Fuzzy Ontology and Intelligent Systems for Discovery of Useful Medical Information

David Tudor Parry

2005

A thesis submitted to Auckland University of Technology in fulfilment of the Degree of Doctor of Philosophy

List of Figures

List of Tables

Attestation of Authorship

I herby declare that this submission is my own work and that, to the best of my knowledge and belief it contains no material previously published or written by another person nor material which to a substantial extend has been accepted for the award of any other degree or diploma of a university or other institution of higher learning, except where due acknowledgement is made in the acknowledgements.

Acknowledgements

I would like to thank my two supervisors - Philip Sallis and Nik Kasabov for their help and thoughtful support during the completion of this thesis. Special thanks go to my wife Emma, for all her patience, and especially for continuing to believe in the various deadliness before they rushed past! Love to Alice who is older than this thesis, and Rosie who is younger – I hope you aren't too traumatised by it. Grateful thanks to Janelle Muller for the proofreading and to AUT library for their almost miraculous interloan response time. I would also like to thank all the people in the Business Faculty, and School of Computer and Information Sciences at AUT who have had to put up with all my demands.

A note on grammar

Occasionally hyphenated words abound in this field. In particular 'evidence (-) based' and 'Internet (-) based' and 'on (-) line' occur in all possible forms in the literature. 'Email' seems secure but 'ecommerce', 'eCommerce' and 'e-commerce' are still in conflict. As a general rule, in this thesis, all compound words have been written to minimise the use of hyphens in order to minimise the disruption to reading flow – so evidence based, online, ecommerce are preferred. The reader's indulgence is sought for those occasions when this conflicts with references used in the text.

Abstract

Currently, reliable and appropriate medical information is difficult to find on the Internet. The potential for improvement in human health by the use of internet-based sources of information is potentially huge, as knowledge becomes more widely available, at much lower cost. Medical information has traditionally formed a large part of academic publishing. However, the increasing volume of available information, along with the demand for evidence based medicine makes Internet sources of information appear to be the only practical source of comprehensive and up-to-date information. The aim of this work is to develop a system allowing groups of users to identify information that they find useful, and using those particular sources as examples develop an intelligent system that can classify new information sources in terms of their likely usefulness to such groups. Medical information sources are particularly interesting because they cover a very wide range of specialties, they require very strict quality control, and the consequence of error may be extremely serious, in addition, medical information sources are of increasing interest to the general public. This work covers the design, construction and testing of such a system and introduces two new concepts – document structure identification via information entropy and fuzzy ontology for knowledge representation. A mapping between query terms and members of an ontology is usually a key part of any ontology enhanced searching tool. However many terms used in queries may be overloaded in terms of the ontology, which limits the potential use of automatic query expansion and refinement. In particular this problem affects information systems where different users are likely to expect different meanings for the same term. This thesis describes the derivation and use of a “Fuzzy Ontology” which uses fuzzy relations between components of the ontology in order to preserve a common structure. The concept is presented in the medical domain. Kolmogorov distance calculations are used to identify similarity between documents in terms of authorship, origin and topic. In addition structural measures such as paragraph tags were examined but found not to be effective in clustering documents.

The thesis describes some theoretical and practical evaluation of these approaches in the context of a medical information retrieval system, designed to support ontology-based search refinement, relevance feedback and preference sharing between professional groups.

Chapter One

Consider the tune, not the voice; consider the words, not the tune; consider the meaning, not the words.

Bhutanese Proverb

This chapter introduces the problem domain and is a critical review of some of the previous work in this area. Section 1.1 covers the aim of the thesis. Section 1.2 discusses the original contribution of this work. Section 1.3 gives some background to this work. Section 1.4 discusses the concept of useful information and gives some pointers for information evaluation. Section 1.5 describes the organisation of the thesis.

1.1 Aim of the work

The major aim of this work is to describe the architecture and elements of a computer system to support the searching needs of professionals and the general public in the medical domain. Development of methods to support the identification of similarities between documents, and discovering personal and group views of knowledge are essential prerequisites for this.

1.2 Original contribution in this work

There are four main strands of original work in this thesis:

1. The identification, derivation and study of the "fuzzy ontology" concept. Publications related to this work include:
   - Parry, D. *A fuzzy ontology for medical document retrieval,* in *The Australasian Workshop on Data Mining and Web Intelligence*, M. Purvis, Editor. 2004, Australian Computer Society: Dunedin. p. 121-126.
   - Parry, D. *Fuzzification of a standard ontology to encourage reuse*. In *The 2004 IEEE International Conference on Information Reuse and Integration (IEEE IRI-2004)*. 2004. Las Vegas, USA.
2. The use of the zipping method for comparison between documents, with extensions to the Kolmogorov distance calculation. See:
   - Parry, D.T. *Use of the Zip method for Author Identification in an Electronic Forum (poster)*. In *Neurocomputing and Evolving Intelligence (NC & EI) 2003*. 2003. Auckland.

3. The identification of the need for useful information to be made available to clinicians at the point of care, and the development of architectures and systems to do this. Publications on this topic include:
    - Parry, D.T. *Finding Useful medical information on the Internet*. In *Annes 2001 Fifth biannual conference on artificial Neural Networks and expert systems*. 2001. Dunedin NZ: University of Otago.
    - Parry, D.T. *Finding medical information on the Internet: Who should do it and what should they know.* SIM quarterly, 1998(4).
    - Parry, D. and E. Parry. *Development of an intelligent system for finding useful medical information*. In *Neurocomputing and Evolving Intelligence (NC & EI) 2003*. 2003. Auckland.
4. The testing of the usefulness of such a system in a clinical setting has been described in:
    - Parry, D.T. *Evaluation of a fuzzy ontology based medical information system* International Journal of Health Information Systems and Informatics (in Press).

1.3 Background

This thesis deals with the area of information retrieval from Internet-based sources. In particular it is based around the notion that there is potential for improving medical care, or any professional activity, by increasing the efficiency of information retrieval from Internet-based sources of knowledge. Medicine is used as a domain for investigation because of personal experience, the availability of large amounts of information, concerns about quality and relevance, and the degree of previous work on the development of effective information retrieval strategies.

Medicine is an extremely knowledge intensive activity. Medical professionals can be defined as “knowledge workers” in that “they themselves are changed by the information they process” (Kidd, 1994) p186. The long tradition of linkage between clinical work and research stretches back as far as the foundation of modern medicine, to Hippocrates and before. Information is transformed into knowledge as part of the clinical research process, by observation, and intervention on patients, and recorded in various ways such as in journal articles, textbooks etc. The same clinicians and others then use this knowledge for the treatment and diagnosis of future patients. The rate of production of medical knowledge is enormous – currently MEDLINE adds 500,000 references a year. The growth of knowledge sources and the number of documents available has not solved the problem of appropriate information being made available for patients- see for example the work of Williamson on the information needs of breast cancer patients (Williamson, 2005). The teaching of information retrieval techniques to medical professionals has been poorly studied – a recent review article found only 3 studies on the effectiveness of teaching of information retrieval skills that could be regarded as well-designed (Garg & Turtle, 2003). Kidd makes the strong argument that such workers are constantly redefining their model of the world in response to new information, and that "passive" information storage is of limited use. She also makes the point that such "knowledge work" is essentially the domain of humans, and the role of information systems should be to support them rather than attempt to replace them. The need for highly available and usable information sources in medical work view is strongly emphasised by the article (Jeremy C Wyatt & Sullivan, 2005b), which points out that each week doctors generate around 45 clinical questions. This article gives some tips for finding the answers to such questions, but the perfect solution is not yet available according to Sullivan and Wyatt. Use of recorded knowledge depends on access to knowledge stores.

This thesis is solely concerned with those knowledge stores that can be accessed electronically. Specifically, in this thesis "electronic sources", mean sources available via the Internet. There are other sources, such as CD-Rom or videotapes, which are not accessible this way. This thesis follows convention and used the term "document" to refer to any individual information source. Rather than use a cumbersome term such as "non-Internet accessible documents" the term "paper document" is used. In many cases, the electronic resources may represent a subset of the information available – for example, some journals have only part of their publication online, or only material published after a certain date. In other cases, the electronic sources act as an index or bibliography for material that is published on paper. In other cases the information may be available only in electronic form.

1.3.1 Electronic Knowledge sources

There is a very wide range of knowledge sources in the medical field. Material may be available at no cost (For example the British Medical Journal), on a subscription basis (such as the Lancet or various online text-books), as part of a professional membership (such as continuing medical education material available to the Royal NZ College of General Practitioners), or supplied as part of an academic course. In practical terms, these sources may demand identification and registration of the user before they can be accessed.

Documents usually form part of a collection. Searching may occur within defined collections or between a number of them. For example, journal articles from the same publication may be available together. Often documents from a particular author or group are available from the same place. These represent primary or original documents. Digital libraries or bibliographies may bring together copies of these documents, or information that allows the user to identify them. Digital libraries always include some sort of editorial control in terms of the selection of the documents that are contained. Search engines and Internet portals use automatic, or semi-automatic means to identify index terms for documents accessible via the Internet, and by doing this make documents accessible by the search engine into a "virtual" collection where membership is defined by the criteria used by the search engine for finding and indexing documents.

There is a degree of mapping between the location of the document, and the access control. Many digital libraries store the document information in a Database management system (DBMS), and generate the document that can be viewed dynamically. Such WWW pages are collectively known as the "hidden web" as they cannot be accessed directly by search engines. This technology also makes access control easier for the owner of the site. General WWW documents, or documents that make up electronic journals may also discourage harvesting by search engine "bots" by means of meta-tags or other techniques.

1.3.2 Examples of Knowledge sources

The National library of Medicine (USA) "MEDLINE" index covers more than 11 million journal articles, published since the 1960's. Historically, MEDLINE can trace its heritage back to the work of Dr. John Shaw Billings (1838-1913) who served as a surgeon in the American Civil War and the director of the National Library of Medicine between 1865 and 1895. During this time "INDEX Medicus" was first published (1879). Index Medicus is an expanding, regularly updated, bibliography covering many medical journals, containing information about author, source, subject and often an abstract of the article. MEDLINE has acted as a successor to this effort. Around 500,000 articles are added to MEDLINE every year, from more than5000 journals. MEDLINE is accessible via a number of sites including "PubMed" and "Ovid".

MEDLINE is available via CD ROM and via the Internet in the PubMed format at no cost to users. Other related databases such as CINHAL (mostly nursing–related) and PSYCHLIT are available from the National Library of Medicine.

Other sources of information include journals not included in MEDLINE, for example basic science, textbooks, WWW pages, databases such as Cochrane, nursing and social science journals, information from newsgroups, Theses and unpublished material. Much of this material is intended to be accessible primarily to professional workers in the field, but there is also a large amount of material specifically written for the public, or for professionals in related fields. Unlike many other sciences, clinical science research is often understandable by non- professionals when certain aspects of vocabulary and convention are understood.

