Semantic Search Engine in Institutional Repository: An Ontological Approach

Maizatul Akmar Ismail, Mashkuri Yaacob, Sameem Abdul Kareem and Ahmad Haris Abdul Halim
Faculty of Computer Science and Information Technology
University of Malaya, 50603 Kuala Lumpur, Malaysia
E-mail: {maizatul, mashkuri, sameem}@um.edu.my; aharis_halim@yahoo.com

Abstract

In Malaysia, the majority of electronically enabled academic related literature have increased in number. As the concept of Institutional Repository (IR) is relatively new, these collections of literature would definitely be a contributor to the IR contents. A recent study shows positive feedback from students towards establishing an IR in local universities. This paper explores the problems in searching text-based content in the context of IR. The problem with generic search engine is that it will retrieve irrelevant results and users need to spend more time filtering the information that will suit them most. The absence of semantic relations in the collections of academic related publication would result in a penalty as the process of exploring the contents will expand to all occurrences of a given word(s) and may retrieve irrelevant information as term(s) can have different meanings. Clearly, a specialize search engine which can elaborate search queries semantically to find conceptual relations between documents (i.e. the ontology) and retrieve related articles in a more efficient way is highly in need. This paper discusses related researches on providing semantic search in text corpus. Next, a proposal in refining an ontology construction process is discussed. Finally, the paper concludes with future directions of research in this area.

Keywords: Institutional Repository (IR; IR Content; Semantic Search; Ontology

1. Introduction

Over the last two decades, the access to the Internet has tremendously changed the way of communicating the information to the end user. Scholars in particular, are among those users who benefits from the advent of the World Wide Web (WWW) where the process of assembling and disseminating the information became easier. Within this, an Institutional Repository (IR) became a free platform for knowledge sharing among the academia around the globe. The content varies from text-based documents such as electronic theses, research report, pre-print, post-print, departmental paper, e-book and samples of successful grant applications.

Semantic Web technology have further extended the capabilities of current WWW by providing access to the "deep web" in which the digital content are not directly accessible by generic search engines. An Institutional repository falls under this category where the semi-structured content of the metadata that describes scholarly content such as author, title and abstract can be retrieved with higher precision. Within this periphery, a semantic relation (ontology) is constructed within the journal collection published by Faculty of Computer Science & Information Technology (FCSIT), University of Malaya to enable semantic search based on specialized content in the Computer Science discipline. An ontology can provide a taxonomy or classification which can improve the query and retrieve a result which is beyond keyword matching. For instance, a query on the article on "Knowledge Modeling" can yield articles on Ontology and Problem Solving Methods as a result even though these phrases are not explicitly mentioned anywhere in those articles. In realizing this, external knowledge from human and other resources are used. The processes involved are discussed in section 3.0.

The studies on ontology construction often start with knowledge modeling and representation according to the expert's requirement. A set of competency questions is usually used as a guide for ontology construction. At a later stage, these questions will be tested to see if the ontology is able to give an accurate answer. For our research, we seek students' requirement in searching academic-based content. These requirements include the identification of:

- Expert(s) on specific area such as Grid Computing, Fuzzy Logic, Information System, and other areas in Computer Science.
- Emerging research trends in Artificial Intelligence, Information Science, Software Engineering and Networking.

Our study also shows that students need more than what is provided by keyword-based search engines which support our hypothesis that students do need extra support while searching for information.

2. Related Work

(a) Open Access Principle and Students' Need

Wilkinsky (2006) pointed out that an open access principle is "*a commitment to the value and quality of research carries with it a responsibility to extend the circulation of such work as far as possible and ideally to all who are interested in it and all who might profit by it*". Considering the fact that postgraduate students are among the active users of IR that gain benefit because of its open access principle, it is important for us to acquire the issues of IR from the perspective of students starting with their concern of searching for information in general, wish-list for the content of IR, their willingness to be part of IR contributor and their specific need while searching for information with regards to their study. For this purpose, the study by Pickton and McKnight (2006) is use as a guide. While their research has considered postgraduate students as contributors, the roles of undergraduate students as possible contributors have yet to be observed. We have extended the study of Pickton (2006) and include undergraduate students to play a role in our study, together with postgraduate students. An analysis of the results will indicate a specific requirement in enhancing the search of scholarly publication.

