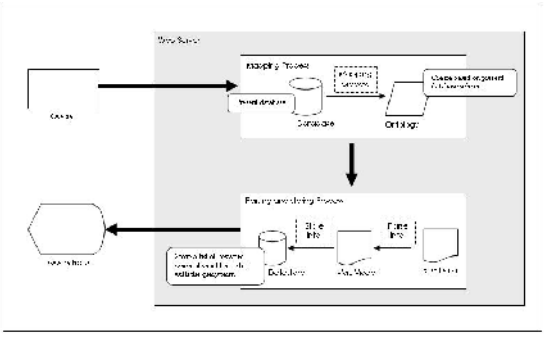

Figure 2: System framework

Figure 2 shows the system framework of the proposed system, which includes a process of mapping current database schema into ontology. The ontology database is converted into RDF data and parsed into XML model. Once the query has been submitted the result is returned in a list of URL links that contain the thesis. After the mapping process is completed, we set the relationship of the class-subclasses and set the synonym relation, create common short form terms for individuals and managed the English-Malay related keywords.

As described in the previous section, it is desired that

the system is able to handle bilingual query. This can be achieved as the system uses synonym to meet this requirement. Users do not have to conduct separate searches using keywords in different languages. For example, the first query is to use the phrase "motivation of learning" and then a second query is issued using "motivasi pembelajaran", which means motivation of learning in Malay (Bahasa Malaysia) if keyword search is used. Instead synonym can be used to identify corresponding words in either language. Similarly synonym can be used to define acronym such as T&L for teaching and learning, PBL for problem-based learning, CAI for computer-aided instruction and so on.

## SYSTEMS OUTPUT

We use the following three scenarios to show the difference between keyword search and ontology search.

**Scenario 1 - Information retrieval based on shortened forms of terms (UHB1412-English for Academic Communication)**

In this scenario the student likes to know any research projects that are related to the subject UHB1412 "English for Academic Communication". The query result should retrieve title that has the keyword either 'UHB1412' or subject name "English for Academic Communication". Figure 3 shows the output result.

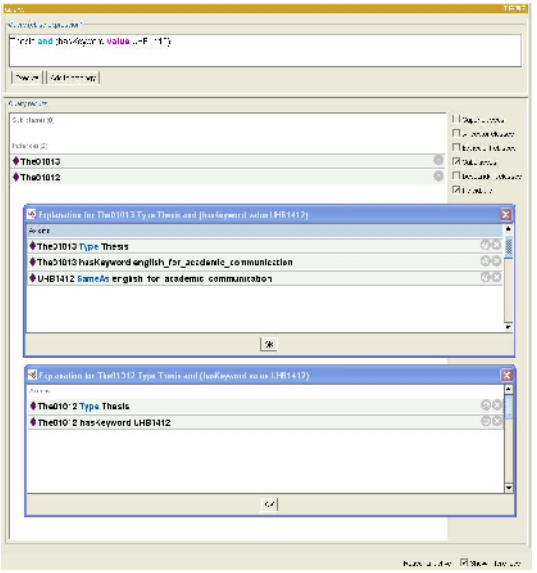

Figure 3: Query 1 - shortened forms of terms

We can see two records have been retrieved from the query syntax in Figure 3 as follow:

***The01012***
*"A Survey on the Language Learning Strategies Used by the UHB1412 Students"*

***The01013***
*"A Survey on the Language Learning Strategies Used by the Students of English for Academic Communication"*

These two records were retrieved from the domain ontology as we give meaning to both individuals/instances which the reasoner can understand the relationship between them. Figure 4 shows how the relationship is represented between these two words which we have defined using the synonyms relation of 'SameAs'.

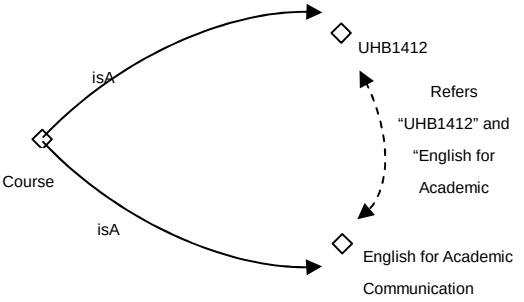


Figure 4: Relationship of "SameAs" Individuals

**Scenario 2 - Information retrieval based on different language or dialects (English-Malay Language)**

In this university we have two cohorts of students, domestic local students and international students. The domestic local students are fluent in the local language of Malay and the international students use English as the medium of instruction and communication.

