Determining and Satisfying Search Users Real Needs via Socially Constructed Search Concept Classification

Dengya Zhu and Heinz Dreher

Curtin University of Technology, GPO Box 1987, 6845 Perth, Western Australia,
e-mail: dengya.zhu@postgrad.curtin.edu.au, h.dreher@curtin.edu.au

Abstract—**The focus of the research is to disambiguate search query by categorizing search results returned by search engines and interacting with the user to achieve query and results refinement. A novel special search-browser has been developed which combines search engine results, the Open Directory Project (ODP) based lightweight ontology as navigator and classifier, and search results categorizing. Categories are formed based on the ODP as a predefined ontology and Lucene is to be employed to calculate the similarity between retrieved items of the search engine and concepts in the ODP. With the interaction of users, the search-browser improves the quality of search results by excluding the irrelevant documents and ontologically categorizing results for user inspection.**

Index Terms—**information retrieval, text classification, search engine, ontological filtering, Open Directory Project.**

I. Introduction

The advent and the explosion of the Web are challenging the Information Retrieval (IR) research community . The first challenge is the huge and dynamic nature of the Web and its store of information. Secondly, more and more people begin to use the Internet as the main approach to obtain needed information. However, as pointed out by [3], most users are not good at expressing their information seeking requests in search term format. Thirdly, people's information needs are diverse, dynamic and fuzzy. They sometimes do not know what exactly they want to search for, especially when they are casual and new search users. They may also change their search topic frequently, or concentrate only on a given topic for a quite long period.

Information explosion and the demands for high quality information by users enlarge the gap between information retrieval services provided by search engines and information consumption requested by the ever increasing population of miscellaneous Web users; the inherent problems of polysemy and synonymy of information retrieval makes the situation even more challenging. [1][14][15][27][39] suggest that more than fifty percent of search results returned by search engines are irrelevant - users' individual information needs are neglected and thus make the search activity and research results of less value.

Much effort has been devoted to improve the relevance of search results to satisfy users' information needs, such as using different IR models and their variations, relevance feedback, information clustering/re-organization, word sense disambiguation [22][31][35], personalization [7], semantic web [11], ontology[6][16][37] based IR, question-answer system (http://www.answers.com/) [12], or more general, natural language processing systems [8], interactive IR [4], and the like.

Among the techniques mentioned above, the promising one is information clustering/classification. Search engines usually return a list of thousands or even millions of retrieved items that are ranked according to the relevance to the search terms by using an IR model. The plain listing of such a large number of search items and the lack of organization of the search results frustrates information seekers. Clusty (http://www.clusty.com) uses information clustering techniques to re-organize the retrieved items according to the subjects/topics formed by the different groups of the search results. Answers.com (http://www.answers.com) also clusters search results in some circumstances while trying to give an accurate definition of the search terms/concepts.

The focus of the research is to ontologically disambiguate search query by categorizing search results returned by search engines [39]. A novel special search-browser has been developed which combines search engine results, the Open Directory Project (ODP) based lightweight ontology as a navigator and classifier, and search results categorization. Categories are formed based on the ODP as a predefined lightweight ontology [6] and Lucene (http://lucene.apache.org) calculates the similarity between items retrieved by the search engine and concepts in the ODP. With the interaction of users, the search-browser is expected to produce more relevant search results by excluding the irrelevant, and thereby improving the quality of information for the user [39].

The rest of the paper is organized as follows. Section 2 is a discussion on the challenges for the current search engines. In section 3, a lightweight ontology driven search results classification and query disambiguation approach is proposed. In section 4, system structure and implementation is presented. Section 5 discusses some alternative technologies which may also be utilized in the research, and finally in section 6, our conclusion.

II. Challenges for Search Engines

The three challenges, in turn, open up numerous issues to be resolved in order to achieve our goal of increasing user satisfaction with regard to their information seeking activities. The main challenges for search engines are discussed below [39].