Before the actual survey was conducted, structured interviews with three postgraduate students were done. We started the interview by giving them an introduction and the purpose of the interview i.e. to develop an understanding of users' perspective about IR concepts and their searching behavior in IR. We need to obtain this information because an informal interview done with selected academic staff shows that they want support while searching, not only getting information from what is queried based on keywords but also hidden information such as emerging trends in research and specialist search. Due to the limitation of space, we will not discuss the overall result of the interview and the questionnaire's analysis in this paper. Thus, from the analysis of the interview, we formulate a questionnaire to elicit students' perspectives towards IR and the need-and-want while searching scholarly contents. The survey took place at the Faculty of Computer Science and Information Technology, University of Malaya with 77 undergraduates (Final Year Students) and 17 postgraduates' students. Data in different formats such as datasets, final year system prototype, subject specific individual and group projects and source codes are marked by students to be part of IR contents. As a whole, we found out that the majority of students understand the concepts of open access and are willing to comply if they were asked to participate in IR.

(i) Searching Strategy

Searching, defined as "to look or inquire carefully" (Merriam-Webster Online Dictionary) is considered as a significant activity in the web and digital library. Guha et al. (2003) further categorized the searching activity into two:

- Navigational search: In this type of search, user will supply the search engine with word or phrase which he wishes to find in the document. For instance, user might supply the combination of words such as "session 2 presentation ICOLIS" in order to obtain a particular document or web page. The query does not provide any concepts or meaning and the user's intention is to merely find a document which contains those words.
- Research search: the user will supply the search engine with a word or phrase about which he intends to find more information. The user might have a clue or no clue at all on the type of documents that he will retrieve as a result. More exactly, the user is making an effort to collect a few documents together which will finally give him the information that he want. For instance, the user might supply a phrase such as "Knowledge Reuse" to search the collection. The query provides meaningful concepts which are related to other concepts such as Knowledge-Based Classification, Knowledge Sharing, Ontology Extraction, Problem Solving Methods etc. The queries will not only yield results based on keywords but also documents which contain those concepts.

The majority of the undergraduates (67, 87%) and postgraduates (13, 76.5%) start searching for information by typing any keyword(s) which they think is/are related to their study. The action here is inline with Guha et al (2003) study which regards this search activity as navigational search. 52 (67.5%) of the undergraduates and 11 (64.7%) of the postgraduates tend to look for keyword under specific domain where they will first search for a phrase, look at the results and refine the keywords for a second search based on the results of the first search. In other words, their searching activity falls under the research search. It is our intention to provide semantic-based search to help students in the information seeking process. The survey shows that students do need support while searching. Jansen's (2000) study points out that the behavior of web searchers follows Zipf's Law (1949) of "least cognitive effort". The term is borrowed by Mann (1987) which describes that in designing libraries, one must provide the most convenient way to assist users in the information seeking process. Griffiths (1996) states that "increasing the cognitive burden placed by the user … can affect great successful retrieval of information". He further explains that less action from the user will decrease the possibility of making errors and greater success will be achieved (in finding relevant information).

The final question is on the reasons that motivate searching the IR content for the undergraduates and postgraduates if it is going to be implemented in the University of Malaya. Among the factors that will motivate students to search in IR are free to use, which are rated the highest by both groups, followed by efficient keyword search, user friendly interface and full-viewed display of the content. It is interesting to observe that emerging research trend, query expansion and expert identification are also marked by more than half of the respondents as motivating factors. These tasks can be achieved by proper construction of domain specific ontology which will provide a periphery that realized a repository which can be searched through the use of multiple terms (Tariq et al., 2003).

(b) Semantic Web

In 1999, Tim Berners-Lee envisioned the concept of Semantic Web as the following: "I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize".