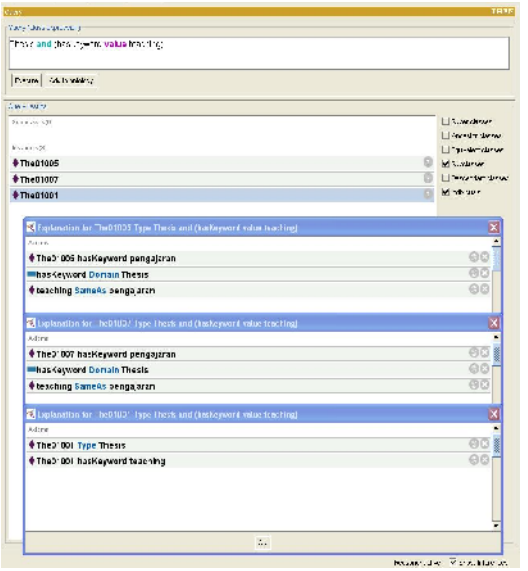

Figure 5: Query 2 - different language or dialects

In this scenario, the student wishes to identify all theses, written in Malay and English, in the research topic of 'teaching'. For local students they will

probably choose to use the keyword 'pengajaran', whereas for the international students they will use the keyword 'teaching'. Thus the query that can handle both languages is desired. Figure 5, shows the result in which three records are returned as follow.

> *"Penggunaan Komputer Dan Internet Dalam Pengajaran Dan Pembelajaran Di Kalangan Guru Sekolah Menengah Di Daerah Pasir Mas, Kelantan"*
>
> *"Pendekatan Pengajaran Yang Digunakan Oleh Guru Sekolah Menengah Di Daerah Johor Bahru Dalam Pengajaran Dan Pembelajaran Matematik"*
>
> *"A Comparative Study On The Learning Styles Of Second Year Education (living Skills) Students And The Teaching Styles Of Their Lecturer"*

**Scenario 3 - Information retrieval based on synonym relation (Intelligent: AI, Smart)**

In this scenario, the student likes to conduct a project development on some form of smart application and would like to know whether any research has been conducted in this area. Thus the student will think of possible keyword such as 'smart application'. However, the keyword of 'smart' is not the only word that can describe smart application. There are some other words that are capable to describe smart application such as AI (artificial intelligence) or intelligent. In our project we use synonym relation to give more meaning to data.

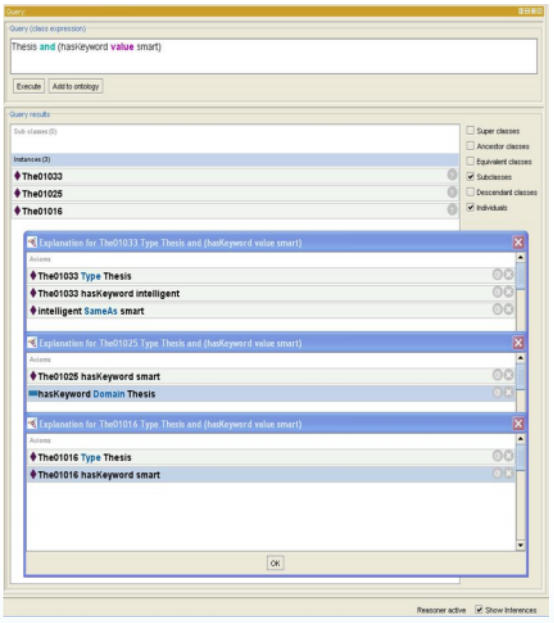


Figure 6: Query 3 – synonym keyword search

Figure 6 shows the results that include all the synonym words that have been refer to the keyword term of 'Intelligent' which list as follow:

> *Aplikasi Sistem Pengurusan Dan Pentadbiran Sekolah Bestari (Smart School Management System - ssms). Tinjauan Di Sebuah Sekolah Rintis Bestari Di Kota Tinggi, Johor.*
>
> *Sistem Pencegahan Kecurian Dan Rompakan Kenderaan 'smart System'.*
>
> *Intelligent software system for CD-ROM project archives.*

## CONCLUSION AND FUTURE RESEARCH

Our research is based on the concept of ontology information retrieval. In our preliminary work, we have utilized existing database system to map into new ontology schema. Synonym-relation word, short-form term and bilingual query term are included in the domain ontology. Outcomes have shown that search can be performed to meet user queries. In the future, we are going to work on developing web user interface to allow an online archive of the database records. We will consider generating a dynamic drop-down keyword search to help novice researchers in their search process. In addition, we will conduct our investigation by examining various structure of domain ontology in the search process.

## REFERENCES

[1] *DAML Ontology Library*. [cited 25 February 2011]; Available from: http://www.daml.org/ontologies/.

[2] *NCBO BioPortal: Ontology Listing*. [cited 26 February 2011]; Available from: http://bioportal.bioontology.org/ontologies.

[3] *Swoogle*. [cited 25 February 2011]; Available from: http://swoogle.umbc.edu/.

[4] *Welcome to the Protege Ontology Library!* [cited 4 February 2011]; Available from: http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library.