Electronic sources can provide variable amounts of information – for example some digital libraries only provide abstracts, some full text and some bibliographic information, Table 1 shows some of these. For example MEDLINE contains abstracts for many of its entries, but not all, and also sometimes provides hyperlinks to the full-text articles. The quantity of information provided may be extremely important, it is a common experience to find that the abstract or title do not provide sufficient details to make a decision on the validity of the work. Fortunately work by (Wilczynski, McKibbon, & Haynes, 2001) has shown that analysis of abstracts for relevance gives similar results to an analysis of the whole document i.e. abstracts are “about” the same thing as the documents they come from.

1.3.3 Use of Knowledge sources in Medicine

The traditional professional model of medicine involves long and rigorous training, followed by a certain degree of continuing medical education (CME) in order to fulfil regulatory requirements. It has been shown that traditional CME is ineffective at providing lasting knowledge gains. In addition it has been shown that knowledge gained in initial training is lost at a fairly high rate (D. Sackett, Richardson, Rosenberg, & Haynes, 1997). On top of this is the very large yearly increase in the amount of scientific and clinical knowledge that is required for effective practice. This knowledge gap has lead to the development of the set of techniques known as “Evidence Based Medicine” (EBM).

Table 1: Types of Knowledge Source

Source	Electronic?	Peer reviewed?	Searchable?	Current?
Journals – Full text	Some	Yes	Sometimes	Yes
MEDLINE Journals	Yes	Most	Yes – sometimes no abstracts	Yes
Textbook	Some (few)	No	Often not	May not be
Commercial WWW pages	Yes	No	Yes	Often not
Other Electronic Databases	Yes	Often	Yes	Yes

EBM involves a process of making clinical decisions based on the most recent, good evidence from clinical studies. This is contrasted with decision making based on personal experiences, old or uncertain evidence as found in textbooks or remembered from training or derived from what seems the most likely outcome from "first principles". The five major tenets of EBM are;

1. Clinical decisions should be based on the best available evidence
2. The search for evidence should be based on the clinical problem,
3. Identifying "best" evidence is done via statistical and epidemiological methods
4. The process is only useful if the answers are applied to clinical problems,
5. Audit of practice is vital

(Adapted from (Davidoff, Haynes, Sackett, & Smith, 1995))

The aim of EBM is to allow clinicians to answer clinical questions quickly and reliably in an environment where clinical research is published in a wide range of sources, and traditional methods of dissemination of results cannot keep up. EBM involves not only searching for information but also assessing it in terms of the quality of the studies and the relevance of the result to the current question. Question formulation is particularly important in EBM where questions of the form "What is the most effective treatment for X.." are preferred over more general questions such as " What is new in the treatment of Y.." There remains continuing concern that medical training does not provide the information retrieval skills needed to successfully access information sources (Shaughnessy & Slawson, 1999). Despite this, projects to allow access to electronic information sources are extremely popular (Moody & Shanks, 1999). More recently, training programs for medical undergraduates have been introduced and evaluated (Gruppen, Rana, & Arndt, 2005). It is interesting to note that in this work, although the training brought improvement there were still large numbers of sub-optimal searches by the students both before and after the training. Other recent work using a focus–group approach has identified lack of access and skills as a reason for abandoning the search for the answers to clinical questions before the information need is satisfied (Green & Ruff, 2005).

There are criticisms of the approach adopted by EBM. In their review article (Cohen, Stavri, & Hersh, 2004) point out a number of potential arguments against EBM. Mostly these relate to its definition of evidence, in particular the hierarchy of quality of evidence, one of the most ironic is that that use of EBM has not been shown to be effective by its own standards – i.e. there has been no controlled trial to demonstrate better patient outcomes in cases where the practitioners use EBM. However, even in this article, the authors are at pains to point out that evidence is important for patient care, and suggest a pragmatic approach aiming towards a "best current answer". One of the major issues in the criticisms of EBM is that it has been too restrictive in its choice of evidence and by widening the range of information sources this issue can be addressed. It should also be pointed out that some of the original authors of EBM pointed out in 1995 that EBM in practice was less prescriptive than some of the original models that had been considered (D. L. Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996), and that clinical work demands both internal expertise and external evidence. A recent editorial by (David C. Slawson & Shaughnessy, 2005), suggests that "information management" by which they mean general information retrieval and assessment skills including 'foraging' for information may be more important than the clinical statistics and meta-analysis side of EBM.

EBM also introduces the concept of partnership with patients since they are the ultimate beneficiaries of the care decisions that are made, which in turn are informed by medical knowledge. Recently the concept of the "smart patient", originally introduced in (R. L. Sribnick & W. B. Sribnick, 1994) has become more popular. In particular the patient is seen as one of the information sources and users in an eHealth network, where personalised information is available for the patient, and the patient is encouraged to collaborate and take ownership of the process of recording symptoms, treatments and outcomes (Jeremy C Wyatt & Sullivan, 2005a). In this model, the patient themselves takes responsibility not only for understanding the choices explained to them by the health professional (Informed consent), but also is encouraged to attempt to discover relevant knowledge about the condition. This approach acknowledges that patients often have access to the same knowledge sources as their carers, as well as additional sources of information, but also that the knowledge requirements of patients may be different, for example that in a chronic disease symptom alleviation may be as important as cure. As is pointed out in the review article by (Powell, Lowe, Griffiths, & Thorogood, 2005), there is still a dearth of work on what consumers actually do with health information. Concerns that the easy availability of health information on the Internet will have negative effects are common, but there has been little work in this area

Within the broad categories of professional and layperson, there are many sub-categories. These can include doctors in training, specialists in the field, generalists, researchers, managers or politicians with responsibility for service provision, friends or relatives of the patients and members of pressure groups or Non-Governmental Organisations hoping to influence policy research or treatment. All these diverse groups can share members and all are ultimately dependent on understanding knowledge that may be obtained from documents concerned with health information. Ultimately, this thesis is concerned with the efficient and effective understanding of this knowledge and the ways that such knowledge can be queried, classified and organised. (Shaughnessy & Slawson, 1999) coined the term “POEMS” (patient oriented evidence that matters), and have developed a commercial system (Inforetriver, available via http://www.infopoems.com/index.cfm) to support the dissemination of these. A more recent article by (Shaughnessy & Slawson, 2003), has shown that such POEMS are not being universally reported in review articles. These comments are interesting because such POEMS as in the case of the diabetes care guidelines described, may be controversial in their interpretation of the data. The compact form of such POEMS – that may be presented in 200 words or less – can conceal uncertainties in the reliability or precise meaning of the evidence.

Recent work (Wood, Benson, LaCroix, Siegel, & Fariss, 2005) has looked at the use of medical information sources. By use of Neilson ratings this work attempted to study the relative use of PUBMED and other NLM information sources. For September 2004, the figures are shown in Table 2

Table 2: Usage of Medical sites (from (Wood et al., 2005)) September 2004

Site	Number of Unique US visitors/Month
WebMD	2.5 million
NIH.gov	2.4 million (see text)
AOL Health (powered by WebMD)	1.7 million
Yahoo! Health	1.2 million
MSN Health (at the time, powered by WebMD)	950000

Of the visitors to NIH.gov it was estimated from internal logs that around 45% accessed the database area that includes PUBMED, and around 39% accessed MedlinePlus – a combination of full text articles and summaries and reviews. Worldwide usage of NIH websites went from 11.7 Million unique visitors in April 2004 to 17.1 Million unique visitors in April 2005. These figures use IP address as a marker of uniqueness and may be overstated by the use of dynamic IP address allocation. This data does not reveal what sort of person is accessing the site, nor whether their access was successful.

1.4 Useful Information

"Useful" is a very broad term. In this thesis it is designed to cover a broader area than "relevant" which has been represented a number of ways in the field of information retrieval. In particular, a useful search is one that helps provide a solution to a clinical problem. This may include failing to find any information or realising that a query returns an enormous number of results. Indeed (Eastman, 2002) points out that when dealing with large retrieval sets, reducing the size of the whole set may be less important than increasing the precision of the "top 10". Relevance is described in more detail later in section 2.1.1. Five criteria for usefulness that apply to primary care doctors are given in (D.C. Slawson & Shaughnessy, 1997) page 949:

> *"Does the information focus on an outcome that my patients care about?*
> *Is the issue common to my practice, and is the intervention feasible?*
> *If the information is true, will it require me to change my practice?"*

A wider definition of usefulness is needed to support all potential users of medical information.

1.4.1 A framework for identifying useful information

A number of attempts have been made to codify the “levels of evidence” available from published materials in the medical domain. These have been described recently by (Harbour & Miller, 2001), and popularised in books such as “Evidence Based Medicine” (D. Sackett et al., 1997). Table 2 shows a simplified representation of these guidelines.

Table 3: Levels of evidence

“Best” evidence	1	Systematic review
	2	Randomised controlled trial (RCT)
	3	Case -control
	4	Cohort
“Worst” evidence	5	Expert opinion

PubMed (National Library of Medicine, 2002) supports the identification of the first two – systematic review and RCT - levels as MeSH index terms. The Cochrane collaboration essentially deals with the first two levels, but it is arguable that documents that do not fall into these categories are potentially useful, if only because not all decisions are amenable to RCT’s for reasons of the rarity of the conditions, difficulty of performing RCT’s, or the nature of the question. This is one of the points raised in (Cohen et al., 2004). The following criteria have been broadly derived from (D. Sackett et al., 1997) and personal experience.

A recent practical guide to evaluating literature in the primary care domain (Alborz & McNally, 2004) introduces ideas relating to filtering. Figure 1 shows the process.