Berners-Lee et al. (2001) further elaborated the term "Semantic Web" as structured information on the Web, which is intended to be read by software agents rather than by human beings. Since then, much research towards realizing Berners-Lee vision had been done to improve the web representation from the mere HTML pages that only provide documents' description and link between them to semantic-based language such as Resource Description Framework (RDF) that can describe arbitrary things such as a relation of a specific document with author (people), institution (place) and conferences (events). In realizing this, it requires a set of associated, standardized rules and terminologies in the form of ontologies.

Ontology, which originated from the branch of philosophy during the era of the great Aristotle, is the science of what is, of the kinds and structures of the objects, properties and relations in every area of reality. In simple terms it seeks the classification of entities (Smith, 2003), (Guarino & Giaretta , 1995). Applying the concept of ontology from the philosophers, the computer scientist definition of ontology is a "systematic arrangement of all of the important categories of objects or concepts which exist in some field of discourse, showing the relations between them". Comprehensively, ontology is a categorization of all of the concepts in some field of knowledge, including the objects and all of the properties, relations, and functions needed to define the objects and specify their actions. A well known definition of ontology given by Grubber (1993) is that ontology is "a formal, explicit specification of a shared conceptualization".

Ontology has a great potential in providing homogeneity amongst heterogeneous resources by offering common access to information. It also allows semantic search which eventually realized specialize search engine in retrieving desired resources, improve precision and lessen the searching time. As the ontology for Computer Science and Library Science discipline is still in the phase of development and testing, we tried to make it rigorous so that it can be used across all databases in the same area. This research is highly motivated by the work done by Zhang et.al (2006) on providing semantic search in MEMS (Microelectromechanical System) journal collection. Zhang constructed the ontology by bootstrapping ontology learning for information retrieval using Formal Concept Analysis and the notion of information anchor. Their research showed that the search result improves significantly after the implementation of ontology on top of the MEMS search engine.

In the case of domain specific collections such as the Malaysian Journal of Computer Science (MJCS) and Malaysian Journal of Library Information Science (MJLIS), the querying of information by the user requires knowledge of the technical terms of the domain and the structure of specific databases. At this moment, the collections can be queried using keyword(s) which utilizes the string matching technique. The research aims at improving the access to this data which resides in disparate databases especially for early-staged researcher such as undergraduate students and postgraduate research students.

The approach chosen is twofold: one is to represent the information in an unstructured data such as text collection alongside the structured data in existing database and integrate the available data by linking collections of journals together with the intention that the researcher can be offered extra information with regards to the subject of interest.

The first approach will be solved by constructing an ontology that will interface the heterogeneous content. The aim is to provide homogenous, properly-organized and easy-to-retrieve structure of information in the Institutional Repository. The constructed ontology will indirectly solve the second task, where the taxonomy provided by the ontology would allow query expansion, thus providing the user with additional information obtained from different collection.

3. Methodology

The chosen methodology of ontology construction is inline with the three-step process of ontology formation (Tariq et al, 2003) starting from phase two until phase four (as depicted in Figure 3.0). Tariq et al. (2003) proposed an architecture for candidate ontology generation given a domain-specific text corpus. The MJCS and MJLIS are samples of domain specific corpus which need to be thoroughly examined for restricted terminology. For instance, the terms in the Artificial Intelligence (AI) domain such as *Knowledge Modeling* and *Expert System* were not found in Wordnet (http://wordnet.princeton.edu), which is the well-known semantic lexicon for English language. We further enriched the specialized corpus with extra resources from the established websites in the field of AI such as Association for the Advancement of AI (http://www.aaai.org) and input from experts.