A. Information Overload of Search Engines

Google claims it has indexed more than 8 billion web

pages (http://www.google.com), and Yahoo announced recently it indexed 20 billion web pages [5]. While these numbers may indicate that the search engines have the potential ability to return more search results, surveys [8] show that most Web users pay much attention to retrieve relevant information effectively and only a few are willing to review more than ten relevant search results. However, search engines tend to return thousands even millions of search results for the short search term query preferred by most users. Searching for the information about the animal "jaguar" using that as the search term, Google (from Google <home> in September, 2006) returns some 82 million items, and in its first returned page, only the seventh listed item concerns cats or animals. The huge quantity of search results is no doubt an information overload, because of which valuable information may be overlooked - information overlook incurs opportunity costs.

B. Mismatch of Search Results of Search Engines

Due to the polysemy problem of natural language and the user search habit of using short search terms, it is very difficult or impossible for search engines to return relevant search results for an individual information seeker without knowing the user's true needs in advance. One searcher may want to retrieve some information about the animal "jaguar", another may want to search for a "jaguar" motor vehicle. In the absence of interaction with the individual user, it is unreasonable for a search engine to return only information concerned with "jaguar" car, or only return search results related to the animal "jaguar". For example, search engines do their best to increase recall by returning as many literal related items as possible; it is therefore not surprising that Google returns some 82 million search items for a searcher who wants to find some information about the animal "jaguar" by using the search term "jaguar". The information seeker will be frustrated when facing such a huge number of items where more than half of those presented on the first page of results may be irrelevant [14][33][39]. This problem can be described as low precision of search results.

C. Missing Relevant Document

Despite millions of returned research results, the low recall issue is still facing search engines in some cases. This is mainly because of the inherent problem of synonymy of natural language. For example, when searching for "artificial intelligence research", search engines will not search for the synonyms "AI" and "machine intelligence" [17]. Another reason for missing relevant documents is search results generally do not include subfields of a general field, for example, when searching for papers on "machine learning", search engines will not return results about "genetic programming", a subfield of machine learning and artificial intelligence.

D. Searcher's Mental Model Mismatch

Some search results are mis-categorized by search engines that cluster the search results. Clusty, ranked number 4 of top 20 search engines by SquirrelNet [34], is a Clustering Engine which organizes search results into folders grouping similar items together. The search result of "jaguar" from Clusty.com is illustrated in Figure 1 (retrieved on September 24, 2006).

Search results are clustered and organized in a hierarchical structure and presented in groups of subject/topic. However, from the point of view of the knowledge structure of human beings, the arrangement of the search results can be confusing. For instance, car, parts and model are all arranged in the first level of the hierarchy, whereas the well-known mental construct for people is that car has parts and different models. The same is true for the arrangement of panthera onca and animal, for panthera onca is a kind of animal. Clustering Web search results and entitling the groups with the extracted topic/subjects usually cannot reflect the hierarchy of knowledge and will thus mismatch the mental model of human beings – a generally accepted basis for knowledge classification.

E. Poorly Organized Search Results

Most search engines arrange search results according to ranking algorithms that rank documents in a higher priority

