A fuzzy ontology for medical document retrieval

David Parry

School of Computer and Information Sciences
Auckland University of Technology
Private Bag 92006 Auckland 1020 New Zealand

Dave.parry@aut.ac.nz

Abstract

Ontologies represent a method of formally expressing a shared understanding of information, and have been seen by many authors as a prerequisite for the "Semantic web". A mapping between query terms and members of an ontology is usually a key part of any ontology enhanced searching tool. However the relative importance of a particular mapping to an overloaded term may be different for different users, and this information is vital for accurate satisfaction of a query.

One way of overcoming this problem is the postulation of a "fuzzy ontology". By adding a value for degree of membership to each term that is "overloaded", for each user or group of users then the recovered documents from ontology mediated search can reflect the likely information need. The author will discuss means of ontology fuzzification, by both analysis of a corpus of documents and the use of a relevance feedback mechanism and some possible extensions to this scheme.

Keywords: Ontology, Ontology Combination, Fuzzy Logic, Information retrieval.

1 Introduction

*"When **I** use a word,' Humpty Dumpty said, in a rather scornful tone,' it means just what I choose it to mean, neither more nor less"(Carroll 1872).*

In Computer Science terms an ontology is used to "express formally a shared understanding of information" (Noy, Sintek et al. 2001). There has been a great deal of interest recently in the construction of ontologies for representing medical knowledge, in deed the construction and use of ontologies has been described as the main task of medical informatics (Musen 2001). A number of Ontologies - such as the Unified Medical language system (UMLS) and its component parts such as the Medical Subject Heading (MeSH) and the Semantic Network already exist in the medical domain. The use, reuse and sharing of information between ontologies, is especially important in the medical field with the growth of evidence-based medicine(Moody and G. 1999), and the consequent requirement for appropriate information to be available to clinicians and patients(Grutter, Eikemeier et al. 2001).

Currently ontologies are seen as one of the key technologies involved in the "Semantic web"(Berners-Lee, Hendler et al. 2001), and representations in dedicated formats such as Protégé and in particular implementations of XML documents have been constructed. Communication and merging of ontologies remains problematic however, although there have been attempts to solve this problem – for example the SMART system (Noy and Musen 1999), which is related to the PROTÉGÉ ontology development and validation tool(Musen, Gennari et al. 1995). In particular there is a pressing need to be able to use multiple ontologies in order to relate knowledge stored in references sources with data collected from clinical records for example. However the constructor of an ontology is faced with an essential paradox – by increasing the suitability of an ontology for a particular part of a domain, the coverage of the ontology decreases and its use as a communication tool decreases as the potential audience becomes more specialised. At the limit, an ontology that perfectly expresses one persons understanding of the world is useless for anyone else with a different view of the world. Communication between ontologies is necessary to avoid this type of solipsism .

Fuzzy set theory has been extensively used in the context of information retrieval (Bordogna and Pasi 2000). A number of different schemes have been devised to implement fuzzy logic in IR. This work has covered such concepts as fuzzy construction of queries, the retrieval of fuzzy sets of documents and fuzzy relevance measures. However this has not been combined with the use of an ontology although in (Widyantoro 2001) the term fuzzy ontology is introduced in terms of the use of the fuzzy combination of query terms.

As previously stated, the UMLS MeSH ontology is a particularly useful one for the medical domain. However, of the 21836 terms within it, 10072 appear in more than one place. Thus the "overloading" of terms in a mature ontology can be seen to be significant. In addition, as ontologies are extended – for example outside their home domain this problem is likely to get worse.

In their comprehensive review of ontology roles and structures, (Lassila and McGuinness 2001), describe a spectrum of structures from a catalogue, to a full ontology with extremely complicated relations between the terms. The UMLS represents a number of points on this spectrum, for a fairly simple hierarchy represented by the MeSH tree, to the rich relationships embodied in the Metathesaurus. One of the drawbacks of the Metathesaurus is that because it is effectively a

compilation of other ontologies using terms drawn form other systems, the exact relationships within the parent systems cannot always be replicated within it.

The aim of this work is to suggest that a fruitful approach to the reuse, and generalisability of ontologies may be made by the introduction of the concept of a "Fuzzy ontology". In this paper the ontology described is a modification of the MeSH hierarchy, which is much simpler than many of the other ontologies described by (Noy and McGuinness 2001). However some future extensions to this work are raised in section 5. - for example Section 2 outlines the nature of the fuzzy ontology; Section 3 gives some indications of ways of assigning membership values within ontology. Section 4 describes the current implementation of a system to support FuzzOnt for information retrieval and Section 5 describes current and future experimental work.