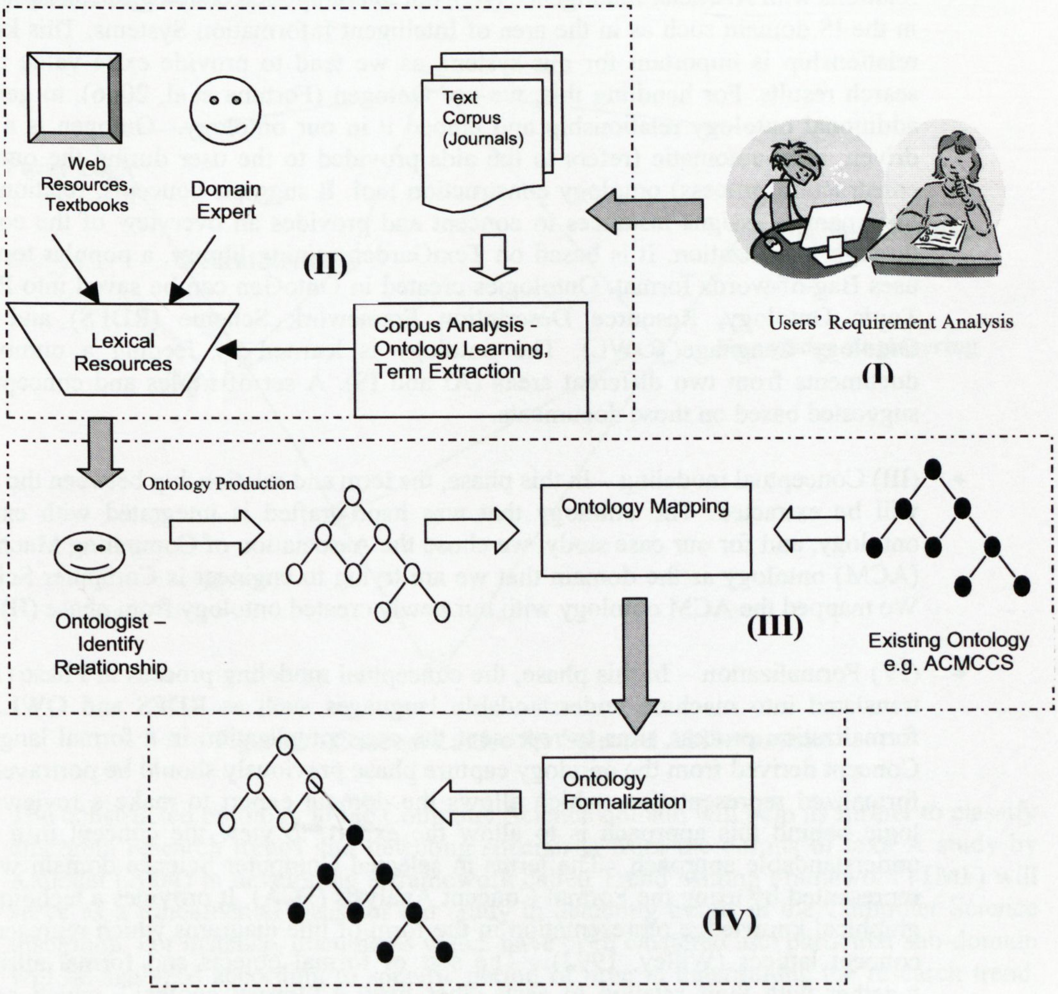

Figure 1: Ontology Construction Process

We extend the approach by Tariq et.al (2003) by adding another phase before the actual process of ontology construction begins.

- (I) Users' Requirement Analysis - Often, knowledge is elicited from the experts or specialists in a particular area. Our study emphasized that before an expert is consulted in order to engineer the knowledge, we must first elicit the need of the user; that is to know what kind of knowledge is required in order to satisfy the user's needs. Our survey conform that users need additional information while searching. Information such as document title, specific type of documents (such as journal, thesis and conference paper) and authors are marked as specific criteria to describe particular documents. In our opinion, the identification of users' needs will make a future built system to be more acceptable and more likely to be used.