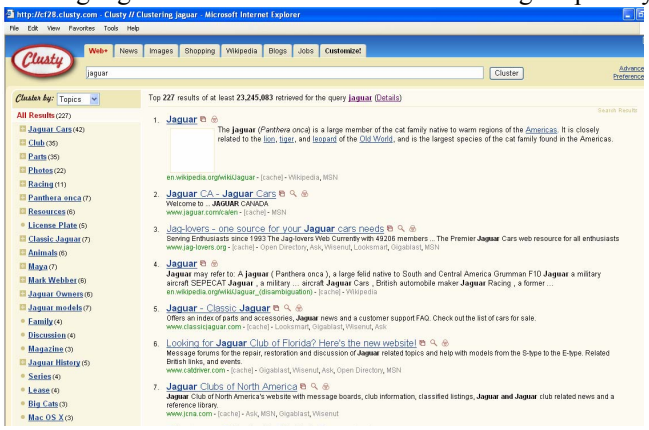

Fig.1 Search results of "jaguar" returned from http://www.clusty.com

according to the document's literal similarity to the given query [2][21]. Ranked documents listed are considered relevant to a user's query in descending order, that is, the first several documents are more relevant to user's query than the rest of the search results. However, because of the problems mentioned above and because search engines frequently return thousands or millions of search results in a list, a user may need to check hundreds of items to retrieve useful information among search results! Finding a relevant document among the returned Web search results is like finding "the needle in the haystack" [3].

Plain lists of search results also deliver no information about knowledge structure related to the search terms; each retrieved item is isolated from the others and is independent. A plain list format of search results is appropriate when the returned items are less than 50, and relevant documents reviewed per session are around ten [8]. Therefore, organizing and classifying the huge amount of search results will facilitate Web information seekers to locate relevant information.

III. Lightweight Ontology Driven Search Results Classification and Query Disambiguation

An "Open Directory Project" (ODP) [26] based interactive ontological search results filtering approach is proposed in this paper. As discussed in section 2.2, polysemy is the main factor that results in search engines returning irrelevant search items, and thus leads to the low precision of the search results. Clustering techniques used by Clusty.com and other search engines usually produce grouped search items that mismatch human mental models because the clustering algorithms used are not based on the human hierarchy of knowledge. By using ODP lightweight ontology to classify search results; then interacting with users to disambiguate search terms; and subsequently filtering the search results, quality search results with higher precision are expected.

A. The Open Directory Project as a Knowledge Hierarchy

The Open Directory Project is the largest, most comprehensive human-edited directory of the Web. It classifies the whole content of the Web into 15 categories with an additional world category that contains the non-English language versions. Figure 2 is the home page of the ODP (retrieved on December 24, 2006).

Fig.2 Home page of the Open Directory Project

The ODP is hierarchically constructed with further information to describe the categories; and for each of the categories, there is a set of submitted web pages which are related to the category. The ODP provides a description for each category which gives specific information about the content and/or subject matter of the category accompanied with some editorial information. For each annotated web page, the submitter of the web page is also asked to provide a description of the submitted page that gives a brief description about the content and subject matter of the submitted site. Figure 3 depicts the "path" of a set of submitted pages in the ODP [39].

B. Extracting Semantic Characteristics of ODP Categories

[39] proposes the semantic characteristics of each category can be represented by the title of the category, the description of the category, the submitted web pages and their descriptions as describe above. All of the data together can form a "category-document". The category-document describes what the category is about and thus makes its semantic information amenable for subsequent algorithmic processing.

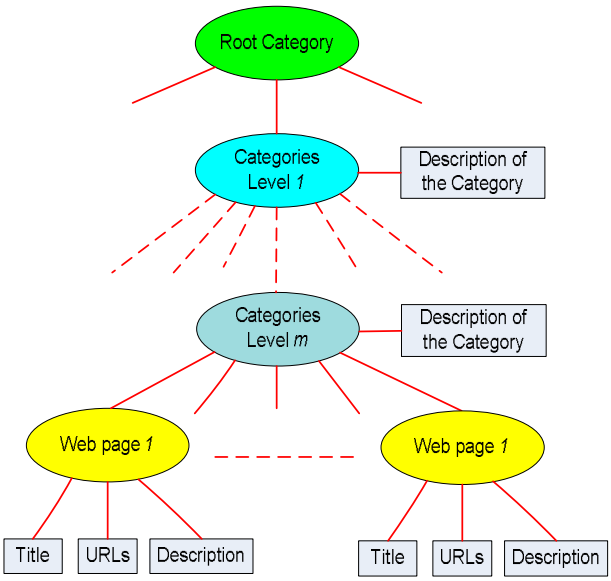

Fig.3 Structure of the Open Directory Project

C. Disambiguating Query Terms using ODP Category Semantics

"In general terms, word sense disambiguation (WSD) involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word" [18]. Query disambiguation based on the ODP means that the returned search items, which are ranked by search engines depending on the term-weighting strategy [3][29], will be re-organized according to their similarities to the semantic characteristics of the different categories of the ODP, and consequently making the different meanings of the query distinguishable by topics/subjects of the related categories of the ODP [39].