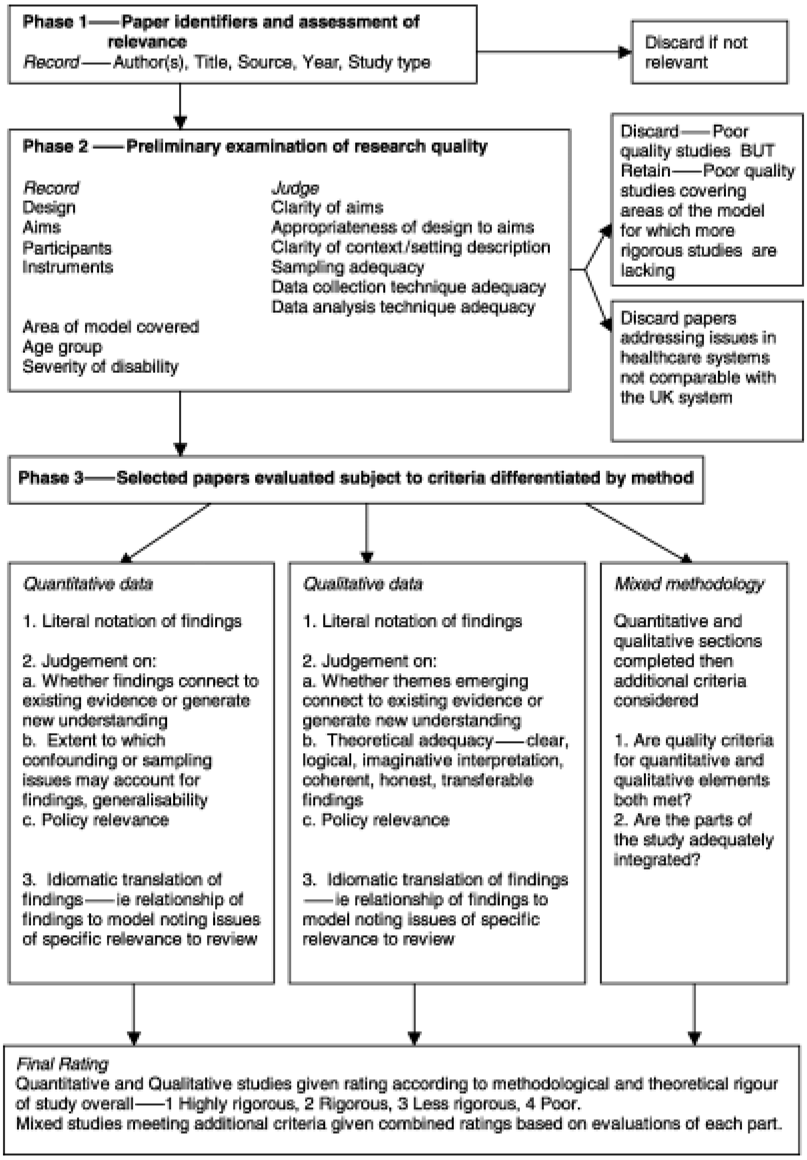

Phase 1——Paper identifiers and assessment of relevance
Record——Author(s), Title, Source, Year, Study type
Discard if not relevant
Phase 2 ——Preliminary examination of research quality
Record
Design
Aims
Participants
Instruments
Area of model covered
Age group
Severity of disability
Judge
Clarity of aims
Appropriateness of design to aims
Clarity of context/setting description
Sampling adequacy
Data collection technique adequacy
Data analysis technique adequacy
Discard——Poor quality studies BUT Retain——Poor quality studies covering areas of the model for which more rigorous studies are lacking
Discard papers addressing issues in healthcare systems not comparable with the UK system
Phase 3——Selected papers evaluated subject to criteria differentiated by method
Quantitative data
1. Literal notation of findings
2. Judgement on:
a. Whether findings connect to existing evidence or generate new understanding
b. Extent to which confounding or sampling issues may account for findings, generalisability
c. Policy relevance
3. Idiomatic translation of findings——ie relationship of findings to model noting issues of specific relevance to review
Qualitative data
1. Literal notation of findings
2. Judgement on:
a. Whether themes emerging connect to existing evidence or generate new understanding
b. Theoretical adequacy——clear, logical, imaginative interpretation, coherent, honest, transferable findings
c. Policy relevance
3. Idiomatic translation of findings ——ie relationship of findings to model noting issues of specific relevance to review
Mixed methodology
Quantitative and qualitative sections completed then additional criteria considered
1. Are quality criteria for quantitative and qualitative elements both met?
2. Are the parts of the study adequately integrated?
Final Rating
Quantitative and Qualitative studies given rating according to methodological and theoretical rigour of study overall——1 Highly rigorous, 2 Rigorous, 3 Less rigorous, 4 Poor.
Mixed studies meeting additional criteria given combined ratings based on evaluations of each part.

Figure 1: Study Selection process from(Cohen et al., 2004)

1.4.2 General Quality

1. Peer –reviewed

   World Wide Web (WWW) sites as well as journals may now have peer-review in place. In Figure 1 peer review is not mentioned because it is assumed to occur for all selected studies. Peer review is not completely standardised, and depends partly on the source of the work, whether conference proceedings, journal article or published book for example

2. Randomised Controlled Trial (RCT)

   This is the gold standard for clinical interventions although many interventions have not been subjected to this process or are suitable for this approach. An RCT generally involves two groups of patients, one of which is given the intervention and one which is not. Members of these groups are randomly chosen. RCT's are prospective – that is the decision as to which group a person is in is made before the start of the intervention. Other approaches are also commonly used, and a good overview is found in (Grimes & Schulz, 2002).

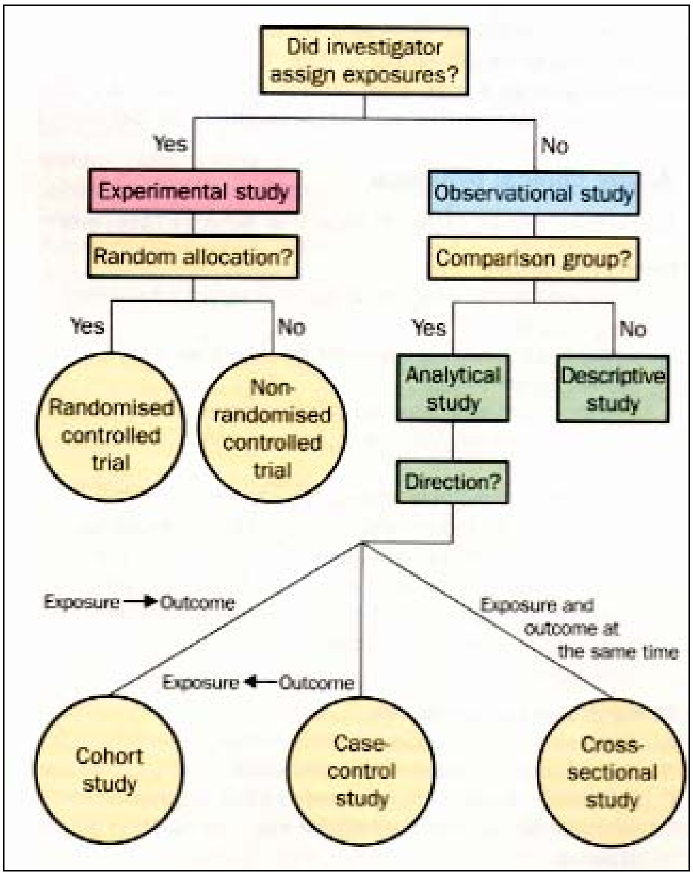


Figure 2: A Taxonomy of Studies – from (Grimes & Schulz, 2002)

There are also issues of the quality and power of a trial. In some cases meta-analysis can cause smaller trials to loose credibility, as larger, inconclusive studies wash-out the results from smaller studies which are focussed on a particular sub-group or were particularly well-performed.

3. High citation number

   This is more of a rule of thumb than an absolute factor. If the source is frequently cited then it indicates that large numbers of authors have found it relevant. It is perfectly possible that a particularly bad study may have a high citation index, or that the index may be inflated for other reasons such as age of the reference. It is possible to infer that references cited in 'good' documents are more likely to be good themselves but this is dangerous to extend too far.

4. Recent

   This depends on the rate of change of the field. Documents in very active research areas are more likely to have a shorter useful life than those in slow-moving or moribund areas.

5. Significant result

   A document containing information that a treatment or diagnostic method is effective, and that this effect is large, is likely to be more useful than one that does not. If there is a traditional treatment that is shown to be ineffective then this also is significant.

6. Authoritative Source

   For electronic sources of information the Health on the Net Code of conduct can give some guidance – otherwise, inclusion by indexes or directories e.g. MEDLINE or Cochrane can lend authority. The author affiliation can be an important issue here. An automated system for "authoritativeness" is described by (Farahat, Nunberg, Chen, & Heylighen, 2002). This system relies on linkages between references in terms of a citation score.

7. Usability – in terms of traditional web usability, for example Neilson's heuristics (Neilson, 2000), and also in terms of technical issues such as plug-ins media etc.

8. Appropriate language;

   Is this information in a suitable language for use by clinicians, or is it designed for lay people? In this situation the information may be too imprecise to be of use. In the opposite case, the information may be too technical.

1.4.3 Clinical relevance

Some of these issues are included in PUBMED – the first three items come directly from the PUBMED filter list.

1. Deals with humans;

   Although animal studies, or theoretical ones may be of great use – for example in the case of poisoning or electric shock, human studies are often essential.

2. Appropriate sex;

   Included in this is whether the interventions are safe for pregnant women, and the variation in body sizes and compositions between the sexes.

3. Corresponding age group – see Table 4. Often broader bands are used, or bands that reflect characteristics of the individual rather than his or her age.

Table 4: Age Bands – from PUBMED

<20 Weeks	Embryo
<40 weeks	Foetus
<1 Year	Neonate
<5 Years	Infant
<10 years	Child
<18 Years	Adolescent
<60 Years	Adult
60+ Years	Geriatric

4. Speciality is appropriate;

   Information designed for one medical specialty may not be appropriate for others, for example between pathologists and other clinicians. Similarly the information requirements of different clinical groups e.g. Physiotherapists and Surgeons treating a patient with an artificial hip may have different needs.

1.4.4 Clinical usefulness

This list is based around classifications found in PUBMED and sources of clinical protocols. One point to realise is that the number of studies that are examined and rejected is very large.