- (II) Knowledge acquisition (KA)– In this phase, a corpus will be created and analyzed. KA is defined as the process of eliciting, modeling and validating knowledge for knowledge engineering purposes (Epistemics, 2007). Once users' needs have been captured, we proceed with gathering related data to form the knowledge-base from various sources such as from text books, established websites of academic-based organizations and consultation with experts in a particular area. The ontology is also learned from text to further extract statement triples by utilizing ontology learning tools to further enrich the ontology. For example, classes in Information Science (IS) such as Information System do not share or have any relations with Artificial Intelligence (AI), but in reality AI technologies can be applied in the IS domain such as in the area of Intelligent Information Systems. This kind of relationship is important for our system, as we tend to provide extra value in our search results. For handling this, we use Ontogen (Fortuna et.al, 2006), to generate additional ontology relationship and embed it in our ontology. Ontogen is a data-driven, semi-automatic (refers to the aids provided to the user during the ontology construction process) ontology construction tool. It suggests concepts, relations and their names, assigns instances to concept and provides an overview of the concept through visualization. It is based on TextGarden mining library, a popular tool that uses Bag-of-words format. Ontologies created in OntoGen can be saved into Proton Topic Ontology, Resource Description Framework Scheme (RDFS) and Web Ontology Language (OWL). The ontology is learned by feeding a number of documents from two different areas (AI and IS). A set of triples and concepts are suggested based on those documents.

- (III) Conceptual modeling – In this phase, the term and relationship between the terms will be extracted. The ontology that was hand-crafted is integrated with existing ontology, and for our case study, we chose the Association of Computing Machinery (ACM) ontology as the domain that we are trying to engineer is Computer Science. We mapped the ACM ontology with our newly created ontology from phase (II).

- (IV) Formalization – In this phase, the conceptual modeling process in Phase (III) is translated into machine understandable languages such as RDFS and OWL. The formalization process aims to represent the conceptualization in a formal language. Concept derived from the ontology capture phase previously should be portrayed in a formalized representation, which allows the domain expert to make a review. The logic behind this approach is to allow the expert, to view the concept in a more understandable approach. The terms in selected Computer Science domain will be represented by using the Formal Concept Analysis (FCA). It provides a technique of graphical knowledge representation in the form of line diagrams which represent the concept lattices (Willey, 1992). The sets of formal objects and formal attributes together with their relation to each other form a "formal context", which can be represented by a cross table. The elements on the left side are formal objects; the elements at the top are formal attributes; and the crosses represent the relation between them. Table 1 depicts the cross table which represent selected concepts from the AI domain.

Table 1: Cross Table Result for AI's Concepts

Object	**Fuzzy Logic**	**Neural Network**	**Knowledge Engineering**	**Machine Learning**
Doc_1	X	X		
Doc_12	X			
Doc_37	X			X
Doc_18	X		X	X

The cross table contains simplified information that were gathered from the generated concepts derived from ontology learning described in phase (II). Based on the formalized data captured in the cross table, we can generate the concept lattice which allows the investigation and interpretation of relationship between concepts, objects and attribute. A concept portrays the whole information in the cross table, for example in our case is the AI classification. The concept lattice is equivalent to the representation of cross table, but it is easier to detect dependencies and relationships between attribute in a line diagram as it provides a "subconcept-superconcept" view (Figure 2). The top nodes present the superconcept of the bottom nodes.