Using the semantic characteristics of the ODP to disambiguate a query is different from the approaches WSD usually employed where dictionaries, a group of features, categories (pure categories without meta-data and relative web pages), associated words (e.g., synonyms, as in a thesaurus), or WordNet (http://wordnet.princeton.edu/) are used to determine the different senses of word. Utilizing meta-data of the ODP can semantically disambiguate a query to overcome the shortcoming facing conventional approaches, described by [18] as "the rather narrow view of sense that comes hand-in-hand with the attempt to use sense distinctions in everyday dictionaries, which cannot, and are not intended to, represent meaning in context." As the OPD is a socially constructed dynamic knowledge representation it informs the disambiguation issue.

D. Improving Precision through ODP Categorical Filtering

One way to improve precision is to reduce the retrieved document set when keeping the retrieved relevant document set unchanged; this is because precision = (relevant documents retrieved) / (document retrieved). By categorizing search results according to their semantic characteristics

based on the ODP as a lightweight ontology, a user may click-select a category to retrieve only search items listed under that category. The answer set is confined within the click-selected categories, vastly reducing the returned results and thus increasing precision. However, an evaluation is needed, since during the process some relevant items may also be excluded and some irrelevant items may not be filtered out.

Suppose the original answer set is A, Ra is the relevant documents in A, and after ontological filtering the answer is A', L is the set of relevant documents excluded (i.e. not returned by the search). The final precision is:

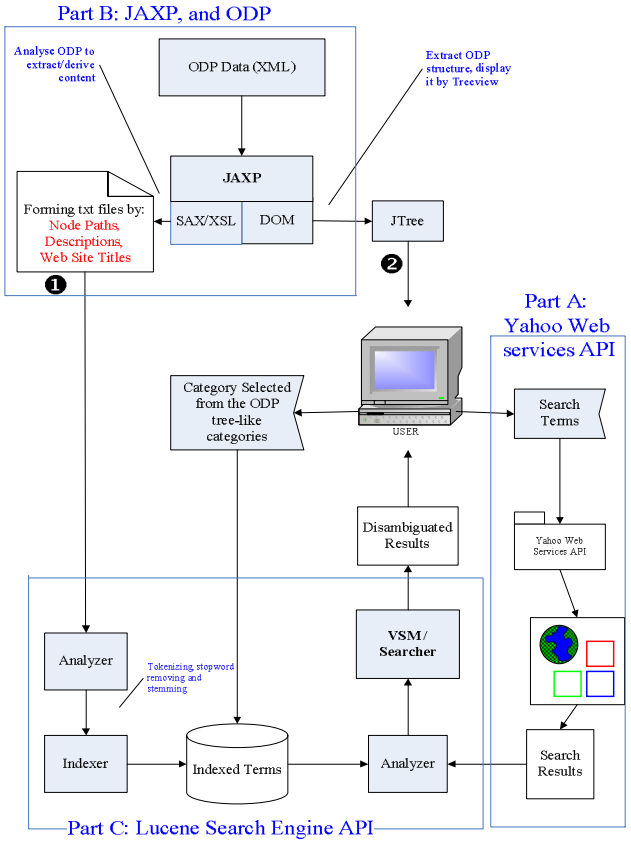

Fig.4 Structure of the specific search browser [39]