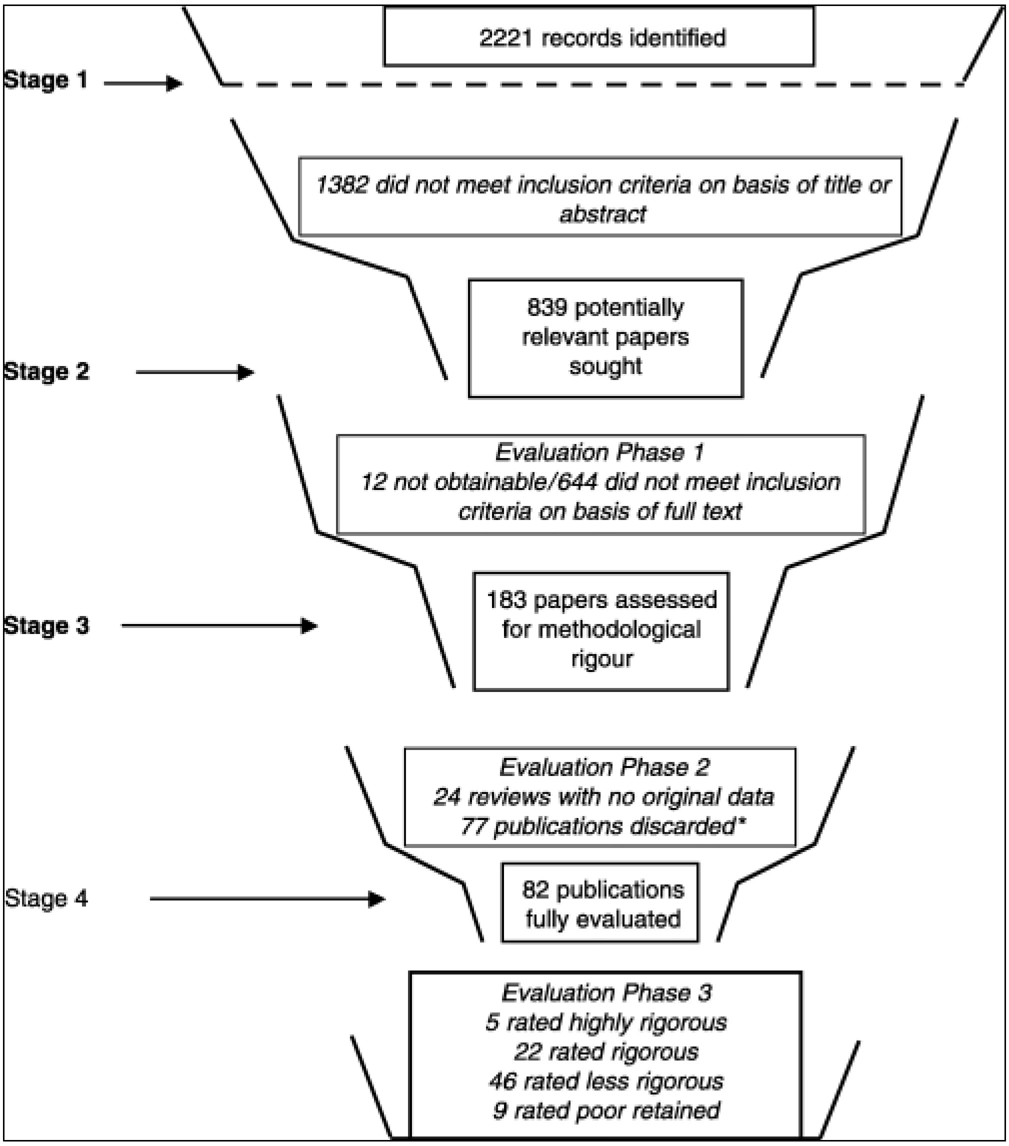

2221 records identified
Stage 1
1382 did not meet inclusion criteria on basis of title or abstract
839 potentially relevant papers sought
Stage 2
Evaluation Phase 1
12 not obtainable/644 did not meet inclusion criteria on basis of full text
183 papers assessed for methodological rigour
Stage 3
Evaluation Phase 2
24 reviews with no original data
77 publications discarded*
82 publications fully evaluated
Stage 4
Evaluation Phase 3
5 rated highly rigorous
22 rated rigorous
46 rated less rigorous
9 rated poor retained

Figure 3: Filtering of studies occurring in(Cohen et al., 2004)

It is interesting to note that even searching within indexed bibliographic sources, (Figure 3); only 82 studies were retained from 2221 original ones. The area of the research of (Cohen et al., 2004), health services delivery, has publications using many different methodologies, and many different speciality bases.

1. Appropriate to stage of encounter (e.g. therapy, diagnosis etc.);

   This also excludes information that is purely of a research nature, if better information for the clinical decision is available. However such information can be useful if it casts doubt on current clinical practice, or can help explain otherwise unexpected results.

2. Deals with available therapeutic or diagnostic tools;

   This includes such aspects as whether the drugs or procedures involved are licensed or available in the location, and acceptable in terms of cultural factors and cost.

3. Suitable format;

   Are the documents or information sources able to be read by the user; correct language, is a machine reader available. Concrete examples of this include different varieties of microfiche, or PDF files that may require large bandwidths for download.

4. Available in a timely fashion;

   Broadly the information may be available immediately (read off the screen- a time period of seconds), quickly (within the library or searching area - a time period of minutes), after a short pause (if documents need to be retrieved from a nearby site- a time period of hours) or after a long time (if the document needs to be specially ordered or generated- a time period of days)

5. Locally available;

   There may be no practical method of obtaining the document at all.

1.4.5 Scope of this thesis

This thesis concentrates on the medical domain and in particular in the specialisation of Obstetrics and Gynaecology. This is partly because of the wide availability of digital knowledge sources in this area and also because of the efforts that have already been made in this field, especially the development of EBM, and tools to support it. In practical terms the author has close links with the local academic department of Obstetrics and Gynaecology. Medicine also has the advantage over some areas – in particular IT that research evidence is seen as central to good practice, although evidence based software engineering has recently been proposed (Kitchenham, Dyba, & Jorgensen, 2004). Crucially, medical practitioners also use a range of knowledge sources, from basic science articles, via clinical investigations in the academic domain, to guidelines and information items produced by 3rd parties, through to information produced by and for patients. However, many of the techniques developed are equally applicable to other fields such as business, law, agriculture/horticulture or engineering. In all these fields, the tide of information threatens to overwhelm the professionals and public. Increasingly in all these fields, the notion of the expert human as the sole and complete interpreter of the knowledge of the field, and retaining that knowledge in his or her memory is becoming obsolete. At the same time, the explosion in the quantity of knowledge has made interpretation even more important. This thesis will attempt to examine ways of finding useful knowledge on the Internet, and in particular identifying ways to make searching more efficient and satisfying for the user.

1.5 Problem outline

Despite the increasing availability of electronic information sources, searching the web for useful information is difficult. A recent large survey of WWW search engines described many approaches – especially the use of links as citations but still concluded that “A substantial amount of work remains to be accomplished” (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001), p.40. The WWW has allowed the cost of publication and access to information to be substantially lowered compared to traditional means of publication via paper and ink. A number of schemes have been devised, for example the BMJ scheme (*BMJ* 2001;323:65 (14 July)) to allow free access to medical journals that are published on the Internet to developing nations. Obviously such schemes are much easier to organise than a comparable distribution of paper-based documents. At the same time large numbers of authors have found it possible to publish material with a much wider audience than before – for example pharmaceutical companies, patient support groups, government agencies and individuals. Although much has been made of the unreliability of many of these sources, the major issue is simply dealing with the vast amount of information available, even when recovered from a search engine. Interestingly a recent study of potential biases in information on the Internet, referenced a study that discovered similar biases in paper based information sources (Slaytor & Ward, 1998)!

Studies showing lack of completeness in material available from the Internet have been criticized in a letter (Eysenbach, 2004), pointing out that consumers of healthcare rarely only look at one source, and that such criticisms have little relation to the real information gathering activities of consumers. A very comprehensive systematic review of studies of the quality of Internet-based information sources for health consumers (Eysenbach, Powell, Kuss, & Sa, 2002), has revealed that such studies have widely varying criteria for quality, many different study methods and variable conclusions. The Eysenbach criteria are shown in Figure 4. Eysenbach explicitly states that “infodemiology” (Eysenbach, 2002) is an extremely unreliable science. His model of the *potential* influences on quality which is what Figure 4 represents is interesting and may lead to some resolution in the uncertainty over the use of quality marks. However, such an approach should be indicative rather than prescriptive, for example an approach that always labelled commercial sites as unreliable because of potential conflict of interest would limit access to such useful sites as (GlaxoSmithKline, 2003), which provides a lookup between US, UK and generic names for drugs.

Other issues that may affect the quality of searches include the caching policy of any search engine, in respect to the dynamic nature of the web. Although looking dated with its description of the web containing 800 million pages, the paper (Brewington & Cybenko, 2000) introduces some very important issues as to the lifetime of pages and in particular the problems of assuming stationarity in a model of web page lifetime.

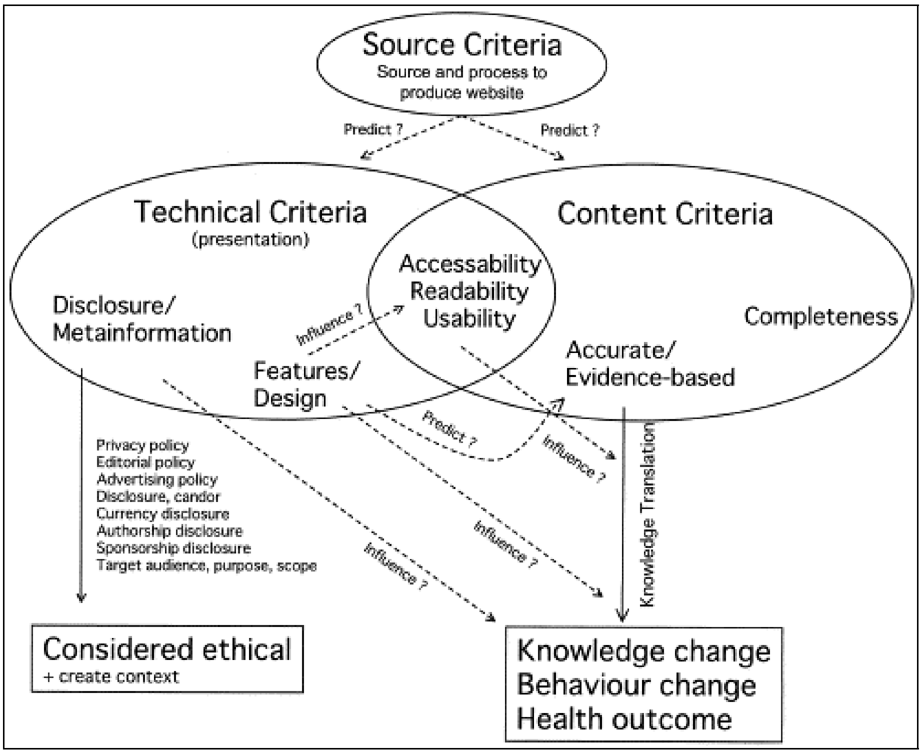


Figure 4: Quality indicators from (Eysenbach, 2002)

Stationarity assumes that each time period is identical to any others but updates are often done with respect to critical dates, for example publication deadlines of periodicals, and most updates occur within US working hours. Interestingly this paper also attempts to use webpage style as a clue to determining age, with older WebPages generally having fewer images, and less data.

Discovering useful information means accessing appropriate information in a suitable format, in a timely manner. Secondary considerations include allowing time invested in searching to be of benefit in future searches, increasing facility of use with a system and better general understanding of the problem domain. A further benefit may be the ability to communicate some of these improvements in search efficiency and conceptualisation to others for example colleagues, students or patients.

Key requirements of the system apart from the main goal include that it be easy to use, quick, able to be used for searching immediately without a long human or machine training period, and use terms and concepts that the user is likely to understand and have facility in using.

1.6 Description of the searching process

As stated previously it is assumed that the user begins with an unmet information need, located in the problem domain covered by the system. It is also assumed that the user has a reasonably large vocabulary of terms that will provide some coverage at least of the desired information. The query process is assumed to be a repeating process with the queries continuing to be modified until the user is satisfied. This process may be constrained by time, so that the level of satisfaction achieved may not be optimal, because the time available for searching has ended. Paradoxically, the opposite is also true – users may continue searching after they have discovered suitable documents because they have time available – and searching for documents on the web is commonly perceived as an intellectually involving and immersive task, see for example the results described in (P. Wang, Hawk, & Tenopir, 2000).