This work has been performed in the context of the development of an intelligent searching system for finding useful medical information(Parry 2001).

One of the issues that arise in ontology construction is why they are needed. Often Semantic Web research focuses on the advantages of having ontologies that artificial agents can use to enhance searching, and reconciling differences between the interpretation of the ontologies used by different documents (Stephens and Huhns 2001). However, this work is focused on attempting to represent the ontology of the users, and particular concentrating on what a user believes his or her query means. A collective ontology is also proposed, which is intended to allow users of a searching system to act as human agents in a collective intelligence, improving the performance of the system for all users by means of a special form of fuzzy ontology updating – described in section 2.

2 Theory of the fuzzy ontology

2.1 Fuzzy Logic

Zadeh originally introduced fuzzy logic in 1965 (Zadeh 1965), in the context of set theory. I use the concept of "membership" as an attribute of an item within an ontology. By use of the membership value, a fuzzy logic can be used, modifying the standard Boolean logic as used in classical information retrieval. To replace the "AND" term the fuzzy MIN term is used and to replace the "Or" term the fuzzy MAX term is used. Briefly, the MAX relation involves assigning the highest membership value found in the antecedents – for example if a document has both " Head" and "Nose" within it and the membership value for the "Anatomy" part of the ontology for these terms is 0.7 and 0.9 respectively, then a query asking for "Head AND Nose" would assign the lowest common value, which is 0.7. However a query asking for "Head OR Nose" would assign a value of 0.9 to the document.

The fuzzy ontology (FuzzONT) is based on modification of an existing crisp ontology. Currently there are ontologies with an extremely rich set of relations between members, for example component parts of the UMLS have over 80 types of relations between ontology members, ranging from the simple "is a" to such specialised relations as "uniquely_mapped_to" and "developmental_form_of". By preserving these relations, an extremely rich set of relations can and form the framework of the ontology when beginning fuzzification. The modification is entirely incremental, conversion to a fuzzy ontology adds membership values to the currently existing relations, and may also add new entries, in the ontology. The ontology membership is normalised in respect to each of the terms in the ontology that is the sum of the membership value of each term in the ontology is equal to 1. This is because it is primarily concerned with mapping from queries to the ontology. This is justified on the basis that that for each term in a query, only one of the meanings will be required, and that these meanings are exclusive.

In the vocabulary of Noy, this is a "merging" process, rather than an alignment because the new ontology contains both of the old ones, but is itself only one. The advantage of this process however is that no information is lost.. The fuzzification process is shown diagrammatically in figure 1.

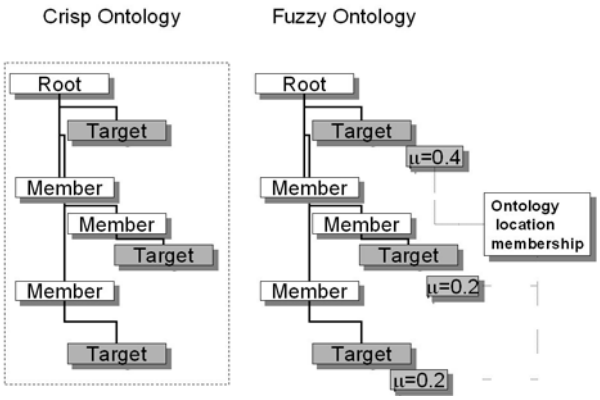

Figure 1: Membership values in a fuzzy ontology

The membership value can be assigned in one of two ways – via a user preference assigned using a membership function, as described in section 3.1, or automatically as described in section 3.2.

The main objection to this scheme is that the ontology becomes more complex. However this complexity already exists in the case of terms that are located in multiple positions in the ontology. Take for example the term "Pain" which occurs in the Mesh ontology 5 times in different trees at different levels – see table 1

Because the term is located in a number of different places, query expansion for this term is difficult, because there are wide numbers of "related" terms. In the case of Pain for example, a fairly standard expansion using the immediate parent, and the immediate "offspring" i.e. terms below Pain in the ontology yields the following potential expansion of 5 Parent terms + 19 Child terms,

giving a total of 24 potential expansions, where an average of 1 parent term and 7 child terms would be needed if a correct term location was known at the start (There is some overlap between children of the different original terms). A simple expansion that does not understand the intended location of the query term may lead to many irrelevant results being returned, as the user presumably has one particular meaning of "Pain" in mind.