[5] Ali, M., F. Esposito, J. Kim, M. Jang, Y.-G. Ha, J.-C. Sohn, and S.J. Lee, *MoA: OWL Ontology Merging and Alignment Tool for the Semantic Web*, in *Innovations in Applied Artificial Intelligence*. 2005, Springer Berlin / Heidelberg. p. 116-123.

[6] Beall, J., *The Weaknesses of Full-Text Searching*. The Journal of Academic Librarianship, 2008. 34(5): p. 438-444.

[7] Bozsak, E., M. Ehrig, S. Handschuh, A. Hotho,

A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. KAON - Towards a large scale Semantic Web. in *Third International Conference on E-Commerce and Web Technologies*. 2002. London, UK: Springer-Verlag.

[8] Bussler, C., J. Davies, D. Fensel, R. Studer, and I. Astrova, *Reverse Engineering of Relational Databases to Ontologies*, in *The Semantic Web: Research and Applications*. 2004, Springer Berlin / Heidelberg. p. 327-341.

[9] Casely-Hayford, L., *A comparative analysis of methodologies, tools and languages used for building ontologies*. 2005, CCLRC: Swindon.

[10] Chen, H., R.T. Wigand, and M.S. Nilan, *Optimal experience of Web activities.* Computers in Human Behavior, 1999. 15(5): p. 585-608.

[11] Chu, H., *Information representation and retrieval in the digital age*. 2003, Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today.

[12] Eng Pwey, L. and G. Dion Hoe-Lian, *In search of query patterns: a case study of a university OPAC*. Inf. Process. Manage., 2006. 42(5): p. 1316-1329.

[13] Finin, T., J. Mayfield, A. Joshi, R.S. Cost, and C. Fink, *Information Retrieval and The Semantic Web*. Proceedings of the 38th International Conference on System Sciences, 2005.

[14] Frederking, R.E., K.B. Taylor, F. Lin, and T. Mitamura, *Keyword Translation from English to Chinese for Multilingual QA*, in *Machine Translation: From Real Users to Research*. 2004, Springer Berlin / Heidelberg. p. 164-176.

[15] Griffiths, J.R. and P. Brophy, *Student Searching Behavior and the Web: Use of Academic Resources and Google.* Library Trends, 2005. 53(4): p. 539.

[16] Guinee, K., M.B. Eagleton, and T.E. Hall, *Adolescents' Internet Search Strategies: Drawing Upon Familiar Cognitive Paradigms When Accessing Electronic Information Sources.* Journal of Educational Computing Research, 2003. 29(3): p. 363-374.

[17] Hitzler, P., M. Krotzsch, M. Ehrig, and Y. Sure, *What Is Ontology Merging?* American Association for Artificial Intelligence, 2005.

[18] Jansen, B.J. and U. Pooch, *Web user studies: A review and framework for future work.* Journal of the American Society of Information Science and Technology, 2001. 52(3): p. 235 - 246.

[19] Kapetanios, E., V. Sugumaran, and D. Tanase, *A parametric linguistics based approach for cross-lingual web querying*, in *Data & Knowledge Engineering*. 2008, Elsevier. p. 35-52.

[20] Kim, T., C.-M. Sim, S. Yuh, H. Jung, Y.-K. Kim, S.-K. Choi, D.-I. Park, and K.S. Choi, *FromTo-CLIR™: Web-based Natural Language Interface for Cross-Language Information Retrieval*, in *Information Processing and Management*. 1999, Elsevier Science Ltd. p. 559-586.

[21] Liu, L., M.T. ÖZsu, and T. Gruber, *Ontology*, in *Encyclopedia of Database Systems*. 2009, Springer US. p. 1963-1965.

[22] Malik, A. and K. Mahmood (2009) *Web search behavior of university students: a case study at University of the Punjab*. 6,

[23] Meersman, R., Z. Tari, S.M. Benslimane, M. Malki, and D.A. Bensaber, *Automated Migration of Data-Intensive Web Pages into Ontology-Based Semantic Web: A Reverse Engineering Approach*, in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*. 2005, Springer Berlin / Heidelberg. p. 1640-1649.

[24] Nyulas, C. and S. Tu. DataMaster – a Plug-in for Importing Schemas and Data from Relational Databases into Protégé. in *In Proceedings of 10 th International Protégé Conference*. 2007.

[25] Wang, Y.-C., T.-H.R. Tsai, and W.-L. Hsu, *Web-based Pattern Learning for Named Entity Translation in Korean-Chinese Cross-Language Information Retrieval.* Expert Systems with Applications, 2008.

[26] Xie, H.I., *Online IR system evaluation: online databases versus Web search engines.* Online IR system evaluation, 2004. 28(3): p. 211 - 219.

[27] Yilu, Z., Q. Jialun, C. Hsinchun, and J.F. Nunamaker. Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal. in *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*. 2005.