$$(|Ra| - |L|) / |A'|$$

Because A' is expected less than A, and L is expected to be relatively small compared to Ra, comparing with Ra / A, a higher precision is achieved. For example, assume a search engine returns 100 search results for a given search term, among the 100 returned search results, a user judges 20 of them are relevant to the information need. In this case, the precision is 20/100 = 20%. This means that 80% of the returned search results are not relevant to the user's information need. Further assume that our search browser categorizes the returned 100 search results into 3 categories, if one of the categories contains 20 returned search results and 15 out of 20 of the results are relevant, that is, |A'| = 20 and |L| = 20 – 15= 5 the precision of the categorized search results is thus (20 – 5) / 30 = 75%. Despite of the 25% recall loss, the precision improvement is 75 – 20 = 55%.

IV. System Structure and Implementation

The specific search-browser which implements the proposed structure is composed of four parts as shown in Figure 4 [39].

Part A (right side of Fig 4) is a web search engine interface which utilizes *Yahoo! Search Web Service API* to implement the search-term based web searching. It accepts user's search-term as input and returns a list of retrieved items.

The goal of Part B (top left of Fig 4) is to produce a set of text files that can then be used by the Lucene search engine, and to extract the hierarchical structure of the ODP to be displayed by a *JTree*. Simple API for XML (SAX) and the eXtensible Stylesheet Language Transformations (XSLT) are employed to analyze the content of the ODP data which can be downloaded from [26]. The content of the "category-document" is stored in text file format. The title of each category is used as the title of each text file, and the content of each category-document is used to form the content of the text file. Document Object Model (DOM) is utilized to map the hierarchical structure of the ODP by a *NodeAdapter* to a *DataModel* which can then be used by *JTree*. An alternative to display the hierarchical structure data of the ODP is to create the *JTree* at the system coding phase by adding the appropriate category nodes.

Part C (bottom left) in Figure 4 implements term indexing and search results classifying. Lucene's *Anlayzer* and *Indexer* classes are used to analyze and index the text files returned from Part B. The classifying process is achieved by comparing the similarities between each retrieved item and the categories of the ODP, as elaborated in the next paragraph. Each returned result has one "most similar" document in the formed "category-document" repository. The category which "the most similar" document belongs to will be marked (Fig. 5) in the corresponding position in the *JTree* to inform the user how the returned search results are classified according to the hierarchical structure of the ODP.

To calculate the similarities between retrieved items and the categories in the ODP, each *category-document* in the document repository *D* is taken as a high dimensional vector which can be denoted as vector $\vec{d}_j$. The search items returned from the *Yahoo! Search Web Services API* are also taken as query vectors, therefore, the similarity between query vector $\vec{v}_q$ (returned item from the *Yahoo! Search Web Services API*) and $\vec{d}_j$ (j = 1, 2, … N) can be measured as:

$$\begin{aligned} \text{sim}(\vec{v}_q, \vec{d}_j) &= \vec{v}_q \bullet \vec{d}_j \; / \; |\vec{v}_q| \times |\vec{d}_j| \\ &= \sum_{i=1\ldots N} W_{i,j} \times W_{i,q} \; / \; \sqrt{\sum_{i=1\ldots N} W_{i,j}^2 \times \sum_{i=1\ldots N} W_{i,q}^2} \end{aligned}$$

For each query vector $\vec{v}_q$, the similarity between this query vector and the *category-document* vector $\vec{d}_j$ (j = 1, 2,

…, N) will be ranked decreasingly by the calculated similarity. The search result item represented by $\overrightarrow{v_q}$ can thus be classified to the category represented by the *category-document* $\overrightarrow{d_j}$ if $\overrightarrow{d_j}$ is the top ranked document in the similarity list of vector $\overrightarrow{v_q}$ and $\overrightarrow{d_j}$ (j = 1, 2, …, N). An alternative and appropriate approach to classify the returned search result is to employ the Majority Voting strategy implemented by [39].