Studies of log files of search tools have shown that users rarely modify their query or look beyond the first page of results, and often use very simple search strings(Silverstein, Marais et al. 1999). In addition users are unlikely or unwilling to undergo the training needed to use "expert" search techniques (Borgman 1996) - for example controlled vocabularies or concept browsers in order to unambiguously identify homonyms.

2.2 Fuzzy ontology and collective searching

A fuzzy ontology membership value can therefore be used to identify the most likely location in the ontology of a particular term. Each user would have their own values for the membership assigned to terms in the ontology, reflecting their likely information need and world view – However, this still requires the appropriate membership values to be assigned to each occurrence of a term for a particular user. This process can be performed in a number of ways, but the most direct approach – of assigning values during the searching process is unlikely to succeed. This is because of the reluctance of users to use existing term browsers – as noted above. The use of relevance feedback is potentially more fruitful, but this has limited application in the context of queries dealing with previously unused terms. A collective searching system may be more successful in this case.

The concept of a collective searching system has been introduced in an earlier paper (Parry 2001). The key element is the creation of groups of users based on common professional group, status and task. These groups can then share a base set of membership values for each term in the FuzzOnt, with modifications via an incremental learning algorithm when queries take place and the relevance score noted. As part of the data structure of a FuzzOnt the number of queries using a term is recorded along with the membership value for that term. Currently the updating of the group FuzzOnt is weighted by an extremely simple algorithm where the new membership (μ_{New}) is determined by the old membership (μ_{Old})) the membership calculated for this query(mi), and the number of queries that have confirmed the intended meaning of this term (Q_{Hist}).

$$\mu_{New} = \mu_{Old} \pm (\sqrt{(\mu_i - \mu_{Old}}\,)^2 / Q_{Hist})$$

The membership value of any other equivalent terms are decreased or increased in proportional amounts in order to maintain normality. For example, consider 3 locations for a particular term, (L_1, L_2, L_3) with membership values (μ_1=0.6, μ_2=0.3, μ_3=0.1). If L_1's membership value is decreased to 0.5 then the extra membership values are split in the , so that the new value for μ_2 is given by:

$$\mu_2(New) = \mu_2(Old) \pm \mu_{Change} x \left(\mu_2(Old) \Big/ \mu_2(Old) + \mu_3(Old) \right)$$

Where μ_{Change} is the change in μ_1 in this case –0.1.

The new values are then (μ_1=0.5, μ_2=0.375, μ_3=0.125).

This scheme is very simple and other incremental learning schemes may be adopted in future

A similar algorithm performs the individuals' FuzzOnt updating, however in this case the number of queries represents the number of queries by that particular user. In the future a "user authority" factor may need to be introduced, related to the degree of experience or the degree of similarity between a user's FuzzOnt and the collective one in other cases.

It is important to note that the updating process will produce different values for the individual and the collective membership values within the FuzzOnt for the same term. Therefore a collective FuzzOnt, and an individual FuzzOnt both need to be maintained.

3 Membership values in a fuzzy ontology

Assigning the membership values of each term in each location is based on the membership function shown in figure 2. Note that each term identified within a document may have its membership value updated when the document is examined.

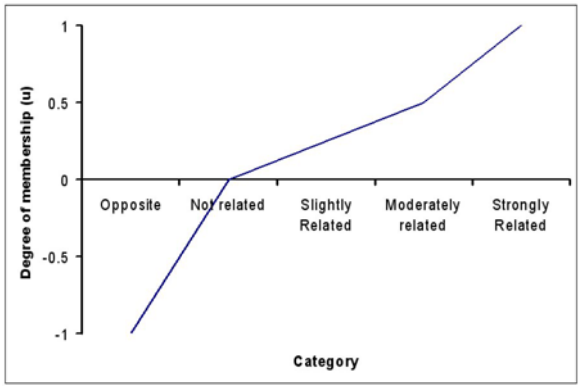

Figure 2: Membership function for fuzzy ontology

3.1 Manual assignment of membership values

This method uses the membership function shown in figure 2. For simplicity, at present the method does not involve combining terms in searches. The user performs a query and a set of documents is recovered (the process is described in more detail in section 4). For each document in the recovered set, terms that exist within the ontology are extracted automatically, affixes removed by stemming where possible using a lookup table. The user can then drag these terms into "related boxes" according to the degree of relatedness to the original query term. There are boxes for "Opposite", "Not Related", "Slightly Related", "Moderately Related" and "Strongly Related". A value for "Usefulness" of the document is also recorded via a slider. "Opposite" is used as a shorthand for " this term is not related to the desired target, and documents

containing this term are unlikely to be relevant – so for example in searching for "Cold" the term "Frigid" would be assigned to the opposite category, if the user was looking for documents about the common cold. The ontology update then takes place in two stages:

1. The intended ontology location of the search term is calculated. Only documents that receive a usefulness value greater than a certain threshold are included in the process. The query term is then compared to the terms in the "related terms boxes".
2. A score is calculated for each potential meaning of each query term by summing the membership values of terms in the "related terms boxes" that are related to each potential location of the query term.
3. The location of the query term with the largest score is assumed to be the location the user intended.

To calculate the membership value of the query term in a particular location, as indicated by the users response to the retrieved document then the following formula is used:

$$\mu_{result} = \frac{\sum_{0}^{i=n} \mu_i}{n}$$

Where μ_i is the membership value for each term the user has put into in the "related boxes". Only terms that are parents or children of the preferred meaning of the query term are included in this part of the calculation. If a term from the retrieved document occurs more than once, then each instance of the term is included. The value n is given by the number of such terms, including duplicates. If the calculation yields a membership value of <0 then the value is reset to 0.

3.2 Automatic Membership value assignment

Initially sets of query terms were derived from the set of MeSH headings present in the UMLS. These were then used as the basis for queries run against both GOOGLE (www.google.com) for the World Wide Web (using the Google API) and PubMed (pubmed.gov) for Medline. In the case of terms with multiple locations, only one search was performed as the queries were not performed using any concept identification as is possible with PubMed. Each search was limited to return 10 documents. The initial term used in the query is known as the "test term".

Before the process begins, each term, which exists in multiple locations, is given an equal and proportionate membership value (i.e. if there are two locations, each would have an initial membership value of 0.5).

Each document was then searched for terms from the ontology. A weighting was introduced so that terms in the keyword section (PubMed) or meta tags (Google) were weighted as 3, terms in the title (PubMed) or headings (Google) were weighted as 2 and terms in the abstract (PubMed) or main body (Google) were weighted as one. A "local term" was defined as one, which occurs either as a parent or child of the test term. For each test term, the automatic membership function for that term occurring in a particular location in the ontology was calculated by summing the number of local terms (L_i) discovered in each section of each document they are discovered in (k) multiplied by the weighting (W_i). This was then normalised by dividing by the sum of all terms discovered (A_i) multiplied by the weighting (W_i) over all documents in the set (n).

$$\mu_{Automatic} = \frac{\sum_{0}^{i=k} L_i x W_i}{\sum_{0}^{i=n} A_i x W_i}$$

As each term has a membership value calculated for each document, the membership value in the base ontology is modified in the same way as with the group ontology in section 3.1. The difference in weight according to the number of queries performed is deliberate in order to use the fact the search engine is being queried with the results delivered in descending order of relevance.

4 Implementation of the Fuzzy ontology technique

Currently the system has been implemented in Visual Basic.Net and the interface is shown in figure 3. The system uses a database to store the base ontology and the group and individual ontologies. The Database management system used is MS SQL Server 2000. A number of other tables include synonyms for concept terms, stop words, and the locations of URL's that have been visited. One of the features of the system is that the ontology construction browser, as shown in figure 3 allows internal links to be followed and URL's opened by the user, so that documents that are not found by an initial query, run via the Google API, can be examined. A similar approach is used with PUBMED, which also allows API access to the Entrez Database.

5 Current and future experimental work

5.1 Current Status of System

Currently an implementation of this system is being trialled in the context of an academic department of obstetrics and gynaecology. Users from a number of different professional groups are being asked to use the system, using the documents retrieved in section 3.2 as a starting point, in order to verify the automatic FuzzOnt. Currently usability tests are being carried out with a number of user groups within the obstetric domain.