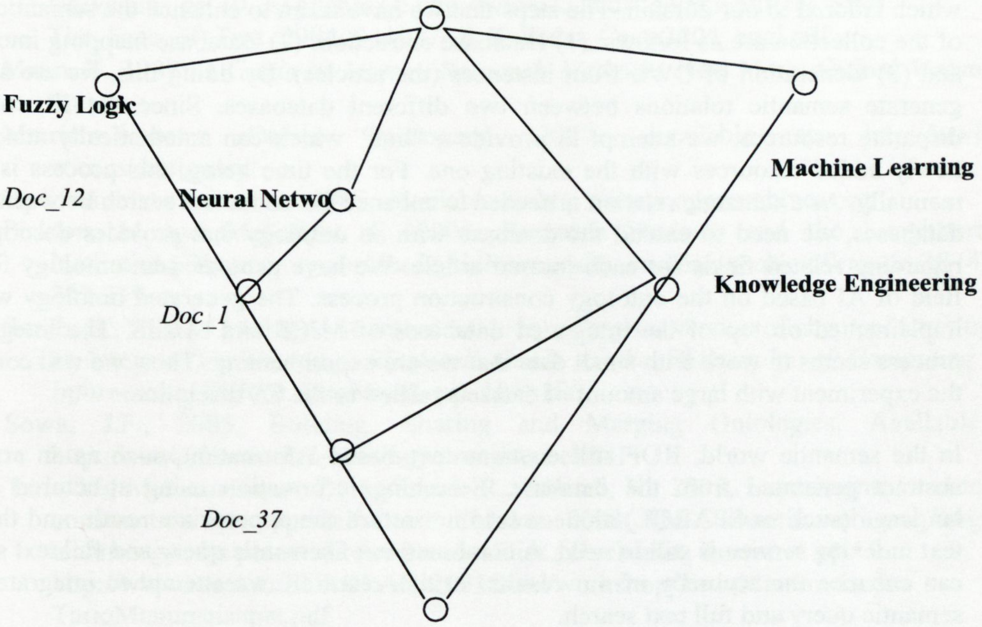

Figure 2 : Concept Lattice for Selected AI Documents

The constructed ontology in the Computer Science domain will help us further to classify emerging research trends and detecting specialists from the corpus of text. A study by Kalledat (2004) in developing a framework called Trend Mining Framework (TMF) will serve as a fundamental basis for our study in detecting trends in the Computer Science discipline. For instance, documents which have been clustered into particular sub-domain will be analyzed according to specific period of time in determining the research trend. The trend detected by utilizing TMF will give us a specific cluster of text which will eventually make the process of identifying the expert for the domain easier. In order to detect a specialist, the structured content of the digital library will tremendously help in identifying the expert in a particular area. For instance, counting the frequency of appearance of an expert's name with respect to certain corpus will yield fast results by just posting a simple SQL in structured database. We will extend the "frequency count method" by learning the nature of name appearance according to the position. For instance, first author will be assigned more weight as compared to second author and so on. A collection of academic articles will be studied to gather information on the maximum number of authors. This will give us a boundary in deducing heuristics for automatically determining an expert in a particular area. The problem that we might face from this approach is in determining an expert who is involved in a multidisciplinary research area.

4. Discussion and Future Work

We have elicited students' requirements towards establishing IR in the local university through an interview and survey with undergraduates and postgraduates students. From the analysis, the specific needs for ontology development i.e. providing common access to information and enabling semantic-based search are identified. Common access to information can be achieved by providing a homogeneous view (from the perspective of human) of databases and multiple applications. Based on the proposed solution for ontology construction process, we have managed to enhance the metadata of the database and provide more relationship thus providing a better description of the article inside the journal databases. We have managed to show the activities done in each of the phase which tailored to our domain. The steps that we have taken to enhance the semantic view of the collection are as follows: (1) Database extraction, (2) Database mapping into RDF and (3) Generation of OWL from instances (the articles). By doing this, we are able to generate semantic relations between two different databases. Since the IR contains disparate resources, we attempt to provide a "link" which can automatically attach the newly added resources with the existing one. For the time being, this process is done manually. As a semantic relation is needed to enhance the document search between these databases, we need to extend the database with an ontology that provides descriptions regarding related fields for each journal article. We have generated an ontology for the field of AI based on the ontology construction process. The generated ontology will be implemented on top of the integrated databases of MJCS and MJLIS. The integration process seems to work with small data that we are experimenting. Thus, we will continue the experiment with large amount of datasets related to the CS discipline.

In the semantic world, RDF still contains text–based information, such as an article's abstract generated from the database. Searching information using structured query language (such as SPARQL) alone would not return the appropriate result, and the full text indexing service is still in need. A combination of semantic query and full text search can enhance the accuracy of our results. In this research, we attempt to integrate both semantic query and full text search.