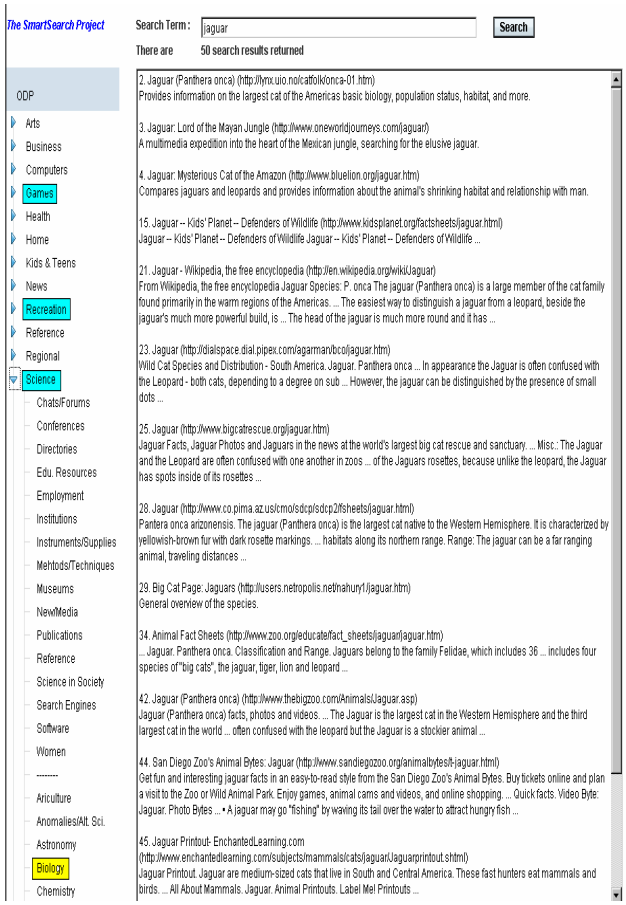

Fig.5 A sample interface of the specific search browser

The last part is the user interface, as seen in Fig 5. The interaction between user and the search-browser is implemented by this part.

The structure of Fig 5 is open, flexible and expandable. Part A can be any meta search engine, or intranet search engine, or any database search engines; Part B can be any light weight ontology, or a domain-oriented knowledge structure; Part C includes a full text search engine and a classifier which may be the VSM classifier, the probabilistic classifier [9], and others [32]; the knowledge structure can be represented by a tree structure, or by other visualization component.

V. Discussion and Future Work

A. Trial the Proposed Approach

There are 15 categories in the ODP, and for each of these categories, a relative search term will be selected and evaluated by calculating the precision of the original search results and the precision after the proposed ontological search term disambiguation based on the ODP. The standard 11 points recall-precision curve [3] will also presented. Further evaluation may be conducted with resources published by TREC (Text REtrieval Conference at http://trec.nist.gov/) or some public web search engines.

B. Probabilistic Model, Machine Learning, and Other Types of Classifier

Lucene uses a modified Vector Space Model [2][3][25][29][30] when calculating the similarity between search term and the document repository. Probabilistic Model [9][19][20][28] can also be used as a classifier based on the ODP to disambiguate search terms. [24] uses a machine learning approach to classify web pages based on *Yahoo! Web category*, for the 14 Yahoo! categories, a separate classifier is built. [23] utilizes n-gram algorithm to automatically classify web pages and the experimental results are encouraging. [13] uses hierarchical knowledge structure (such as Yahoo! Web category) to achieve a better categorization of Web search results.

The data of the ODP is dynamic, increasing, and huge. To reduce the high dimensionality of the vector space created by this ODP data, two techniques, the Latent Semantic Indexing algorithm [10][32], and a recent and computationally more efficient technology named Normalised Word Vector [38], are to be employed in the next stage of the research. Machine learning approaches and probabilistic models will also be evaluated as a measure to ontologically disambiguate search terms based on the ODP.