Term	Concept ID	Parent	Depth	Root term
Pain	G11.561.796.444	Sensation	4	Musculoskeletal, Neural, and Ocular Physiology
Pain	F02.830.816.444	Sensation	4	Psychological Phenomena and Processes
Pain	C23.888.646	Signs and Symptoms	3	Pathological Conditions, Signs and Symptoms
Pain	C23.888.592.612	Neurologic Manifestations	4	Pathological Conditions, Signs and Symptoms
Pain	C10.597.617	Neurologic Manifestations	3	Nervous System Diseases

Table 1: Locations of the term "Pain" in MeSH

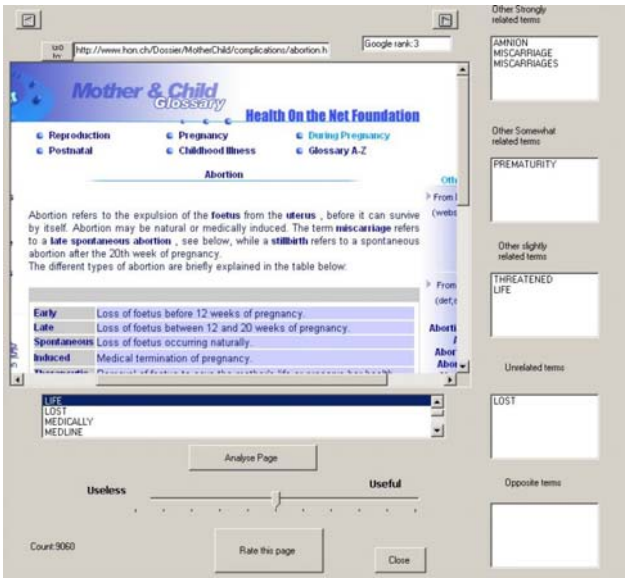

Figure 3: The user interface

5.2 Future work

Once a reasonable proportion of the ambiguously located terms have had membership values attached to them, then the resultant XML documents can be used as a filter for a search engine. The appropriate ontology for the users professional or other group is chosen. When a user searches using such a term, then expansion can be performed so that the most likely meaning is used for expansion. If this does not result in suitable information then the next query can use the next most likely meaning, or present a list of possibly related terms from the ontology. It is intended to run a clinical trial of this system where users understanding of a particular clinical question is compared when using either this system or conventional search techniques. Fortunately in Obstetrics there is a "Gold Standard" for information in terms of the Cochrane Database, (Cochrane Collaboration 1997).

The Cochrane collaboration library is a repository of "meta-analyses' that collect all published research on a particular question, rate it's validity and then combine results of suitable methodological quality. In many ways Cochrane is the source of best possible evidence, and it is presented in terms of guides to the best action in certain clinical circumstances – for example "When should steroids be given in premature labour?". Cochrane also provides a comprehensive bibliography of the topic area. Unfortunately Cochrane does not cover every possible question and tends to be rather narrowly focussed on well researched areas but for the purposes of the trial such topics can be used. Answers that the users discover can be compared to the results in the Cochrane database, and scored by experts in terms of whether such answers are consistent with best practice, and whether the information found by the users refers to good evidence.

6 Discussion

The fuzzy ontology approach may provide a simple method of reusing ontologies, where terms occur more than once in the ontology. Currently this system relies on a limited vocabulary, aligned with the ontology being used, but it is possible to imagine a scheme whereby new terms may be introduced into the ontology via the automatic assignment of a membership value, and an adaptation of the method of (Kruschwitz 2003). The key advantage of this approach is the fact that all users may are able to utilize the same ontology, but their differences can be communicated by means of the difference in the membership values of items in their ontology. This may simplify ontology reuse and communication. It may also be that the changes in membership function are asymptotic when large numbers of users or documents are involved, and the updating mechanism may not be needed after a suitable period except for novel terms. If so, then search engines may be able to present a suitable mix of results to the user based on the likelihood of the intended meaning of the ambiguous search term. The concept of "More like this" links can then be used to allow the search engine to perform an expansion using the most

likely local terms in the ontology. This approach is specifically designed to act on different groups of results returned by search engines or other tools, and so is not concerned with storage or indexing of documents. This allows it to be flexible in terms of being able to attach to the most suitable searching tool, which can be an advantage as there are specialist bibliographic systems for particular groups. For example CINAHL contains more nursing related documents and a FuzzOnt automatically constructed from that may have different location membership values than that discovered from Medline.