Three examples are given below to illustrate some of the actions and processes involved:

Example 1 – Clinical use in a consultation

Using the framework for the study of interactions described by Preece in “Interaction Design” (Preece, Rogers, & Sharp, 2002) an environment, task and user profile can be identified along with the goals of the user. The goal of the user, a clinician is to discover specific information about a treatment for a particular condition – for this scenario, the clinician is assumed to have a condition that may be being exacerbated due to adverse interaction between a number of prescribed drugs and over-the-counter remedies. The environment comprises a consulting room with Internet access, in an English speaking country that is not the US. Salient points include the fact that the screen is visible to the patient and that access may be available to some ‘pay-for’ databases and local repositories of information – as in the CIAP project (see Figure 6). The task involves finding appropriate information, ascertaining its validity and appropriateness for the current situation and using this information to advise the patient and possibly change the medication prescribed. In this case the task constraints include limited time for the process and potential uncertainty as to the most appropriate source of the information. A major sub-task would involve accurately identifying the appropriate search terms given the non- US source and the fact that although some databases such as PubMed automatically include non-US spellings in searches, this is not generally the case. In addition, the proprietary name of the over-the counter remedy may be different in different countries. The user profile of the clinician would include a fairly comprehensive knowledge of appropriate terms within their speciality. The degree of experience in general computer use – for example facility in typing, mouse movement etc. may vary between individuals. Facility in the use of searching systems is likely to be low, comparable with the general population, because of the general lack of experience in the use of such systems and the paucity of training available to working clinicians. In terms of the hardware and software environment it is possible that the clinical users will have access to high speed connections at their workplace. A display resolution of 800x600 seems reasonable, with a computer age of around 2 years or so. However a recent survey by Cullen (Cullen, 2002), indicated that although 48.6% of New Zealand General Practitioners used the Internet to access medical information, the majority of these did so via home systems rather than at the workplace (only 36.8% had access available at the workplace). This number may be higher for those working in secondary and tertiary care, but it would be much lower for those involved in domiciliary visits. It is likely that such users will have at least some beliefs about the validity of certain sites, although their knowledge of the range of information sources may be limited.

A system designed to support such consultation use should be easy to use, quick, and able to interpret technical queries.

Example 2 – Patient information

A patient may wish to access before or after a consultation. Cullen (Cullen, 2002), records that 87.8% of General Practitioners in New Zealand have experienced patients arriving at consultations with information the patient has downloaded from the Internet, although in most cases less than 10% of patients do this. Post- consultation searching may be done in the context of recommendations from a health worker or independently. Again, national context is important, in this case not only for drug availability and comprehension, but also for contact details of any support or information groups. Patients may have less understanding of technical terms and may wish to consider alternative treatments. Patient information is not always sort by the patient themselves, and in the case of children, it may involve different presentation for children and parents. A similar situation may occur when the patient does not speak the same language as the clinician, but is accompanied by a relative or friend who does speak the clinicians language or where disability prevents the patient from accessing the information directly. These users are unlikely to have access to paid-for information sources, are likely to have limited understanding of the technical vocabulary, but their experience of information retrieval may vary widely between individuals. In terms of hardware and software it is likely that the users will have dial-up access to the Internet. Such users may be naïve in terms of the validity and worth of various sources of information.

For patient information use, a system has less requirement for sped, but more need for comprehensibility, and the ability to cope with non-technical questions.

Example 3 – Research Worker

A research worker is likely to have access to the Internet via a high-speed link, whether at their desktop or via a shared space. Because of the type of questions being asked it is possible that national and linguistic differences between terms are not such a problem for this group because of their narrower, but deeper understanding of the vocabulary used. The research worker would be likely to have an understanding of the use of query terms and the quirks of particular databases, because of the prevalence of “introduction to databases” courses, and the limited set of databases generally needed for a particular research topic. The research worker is likely to have high demands in terms of the completeness and coverage of the search compared to the other two groups. In terms of the question “what do I need to know about this?” if the task involves reviewing a particular area of research – for example for an editorial or Cochrane group, then the answer is likely to be “everything”.

The system supporting the research worker needs to be able to work across a large number of databases, accept sophisticated queries and should also allow easy storage of results.

1.7 Organisation of this thesis

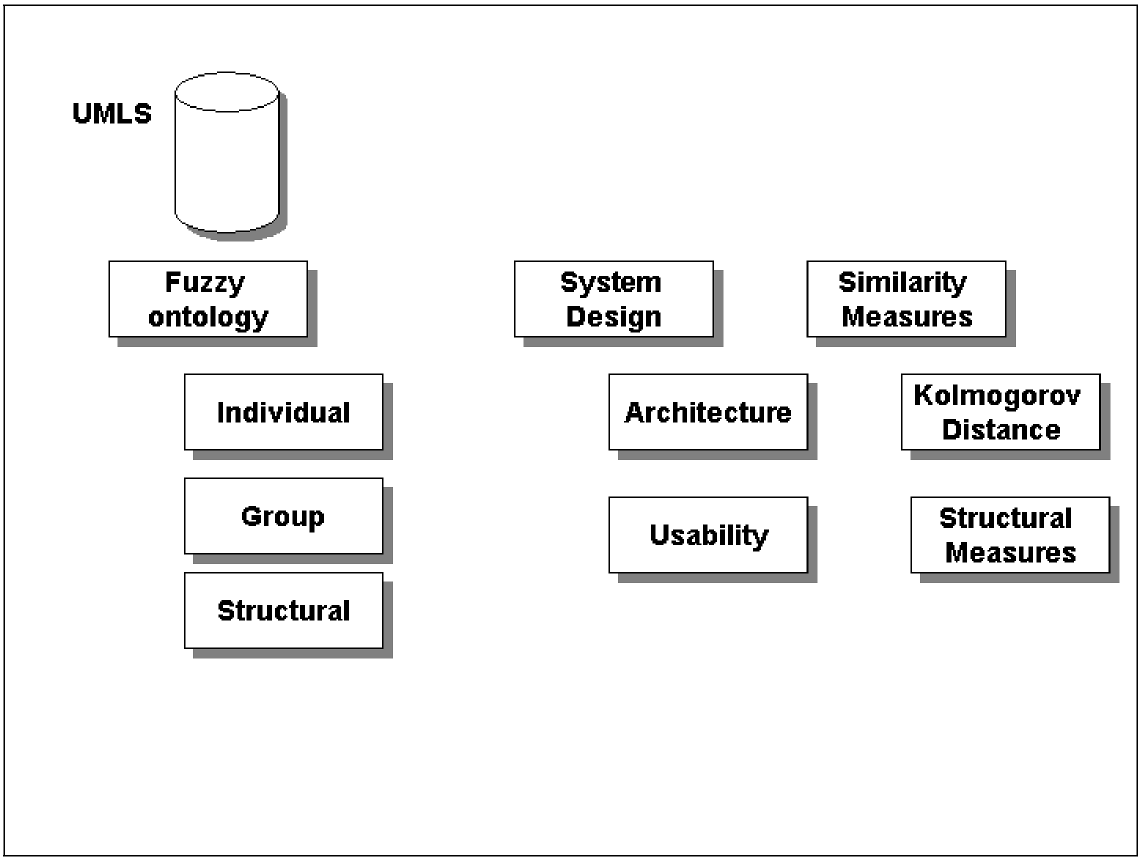


Figure 5: Elements of the work

This thesis is organised into chapters. Chapter 1, the introduction, describes the original work of this thesis, describes some of the essential concepts in the domain of medical information retrieval, and emphasises the importance of intelligent support for searching by clinicians and introduces some hypotheses. These hypotheses are revisited in the final chapter.

Chapter 2 includes a literature review and background to the problem. The review covers some of the basic concepts of information retrieval, including relevance, and links this to the process of querying. The nature of "useful" information is discussed in the context of the medical domain. Systems for organising information and knowledge are examined, including coding schemes and ontologies. The concepts of Fuzzy search, and the semantic web are described.

Chapter 3 covers the concept of a collaborative searching approach and introduces the concept of a fuzzy ontology. Fuzzy ontology is justified in terms of cognitive psychology and the use of fuzzy ontology is explained in practice. A number of learning schemes for fuzzy ontology are proposed.

Chapter 4 deals with the use of information theory, and particularly the Kolmogorov distance measure to characterise and cluster diverse documents. It also covers some of the learning approaches used to characterise structural information. The initial experiments validating Kolmogorov distance measurement for author identification are described here.

Chapter 5 deals with the practicalities of the development of a system (known as fSearch) for discovering useful medical information. This chapter deals with the implementation of fSearch and programming issues associated with it. This includes practical aspects including the use of web services and the modification of an existing ontology.

Chapter 6 includes the majority of simulation and validation work, learning fuzzy ontologies from various sources including medical websites and the Reuters-21578 corpus. The majority of experimental results are described here.

Chapter 7 deals with a case study of the use of fSearch and includes the results of usability testing. The results of the ontology learning from the users are also described here.

Chapter 8 includes the discussion and conclusions along with future work. The hypotheses put forward in chapter 1 are revisited.

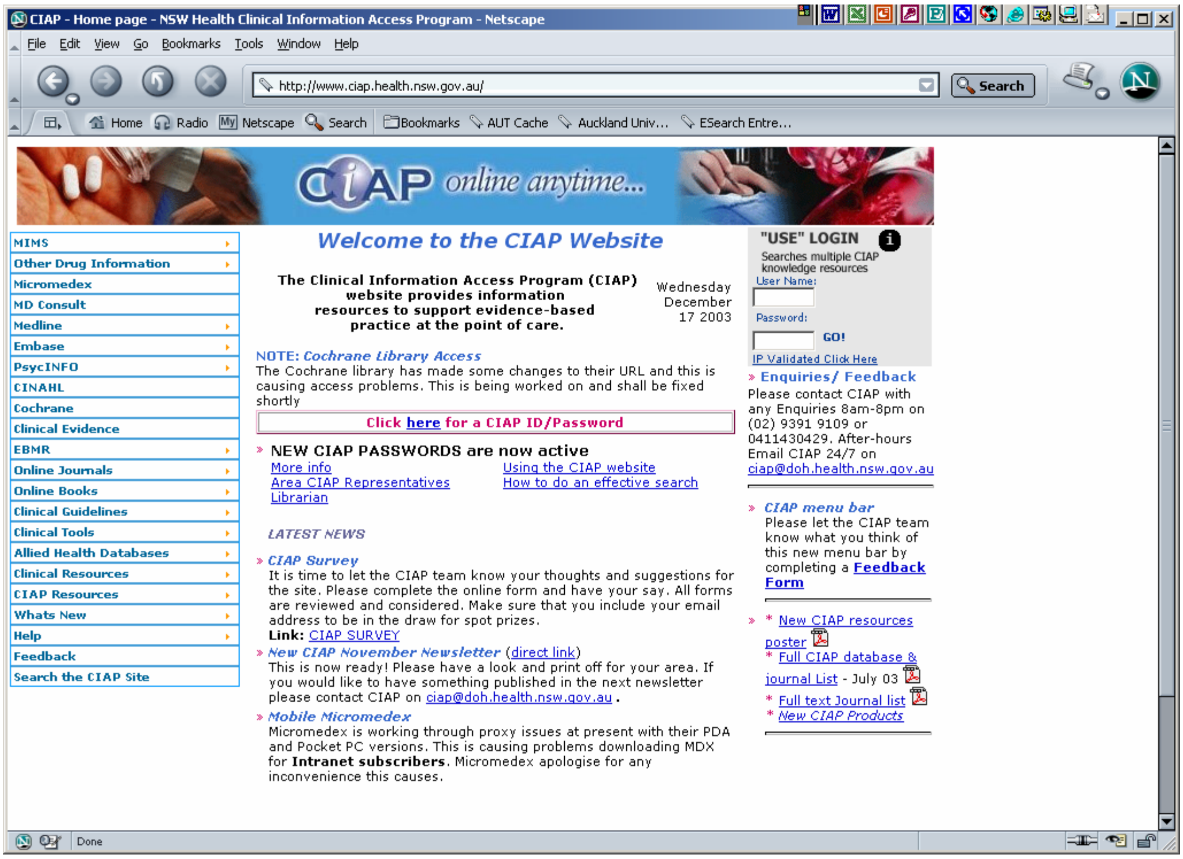
The appendix includes some material related to the terms used in the fuzzy ontology construction in Section 0, 8.8 and 8.9 show some XML structures used in this work, and section 8.10 covers some of the material from the BMJ Corpus. Section 8.11 gives a more detailed description of the fSearch screens, the questionnaire used is described in appendix section 8.13.