VI. Conclusion

Search engines are facing challenges of irrelevant search results, poor search results organization, and lack a commonly shared knowledge structure for search results classification. The paper proposed a novel approach to ontologically disambiguate search terms based on the knowledge structure of the ODP. The meta-data in the ODP are used to form a "category-document" collection, the search items returned from search engines are also treated as documents and a classifier is developed by using Lucene to align the returned items to the most "similar" category in the ODP. By interaction with users, a higher precision of search results is expected. In this way we hope to accommodate the Web's diverse information seekers, with their diversity of information needs, even while the web changes and expands.

VII. References

[1] R.B. Almeida and V.A.F. Almeida, "A Community-Aware Search Engine", in: *Proceedings of the 13th International Conference on World Wide Web* (New York, NY), 2004, 413-421.

[2] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan, "Searching the Web", *ACM Transactions on Internet Technology,* vol. 1, no. 1, 2001, 2-43.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press, New York, 1999.

[4] N. Belkin, C. L. Borgman, S. Dumais and M. Hancock-Beaulieu, "Evaluating Interactive Retrieval Systems", in: *Proceedings of the*

17th annual international ACM SIGIR conference on Research and Development in Information Retrieval (Dublin, Ireland), Springer-Verlag, New York, 1994, 361.
[5] Blogcritics.org http://blogcritics.org/archives/2005/08/09/030442.php
[6] J. Bruijn, "Using Ontologies: Enabling Knowledge Sharing and Reuse on the Semantic Web", *DERI Technical Report DERI-2003-10-29*, October, 2003. Retrieved May 8, 2005, from http://whitepapers.zdnet.co.uk/0,39025945,60120712p-39000589q,00.htm
[7] P. Chirita, W. Nejdl, R. Paiu and C. Kohlschutter, "Using ODP Metadata to Personalize Search", in: *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '05)*, (Salvador, Brazil, August). ACM Press, New York, 2005, 178-185.
[8] G. G. Chowdhury, *Introduction to Modern Information Retrieval.* Facet Publishing, London, 2004.
[9] F. Crestani, M. Lalmas, C.J. Rijsbergen, and I. Campbell, "Is This Document Relevant? … Probably: A Survey of Probabilistic Models in Information Retrieval", *ACM Computing Surveys,* vol. 30, no. 4, 1998, 528-552.
[10] S. Deerwester, S.T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic indexing" *Journal of the American Society for Information Science,* vol. 41, no. 6, 1990, 391-407.
[11] Y. Ding, C.J. Rijsbergen, I. Ounis, and J. Jose, "Report on ACM SIGIR Workshop on 'Semantic Web'", *ACM SIGIR Forum (SWIR 2003),* vol. 37, no. 2, 2003, 45-49.
[12] H. Dreher and B. Williams, "Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering", in *Proceedings of the 7th International Conference on Flexible Query Answering Systems H. Legind Larsen et al. (Eds.) (FQAS 2006)*, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 282-294.
[13] I. Frommholz, "Categorizing Web Documents in Hierarchical Catalogues", in: *Proceedings of the 23rd European Colloquium on Information Retrieval Research,* 2001.
[14] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing", *Web Intelligence and Agent System,* vol. 1, no. 3-4, 2003, 219-234.
[15] E.J. Glover, S. Lawrence, M.D. Gordon, W.P. Birmingham, and C.L. Giles, "Improving Web Search with user preference: Web Search—Your Way", *Communication of the ACM,* vol. 44, no. 12, 2001, 97-102.
[16] T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing" *International Journal of Human-Computer Studies*. vol. 43, no. 5-6, 1993, 907-928.
[17] N.Guarino, "Ontology-Driven Information Retrieval", in: *Invited talks in International Conference on Digital Libraries,* (New Dehli, India). Retrieved Nov 6, 2005, from http://www.teriin.org/events/icdl/presentation/day3/ng.ppt