The concept of collaborative intelligence is a key part of this work. By allowing modification of the ontology, within limits, in terms of its success or otherwise in retrieving relevant documents, the process should improve the quality of queries. The success of the semantic web project depends not only on suitable ontologies, but also the correct application of these, to documents, and the understanding of the users requirements. There must be a form of communication between the creators of the ontologies, the document creators and the user.

This work could be extended to a much richer relationship set. Currently each term is represented in a hierarchy with a value for its membership within a simple is-a tree. However, each allowable relationship between terms in a hierarchy could have a membership value assigned to it. For information retrieval purposes, it may be useful to allow the user to select the relationships of interest and then normalize the membership value to one over those rather than over all relationships as this would penalize terms which happen to have many relationships associated with them. For example “leg” could have relationships including ‘is-a’ (limb) ‘part-of’ (body) ‘comprises’ (knee, shin etc.), ‘connects’ (foot and torso) etc. By specifically addressing the overloaded nature of terms in ontology, the Fuzzy Ontology allows the use of ontologies for information retrieval may be extended.

7. References

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The semantic web." *Scientific American*(May 2001): 29-37.

Bordogna, G. and G. Pasi (2000). Application of Fuzzy Set Theory to extend Boolean Information Retrieval. Soft Computing in *Information Retrieval: Techniques and Applications*. F. Crestani and G. Pasi. Heidelberg, Physica-Verlag: 21-47.

Borgman, C. (1996). "Why are online catalogs still hard to use ?" *Journal of the American Society for Information Sciences* **47**(7): 493 - 503.

Carroll, L. (1872). *Through The Looking Glass, and what Alice found there*. London, Macmillan and Co.

Cochrane Collaboration (1997). Cochrane Pregnancy and Childbirth Database. Oxford, UK, Update Software.

Grutter, R., C. Eikemeier and J. Steurer (2001). Up-scaling a semantic navigation of an evidence-based medical information service on the Internet to data intensive extranets. User Interfaces to Data Intensive Systems, 2001. *UIDIS 2001. Proceedings..*, Univ. of St. Gallen, Switzerland. 36-42

Kruschwitz, U. (2003). "An adaptable search system for collections of partially structured documents." *Intelligent Systems*, IEEE 18(4): 44-52.

Lassila, O. and L. McGuinness (2001). The Role of Frame-Based Representation on the Semantic Web. *Stanford Knowledge Systems Laboratory Technical Report* KSL Tech Report Number KSL-01-02. Stanford University

Moody, D. and S. G. (1999). Using Knowledge Management and the Internet to Support Evidence Based Practice: A Medical Case Study. *The 10th Australasian Conference on Information Systems*, Victoria University Wellington.

Musen, M. (2001). Creating and using Ontologies: What informatics is all about. *Medinfo 2001*, 1514. London.

Musen, M. A., J. H. Gennari, H. Eriksson, S. W. Tu and A. R. Puerta (1995). "PROTEGE-II: computer support for development of intelligent systems from libraries of components." *Medinfo*. 8 Pt 1: 766-770.

Noy, N., M. Sintek, S. Decker, M. Crubézy, R. Ferguson and M. Musen (2001). "Creating Semantic Web Contents with Protégé-2000." *IEEE Intelligent Systems*(March/April 2001): 60-71.

Noy, N. F. and D. McGuinness (2001). Ontology Development 101: A guide to Creating your First Ontology. *Stanford Medical Informatics Technical Report SMI-2001-0880* **.** Stanford University

Noy, N. F. and M. A. Musen (1999). SMART:Automated support for Ontology Merging and Alignment, Stanford University. *Stanford Medical Informatics Technical Report SMI-1999-0813.* Stanford University

Parry, D. T. (2001). Finding Useful medical information on the internet. *ANNES 2001 Fifth biannual conference on artificial Neural Networks and expert systems*, Dunedin NZ, University of Otago.

Silverstein, C., H. Marais, M. Henzinger and M. Moricz (1999). "Analysis of a very large web search engine query log." ACM SIGIR Forum **33**(1): 6--12.

Stephens, L. M. and M. N. Huhns (2001). "Consensus ontologies. Reconciling the semantics of Web pages and agents." *Internet Computing*, IEEE **5**(5): 92-95.

Widyantoro, D. H. Y., J. (2001). Using fuzzy ontology for query refinement in a personalized abstract search engine. *IFSA World Congress and 20th NAFIPS International Conference, 2001*

Zadeh, L. (1965). "Fuzzy Sets." *Journal of Information and Control* **8**: 338-353.