The complete framework of the ontology for IR should comply with OAI-PMH metadata harvesting standard. OAI-PMH is a low-level protocol written in XML for metadata harvesting and is widely use in IRs. It provides standard encoding in request and respond messages between the database and the harvester (Open Archives Initiatives, 2007). The architecture of OAI-PMH relies on Dublin Core, and it should be able to maintain any future migration of the databases. It is not the focus of this research to implement OAI-PMH, as there are a lot of existing tools that can be used for future migrations. Other than that, our future prototype will provide similar functionality like a harvester, making it compatible with current proposed metadata harvesting system architectures.

References

Berners-Lee, T., Fischetti, M., 1999. *Weaving the Web*. Harper San Francisco, Chapter 12. ISBN 9780062515872.

Berners-Lee, Tim., Hendler, J., and Lassila, O. 2001, The Semantic Web. *Scientific American,* 284:34-43

Epistemics Inc. 2007, Knowledge Acquisition, at URL http://www.epistemics.co.uk/Notes/63-0-0.htm

Fortuna, B., Grobelnik, M., Mladenic, D., 2006, Semi-automatic Data-driven Ontology Construction System, Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia

Griffiths, J.R., 1996, Development of a specification for a full text CD ROM user Interface. MPhil thesis : MMU

Guarino N., and Giaretta P, 1995. Ontologies and Knowledge Bases: Towards a Terminological Clarification, Toward Very Large Knowledge Bases, IOI Press. pp 25-32

Guha, R., McCool, R., Miller, E., 2003, Semantic Search, *Proceedings of the 12th International Conference on World Wide Web,* Budapest, Hungary, ACM Press, pp:700 – 709, ISBN:1-58113-680-3

Jansen, B.J et. al. 2000, The effect of query complexity on Web Searching Results, *Information Research 6* (1). Available at: http://www.shef.ac.uk/~is/publications/infres/paper87.html

Kalledat, T. 2004, Automatic Trend Detection and Visualization using the Trend Mining Framework (TMF) , *Canadian Symposium on Text Analysis Conference*, Macmaster University, pp 10. Available at: http://www.kalledat.de/Scientifical_Stuff/The_Face_of_Text_2004/ 041027_T_Kalledat_Casta2004_engl.pdf

Mann, T. 1987. *A Guide to Library Research Methods*. New York: Oxford University Press.

Merriam-Webster Online Dictionary. 2007. Available at: http://www.m-w.com/dictionary/

Open Archives Initiatives. 2007. Available at: http://www.openarchives.org/

Pickton, M and McKnight, C. 2006, Research Students and the Loughborough Institutional Repository, *Journal of Librarianship and Information Science,* 38 (4), pp 203-219

Smith, B., 2003, Ontology. An introduction by - Preprint version of chapter "Ontology", in Luciano Floridi (ed.), *Blackwell Guide to the Philosophy of Computing and Information,* Oxford: Blackwell,, pp. 155–166.

Sowa, J.F., 2005, Building, Sharing and Merging Ontologies. Available at: http://www.jfsowa.com/ontology/ontoshar.htm#s1

Tariq M., Manumaisupat, P., Al-Sayed, R., Ahmad, K., 2003, Experiments in ontology construction form specialist text, EuroLan 2003, The Semantic web and language technology : Its potentials and practicalities., Bucharest, Romania, pp 10 http://www.racai.ro/EUROLAN-2003/html/workshop/TariqManumaisupat/TariqManumaisupat.pdf

Willinsky J. 2006. *The access principle: the case for open access to research and scholarship.* Cambridge, Mass.: MIT Press,

Wille, R. 1992 Concept lattices and conceptual knowledge systems. In: *Computers and Mathematics with Applications*, 23 , 493-515.

Zipf, G. 1949, *Human Behaviour and the Principle of Least Effort.* Reading MA: Addison-Wesley

Zhang, G-Q., Troy, A.D., Bourgoin, K. 2006. Bootstrapping Ontology Learning for Information Retrieval Using Formal Concept Analysis and Information Anchors, *14th International Conference on Conceptual Structures*, Aalborg, Denmark, July 2006, pp 14. Available at: http://newton.cwru.edu/papers/Zhang_ICCS06_ontology_final.pdf