Chapter Two

This chapter gives a broad introduction to the literature of information retrieval and identifies relevant work. Section 2.1 introduces some of the concepts and terms used in information retrieval, and section 2.2 discusses metadata. The “Semantic web” and the implications for information retrieval are covered in section 2.6. Section 2.3 describes some techniques for dealing with web-based documents. This chapter also discusses collective intelligence as an emergent property of systems in section 2.4. In particular this chapter deals with the application of fuzzy logic to the problems of information retrieval and the use of ontologies in information retrieval in sections 2.5 and 2.7.

2.1 Information retrieval

Searching for information can be an extremely complex task. Many searches performed are inadequate or cursory as (Hertzberg & Rudner, 1999) show in the context of the ERIC database. Successful searching relies on a high level of cognitive effort, for example by using the techniques of critical reflection (Brem & Boyes, 2000). However there is data to support the idea that users are loath to spend more than 30 minutes learning a system such as a catalogue (Borgman, 1986). Indeed if the concept of just-in-time information retrieval, as an aid to clinical decision making at the point of care is to be realised (Gardner, 1997), then complex time-consuming strategies performed by trained users are not possible. Recent work, looking at the usage of the Clinical Information access programme, (CIAP) in new South Wales (Gosling, Westbrook, & Coiera, 2003), has emphasised cultural barriers to use of online sources of information in a clinical setting, and this includes a perceived lack of skill in information retrieval by clinicians. Interestingly, commonly quoted technical reasons, such as slow access and the ever-present “lack of time” were not found to be of importance. Gratifyingly, recent evidence from analysis of the online logs of the system (Westbrook, Gosling, & Coiera, 2004), appear to show that the system is used primarily for answering questions of clinical importance, however the method used for showing this – largely based on correlation between busy admission periods and usage of the system is open to some question. However it is undoubtedly true that such systems are being used increasing often in clinical environments and that the presence of such systems is being seen as the norm rather than as an exception – if only for the reason that digital libraries are increasingly common tools for all knowledge workers. The CIAP system, described by (Moody & Shanks, 1999), is particularly interesting as a “top-down” approach to providing evidence at the point of care. One of the themes that emerges from the literature is the importance of champions in both the installation and use of such systems. In particular the different uses to which different groups of medical professionals use online tools, and hence the different types of resources required are shown in the work by Gosling (Gosling et al., 2003) and work such as (Wozar & Worona, 2003), which showed that although nurses were enthusiastic uses of online information sources once trained in them, their usage tended to be of patient-focussed materials rather than academic bibliographic systems such as CINHAL or MEDLINE. Having multiple database systems, with many different interfaces and means of searching can only increase the obstacles to effective use of these tools. Even the CIAP system has over 40 different, searchable, databases available, each with its own quirks, not to mention the individual journals, and tools such as Google (see figure 1).



Welcome to the CIAP Website

The Clinical Information Access Program (CIAP) website provides information resources to support evidence-based practice at the point of care.

Wednesday December 17 2003

NOTE: *Cochrane Library Access*
The Cochrane library has made some changes to their URL and this is causing access problems. This is being worked on and shall be fixed shortly

Click here for a CIAP ID/Password

» NEW CIAP PASSWORDS are now active
More info
Area CIAP Representatives
Librarian
Using the CIAP website
How to do an effective search

LATEST NEWS

» *CIAP Survey*
It is time to let the CIAP team know your thoughts and suggestions for the site. Please complete the online form and have your say. All forms are reviewed and considered. Make sure that you include your email address to be in the draw for spot prizes.
Link: CIAP SURVEY

» *New CIAP November Newsletter* (direct link)
This is now ready! Please have a look and print off for your area. If you would like to have something published in the next newsletter please contact CIAP on ciap@doh.health.nsw.gov.au .

» *Mobile Micromedex*
Micromedex is working through proxy issues at present with their PDA and Pocket PC versions. This is causing problems downloading MDX for **Intranet subscribers**. Micromedex apologise for any inconvenience this causes.

» **Enquiries/ Feedback**
Please contact CIAP with any Enquiries 8am-8pm on (02) 9391 9109 or 0411430429. After-hours Email CIAP 24/7 on ciap@doh.health.nsw.gov.au

» *CIAP menu bar*
Please let the CIAP team know what you think of this new menu bar by completing a **Feedback Form**

» * New CIAP resources poster
* Full CIAP database & journal List - July 03
* Full text Journal list
* *New CIAP Products*

Figure 6: The CIAP screen - note the large numbers of databases and information sources

Log file analysis of search engines has been performed e.g. (Westbrook et al., 2004), (Jansen & Spink, 2005). These approaches provide large amounts of data, but do not really give any information on the intensions of the searcher, whether their search was successful, and of course whether the information they retrieved was accurate and useful. The more recent data (Jansen & Spink, 2005) does show that queries seem to be on a broader range of topics, but use of advanced query construction remains small. As is pointed out in (Broder, 2002), queries to web search engines may not be for direct information need satisfaction, but may also be "Navigational", where a particular site or hub is sought, "Transactional" , where the intention is to perform an activity, or the more traditional "Informational", where information is expected to be found.

In order to build a successful system for information retrieval it is necessary to have an understanding of the process from both the systems, and the users point of view. Classic texts such as (Jardine & van Rijsbergen, 1971) and more modern works such as (Belew) have covered both areas. Logs alone will not provide this information or guidance.

Kagolovsky's (Y. Kagolovsky & J.R. Moehr, 2000) view of the IR process is given in Figure 7. Modern search engines attempt to automate some of the relevance, and query formulation aspects of the process.

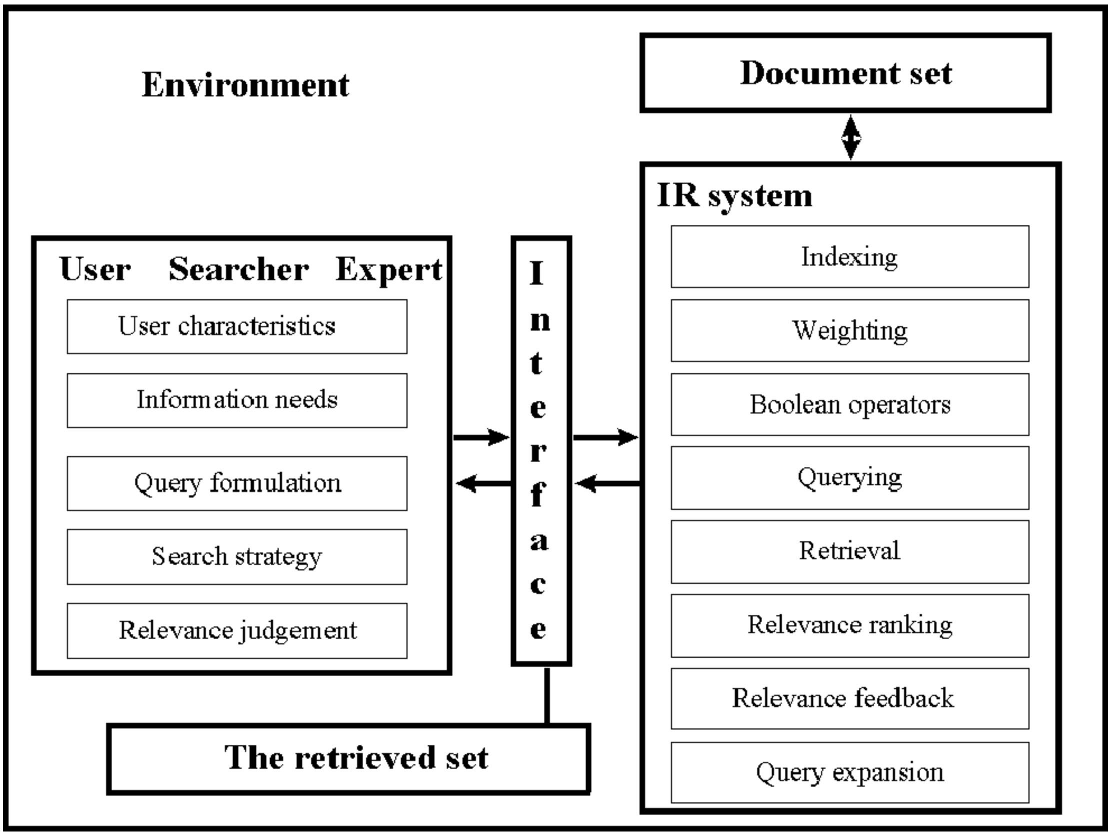


Figure 7: Aspects of the IR Process from (Y. Kagolovsky & J.R. Moehr, 2000)

The large number of attributes assigned to both the user and the system, emphasise the fact that determining the success or otherwise of a particular searching system does not rely entirely on relevance scores. (Jardine & van Rijsbergen, 1971) does give a useful summary of some parameters that may be important for assessing an information retrieval system. These are shown, with comments, in Table 5.

Table 5: Attributes of information sources, adapted from (van Rijsbergen, 1979)

Attribute	Description	Comments
Coverage	The degree to which the collection being searched includes material of use	With multiple information sources being queried, this is now extremely broad.
Time Lag	The time taken for the search to take place	Largely a function of bandwidth available and activity at search sites.
Presentation	The form of the output	The use of HTML and PDF has standardised the format, although not the appearance of returned results.
User Effort	The work expended by the user in obtaining the results	Covers the time spent formulating queries, as well as the effort in thinking about them
Recall	The proportion of relevant material in the collection actually retrieved in answer to a search request	Especially important for review tasks – has every study been found?
Precision	The proportion of retrieved material that is actually relevant	Dealt with in more detail in section **Error! Reference source not found.**.

This problem is not trivial. The last of Swanson's "Postulates of Impotence" (Swanson, 1988) says that "In sum, the first eight postulates imply that consistently effective fully automatic indexing and retrieval is not possible. The conceptual problems of IR – the problems of meaning – are no less profound than thinking or any other form of intelligent behaviour", (page 558). Assuming that he is correct, then the correct role of any system that aims to contribute to IR is one that supports the user, and in particular the cognitive and metacognative processes, rather than trying to replace them.

2.1.1 Relevance

Relevance is a key concept in information retrieval. The complete definition of relevance continues to elude the discipline, but Saracevic gives one as "Communication of knowledge is effective when and if information that is transmitted from one file results in changes in another. Relevance is the measure of these changes"(Saracevic, 1975) page 326. One definition of relevance is the degree to which the retrieved document set matches the user's requirement for information. More sophisticated approaches have used relevance theory (D. Wilson & Sperber, 2002), where an item is relevant if it has some "Positive Cognitive Effect" – that is it matters to the user.