[18] N. Ide and J. Veronis, "Word Sense Disambiguation: The State of Art", *Computational Linguistics,* vol. 24, no. 1, 1998, 1-40.
[19] K.S. Jones, S. Walker, and S.E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments", *Information Processing and Management,* vol. 36, no. 6, 2000, 779-840.
[20] K.S. Jones, "Document Retrieval Shallow Data Deep Theories Historical Reflections Potential Directions", *Lecture Notes in Computer Science*, vol. 2633. Springer-Verlag, Berlin Heidelberg, 2003, 1-11.
[21] M. Kobayashi and K. Takeda, "Information Retrieval on the Web", *ACM Computing Surveys,* vol. 32, no. 2, 2000, 144-173.
[22] R. Krovez and W.B. Croft, "Lexical ambiguity and information retrieval", *ACM Transactions on Information Systems,* vol. 10, no. 2, 1992, 115-141.
[23] Y. Labrou and T. Finin, "Yahoo! As an Ontology: Using Yahoo! Categories to Describe Documents", In: *Proceedings of the eighth international conference on Information and knowledge management* (Kansas, Missouri), ACM Press, New York, 1999, 180-187.
[24] M. Mladenic, "Turning Yahoo into an Automatic Web-Page Classifier", in: *Proceedings of the 13th European Conference on Artificial Intelligence Yong Research Paper* (Brighton, U.K.), John Wiley & Sons, Ltd., 1998, 473-474.
[25] MOLE – Text Analysis Group: Vector Space Model, Retrieved. Nov 4, 2005, from http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/index.html
[26] Open Directory Project, http://ww.dmoz.com
[27] J.. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar and T. Breuel, "Personalized Search: A contextual computing approach may prove a breakthrough in personalized search efficiency", *Communications of the ACM,* vol. 45, no. 9, 2002, 50-55.
[28] C.J.V. Rijsbergen, *Information Retrieval*, 2nd edn, Butterworth-Heinemann Newton, 1979.
[29] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval", *Information Processing & Management,* vol. 24, no. 5, 1988, 513-523.
[30] G. Salton, J. Allan and A. Singhal, "Automatic Text Decomposition and Structuring" *Information Processing & Management*, vol. 32, no. 2, 1996, 127-138.
[31] M. Sanderson, "Word sense disambiguation and information retrieval", in: *Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (Dublin, Ireland), Springer-Verlag New York, 1994, 142-151.
[32] F. Sebastiani, "A Tutorial on Automated Text Categorisation", in: *Proceedings of the 1st Argentinian Symposium on Artificial Intelligence* (Buenos Aires, Argentina), 1999, 7-35.
[33] U. Shah, T. Finin, A. Joshi, R.S. Cost and J. Mayfield, "Information Retrieval on the Semantic Web", In: *Proceedings of the eleventh international conference on Information and knowledge management* (McLean, Virginia), 2002, 461-468.
[34] SquirrelNet, http://www.squirrelnet.com/BestSearchEngines/top20.asp
[35] C. Stokoe, M.P. Oakes and J. Tait, "Word Sense Disambiguation in Information Retrieval Revisited", in *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, (Toronto, Canada), ACM Press New York, 2003, 159-166.
[36] V.S. Subrahmanian, *Principles of Multimedia Database Systems.* Morgan Kaufmann Publishers Inc. San Francisco, 1997.
[37] M. Uschold and M. Gruninger, "ONTOLOGIES: Principles, Methods and Applications", *Knowledge Engineering Review*, vol. 11, no. 2 1996, 93-136.
[38] R. Williams, "The Power of Normalised Word Vectors for Automatically Grading Essays", *Issues in Informing Science and Information Technology*, vol. 3, 721-729. Retrieved April 14, 2006, from http://proceedings.informingscience.org/InSITE2006/IISITWill155.pdf
[39] D. Zhu, "Improving the Relevance of Search Results via Search-term Disambiguation and Ontological Filtering", Master Thesis, Curtin University of Technology. To be submitted in January, 2007.