One very important method used in information retrieval is the “vector space” model (Salton, Wong, & Yang, 1975),. In this approach documents are seen as vectors within a space made up of appropriate index terms. A “good” selection of index terms results in each document having a unique vector, as different from the others as possible. This can be achieved by using both index terms and weights associated with such terms. The index terms can be stored in an index, that is one that lists the terms and the documents that contain those terms and a reverse index that contains the document names and the terms each contains. Less common index terms should then be given priority in the search, as they are more likely to distinguish between documents. Relevance scores may be calculated in terms of the vector-space model which uses such a weighting scheme.

These is a very useful discussion of these approaches in (van Rijsbergen, 1979). More recently methods including the likelihood of the query (Bodoff, 2004). As a working definition it seems reasonable to suppose that a relevant document is one that results from a search that in the opinion of the user is related to the topic of the search. However, a useful document must not only be relevant to the search, but the context in which the search was conducted, or be a document that assists in the successful conclusion of the activity that the search was conducted in relation to. As an example of the second case, a useful document may be one that assists navigation by for example showing that a particular page no longer exists. A successful query – that is one that satisfies the requirements of the user may include the retrieval and use of a number of useful documents. The following diagram (Figure 8) may be helpful:

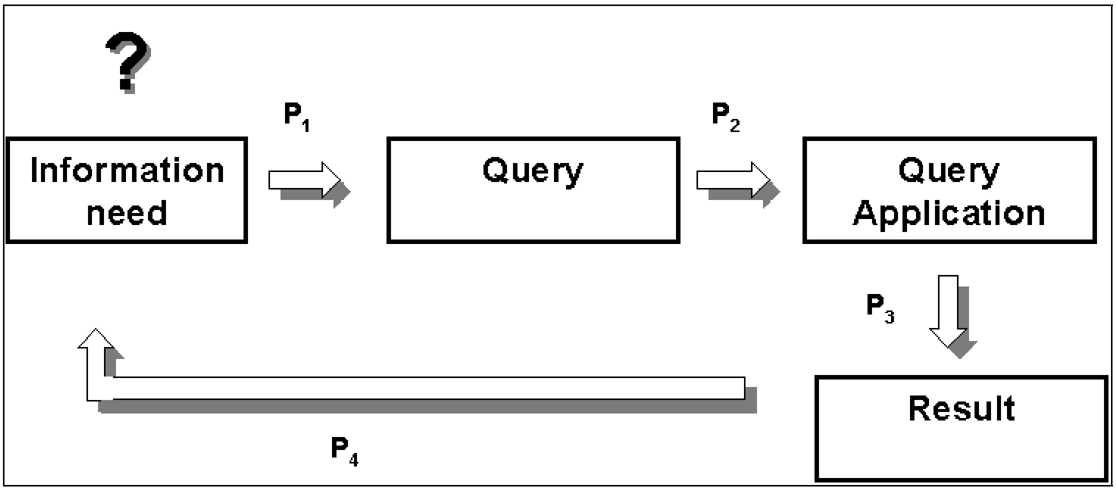


Figure 8: Information need satisfaction

Each process, (P_1..P_4), contributes to the overall efficiency of the search for information, where the efficiency represents the degree to which the information need was satisfied divided by the amount of effort needed to answer the question. It does not reflect the importance of answering the question, nor the novelty of any information discovered.

P_1- The query formulation process.

P_2- The parsing of the query by the searching device

P_3- The database selection and searching process

P_4- Result presentation.

This represents a rather mechanistic and “systems” approach to the problem. Kagolovsky and others have proposed that rather than considering the user as a person with an information need to be satisfied, they may be thought of as being in a “anomalous state of knowledge” (ASK) (Kagolovsky, 2001). This approach is shown diagrammatically in Figure 9.

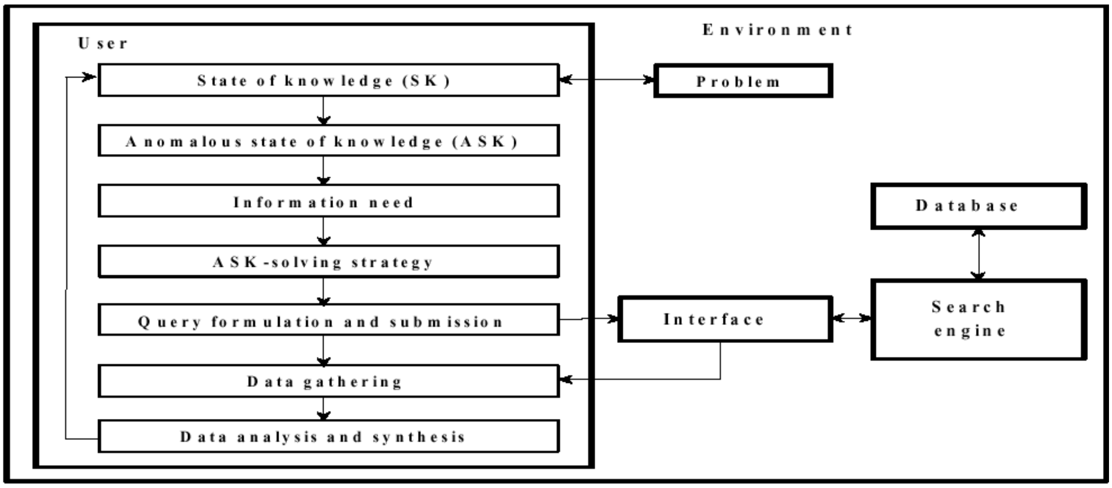


Figure 9: The anomalous state of knowledge in IR from (Kagolovsky, 2001)

This reflects the fact that the extent of the information need may not be clear to the user when they begin their search. This concept is related to Swanson’s notion (see above) that the information need is hard to measure at the start of the process, and only becomes clear at the end, when it has been satisfied. This approach is particularly relevant when considering the actions of patients in searching for information, because of the lack of information as to what is a reliable or authoritative source, or even an understanding of the organisation of the information. Such “personal health information seeking” (Stavri, 2001) may have much more in common with a long-term learning process rather than a single, information deficit.

2.1.2 Query formulation

Query formulation is often particularly difficult for users that do not have training in information retrieval. This is an old problem (Borgman, 1986) that has continued despite improvements in interfaces “computer literacy” (Borgman, 1996). Essentially the message of Borgman is that understanding of the underlying principles of the organization of knowledge databases including for example the use of index terms and Boolean logic is necessary to use them efficiently. However users are reluctant to undergo the necessary training for this, indeed training for longer than half an hour is seen as excessive. In addition, the rapid rate of change of interfaces, and the wide variety of interfaces used have made the users job harder. Problems that continue to occur include:

- Choice of index term
- Expansion of terms
- Combination of terms

Of particular interest are systems like Inquirus2 (Glover, Lawrence, Birmingham, & Giles, 1999; Glover, Lawrence, Gordon, Birmingham, & Giles, 2001), that are designed to answer particular clinical queries rather than act as a general- purpose search tool. The drawback of these systems is the degree to which constant editing and updating is required.

Another approach uses the traditional IR approach of allowing complex Boolean statements to be entered into the system. However, examination of the log files of search engines reveals that the use of “advanced” search tools comprises less than 10% of all queries (Spink, Jansen, Wolfram, & Saracevic, 2002), and some of these are thought to be errors. Indeed most queries are made using one or two words with an implicit “AND”. (Wolfram, Spink, Jansen, & Saracevic, 2001) has found that the mean number of words in each query did not change between 1997 and 1999 and remains at 2.4. These studies dealt with log files from European users of the “all the web” general-purpose search engine. Some systems such as “Ask Jeeves” and GOOSE (Liu, Lieberman, & Selker, 2002) attempt to parse natural language queries and construct queries from them in order to avoid the distaste users appear to have for query construction. GOOSE is interesting because it attempts to build a model of what the user ‘means’ – by using a repository of ‘commonsense’ as part of its parsing. Interesting work continues at the MIT media lab in the Commonsense Computing project in trying to elicit and represent commonsense from both text documents and visitors to their “Open Mind” website located at (http://openmind.media.mit.edu/) Other methods use quite simple parsing for the identification of potential keywords and in the case of “Ask Jeeves”, by attempting to match the user’s query to a library of preformulated questions.

However, the argument that users are to blame in effect for the perceived poor performance of search engines and that better education would improve this appears to be contradicted by the results of the work of (Eastman & Jansen, 2003). These authors took a selection of queries that had been logged in the Excite search engine that used Boolean operators. They then ran these queries on a number of search engines (AOL, MSN and Google), with or without the Boolean operators included. They then examined the retrieved documents for relevance and ranking. Somewhat surprisingly the results showed little difference between queries using Boolean linkages, and those that did not. Before discarding the use of information retrieval training however, it should be remembered that search engines often parse queries before executing them – for example Google may include implicit “AND” statements. Crucially for this work the authors acknowledge that this work has been done on general purpose search engines, with quite commonly used, and certainly non-technical, terms. Obviously any successful search engine in general use will be optimised to satisfy likely queries, and may indeed use precalculated results for queries that it judges to be similar to the current one, rather than performing a Boolean operation at all. The case of technical terms, and rarely used query terms may be very different, and sources of information other than search engines – such as PubMed - still use Boolean approaches.

MEDLINE currently contains various filters for terms such as ‘clinical trial’ and ‘review’. These filters have become increasingly sophisticated and a study published recently explains some of the background to this process (Wilczynski & Haynes, 2004). In order to assess the best way of returning results related to the concept of ‘prognosis’, the authors examined large numbers of journal articles by hand. Articles were then rated for quality and relevance to the goal concept. Searches were then performed on MEDLINE and the specificity, sensitivity, accuracy and precision calculated. The combination of search terms that were required to get high, equal levels of sensitivity and specificity were derived from candidate terms and combinations of terms obtained from experienced MEDLINE users. The aim of this work was to allow a search filter to be constructed, using search terms, that would identify methodologically high quality articles to be retrieved. In practice the filters should be used in conjunction with other, clinically relevant terms in order to obtain better accuracy. A similar article was published in the (Haynes & Wilczynski, 2004), outlining the changing precision, sensitivities and specificities of searches, from MEDLINE using various terms. The information in table 5 is derived from this information.

If U are the articles in the database, and R are those retrieved by the chosen search strategy where U_{high} are high quality articles and U_{low} are low quality articles and R_{high}, and R_{low} are corresponding classes for the retrieved articles then – the following table describes the values calculated.

Table 6: Derivation of terms regarding effectiveness of searching

Measure	**Meaning**	**Calculation**
Sensitivity	Proportion of high quality articles retrieved	$=N(R_{high})/N(U_{high})$
Specificity	Proportion of low quality articles not retrieved	$=(N\ (U_{low})-N(R_{low}))/N(U_{low})$
Precision	Proportion of retrieved articles of high quality	$= N(R_{high})/N(R)$

However, although sensitivities and specificities of particular searches can go into the 80%+ range, the low precision of such searches (around 10%) may present the greatest barrier to clinical usage at the point of care. The reason for this is neatly explained in the article by (Bachmann, Coray, Estermann, & ter Riet, 2002), which introduces the concept of "number needed to read" (NNR), in analogy to the concept of "number needed to treat" (NNT). NNT is a familiar concept from EBM and refers to the number of patients you would need to treat with a particular intervention in order to gain a certain benefit for one patient. It is used particularly in the evaluation of preventative or screening medical interventions (Altman & Andersen, 1999). The NNR is calculated as 1/precision, and calculates the number of articles or abstracts that would be read in order to obtain one good one. For 10% precision this number is 10. This raises particular problems for clinical users of information systems. The typical "first page" presentation of abstracts in a system such as MEDLINE will have between 1-3 abstracts that are of high quality, even when a filtering system is in place, with reasonable sensitivity. Bachmann's work emphasise that with less sensitive searching, the precision rises, but the important point is made that such a filtering system may well not be randomly excluding articles, but may do so with a particular bias, which implies that the result set may not reflect the true picture. One can imagine such a situation in particular, where a term begins to gain a more or less specific meaning over time, a strategy based around searching for such a term may emphasise documents produced at the point of the terms most popular use. An example may be the term WEB, where the combination of

WEB+computer may be a particularly useful one before 1994 for discovering computer simulations of spider web constructions, but not so after that date. A similar case may exist with obsolete diagnoses or deprecated terms in medical searching. Medical subject headings (MeSH) index terms are maintained and assigned by the National Library of Medicine in particular for use in PUBMED, and represent a hierarchy of concepts, associated with the medical literature.

One of the interesting aspects of the work of (Wilczynski & Haynes, 2004) is that search techniques based around formal traversing of the MeSH tree may not be as successful as the use of text words combined in a more empirical way. Thus the work done in order to avoid systematic bias by the use of specific search terms (such as the specialist lexicon), while valuable, may become somewhat irrelevant if only applied to keyword-based query expansion and refinement.

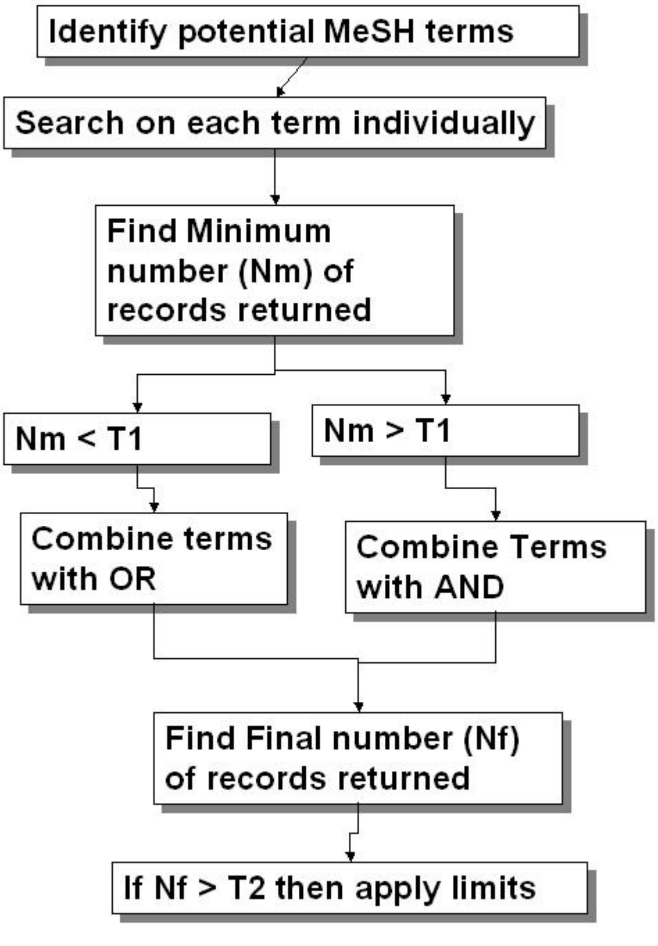

Identify potential MeSH terms
Search on each term individually
Find Minimum number (Nm) of records returned
Nm < T1
Nm > T1
Combine terms with OR
Combine Terms with AND
Find Final number (Nf) of records returned
If Nf > T2 then apply limits

Figure 10: A simple MeSH search

A very simple approach to the use of limits and Boolean combinations is shown in Figure 10. In this case the user has a particular number of query results in mind (represented as T1). The use of OR increases the number of results, while AND reduces the yield. Applying limits at the final stage can reduce the number again. However, the biggest problem with this model, although often taught, is that there is no real certainty about what the value of T1 should be. When online searching was in its infancy, and especially when search results were priced according to the number of documents retrieved, such an approach was likely to be more useful. There is information available that indicates that the second and subsequent pages of search results are rarely viewed, but as pointed out above, “advanced” searching is uncommon. This leads to the possibility, which is not really surprising, that the number of results returned by a particular query is actually unimportant to the general user of a searching tool, even though the number of results returned for a particular query may be of great interest to search tool designers. In effect, if the user doesn’t examine them, then there may be one or a thousand unreviewed search results, and sophisticated means of query expansion and refinement are of little interest compared to the possibility of a “useful” hit on the first page of results. In addition, most systems do not provide feedback on the likely number of documents that will be retrieved before the query is run so even a considerate user of a system, or one that simply wishes to have the quickest response is unlikely to be able to discover the potential effect of the search strategy they use until they use it.

2.1.3 Quality marks and other approaches

It is undoubtedly true that much of the information available on the Internet is incorrect or misleading. Finding information that is useful is surprisingly difficult – for example the work of (Berland et al., 2001), demonstrated that few of the pages discovered using medically focussed search engines contained useful information. In addition they found that many of the sites did not cover all important aspects of a condition, and that the reading age demanded for comprehension was too high for many potential users of the material. Their study was conducted in English and Spanish and included navigating down from the page found by a search engine to investigate the comprehensiveness of the information. It can be argued that some material is not intended for general use, however within the context of the web there are fewer barriers to accessing inappropriate material, for example cover appearance or location in a specialist library that occur in paper documents. Writing text that is both precise and accessible to people with limited reading skills is difficult. Simple lack of awareness of the problem, or lack of time may be important in reducing the amount of material at a suitable reading age. An analogy may be drawn with disabled access to websites, which is still problematic, despite tools being available to analyse compliance with disability access standards.

Neilson's guidelines for writing on the web include writing in a simple and direct manner, but he still rates "Content not written for the Web" (Neilson, 2005) as one of the top 10 design mistakes for the Web in 2005.

The problem of coverage seems especially difficult to overcome by search-engine based approaches as a query-based approach cannot generally tell you what you don't know you don't know – as Donald Rumsfeld put it:

> **"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know."**

Cited by (Plain English Campaign, 2003) in their "Foot in mouth awards".

Unreliable information is seen as especially dangerous in the medical field, early work on this included the paper by (Silberg, Lundberg, & Musacchio, 1997). This has seen the emergence of quality control "kite marks" such as the Health on the net (HON) Code (Health on the Net Foundation, 2003) or the code proposed by Silberg. However such codes do not appear to be particularly useful when comparing information given on websites with definitive information – for example (Hardwick & MacKenzie, 2003) found that in the area of miscarriage there was no correlation between the HON or Silberg score and the quality of information assessed in respect to the RCOG guidelines. However they did discover that the RCOG guidelines rated highly on both scales. This illustrates again the drawback of "general purpose" guidelines – the HON guidelines – shown below - do not concern themselves with completeness, and are biased towards avoiding advertising and "passing-off" rather than the quality or coverage, or usefulness of the information. A more general approach, including evaluating the quality of the links etc. in a website was suggested in (J C Wyatt, 1997) including work by (Silberg et al., 1997).

Table 7: Evaluation of Medical websites based on (J C Wyatt, 1997)

Aspect	Evaluation method
Credibility, conflicts of interest	
Web site owner or sponsor, conflicts of interest	Inspect site (Silberg et al's criteria)
Web site author, credentials	Inspect site (Silberg et al's criteria)
Structure and content of web site	
References to sources	Inspect site (Silberg et al's criteria)
Coverage, accuracy of content material	Inspect site (Silberg et al's criteria; compare with current best evidence)
Currency of content material	Inspect site (Silberg et al's criteria; compare with current best evidence)
Readability of material	Calculate reading age, readability indices (word processor grammar checker)
Quality of links to other sites	Inspect site, judge if appropriate
Media used to communicate material	Inspect site, judge if appropriate
Functions of web site	
Accessibility of site via search engines	Laboratory test with users
Use of site, profile of users	Web server statistics, online questionnaires
Navigation through material	Laboratory test with users
Impact of web site	
Educational impact on users	Laboratory test, field trial
Impact on clinical practice, patient outcome	Laboratory test, field trial

An alternative approach to the identification of sites of high quality has been taken (Eysenbach & Diepgen, 1998a, 1998b). These approaches involve the use of tags and metadata to at least work out what the page is about, as well as allowing users of the data to rate the quality of the information, and disseminate these ratings. This project is currently being funded by the European union -(Mayer et al., 2003), under the name MedCIRCLE. The work involved the construction and use of a description language using XML/RDF - Health Information Disclosure, Description and Evaluation Language (HIDDEL). As with all metadata projects, the coverage in terms of the number of information sources that agree to use it, and the accuracy, in terms of the agreements in definition, will determine success or failure. This work is extremely interesting in terms of its collaborative approach, and the work discussed in this thesis could be integrated into it.

Any system of reliability rating will only be as reliable as those doing the rating. One of the issues that occurs with this approach is whether the raters themselves are reliable. In this work the approach taken has been to try and minimise the impact on others of particularly odd or perverse decisions, but these can still occur.

2.1.4 Browsing and searching

Often these two terms are used interchangeably, however browsing and searching behaviour can be very different. In brief, browsing involves moving between documents, or at least choosing from a list of available documents with some information about each, whereas searching involves the user in explicitly choosing a set of terms or a query to search on – without perhaps being fully aware of alternative, related terms. As is pointed out in (Wollersheim & Rahayu, 2002), the browsing process supports recall, rather than recognition. An analogy would be “browsing” in a library, by actually visiting the bookshelves as opposed to searching in a catalogue. When information retrieval was largely concerned with individual well-planned searches, or the use of catalogues then “search” was the dominant modality. The use of the world wide web, with its dominant “browser” interface, and especially the presence of hypertext, lead to a great deal of discussion as to the best way to continue to include the knowledge built up in previous IR applications that had been largely based around search. This problem was identified as being of particular importance to the WWW as early as (Mackinlay & Zellweger, 1995). The phrase “information serendipity”, introduced in this brief panel paper is particularly important to the understanding of the attractiveness of browsing. In theory, searching is a much more efficient method of finding information, because of the array of techniques that can be used to increase relevance, but it requires skill in term selection and use, along with knowledge of the appropriate terms to use. In contrast, browsing, while inefficient for a query with a known endpoint and correctly selected query terms, has the advantage of allowing the user to refine the query much more naturally, and in the case of hyperlinks, the user may be able to benefit from the knowledge of authors about the wider field rather than just the document that they themselves have written. In addition, the process of browsing may be more likely to reveal unexpectedly relevant information, or even the lack of information about a particular topic. A browsing process may be much more of a journey, or conversation rather than a simple question and answer session.

This thesis is concerned primarily with a searching process, with some constrained browsing processes – for example following links, or picking potential query terms from a list.

However, unconstrained and unassisted browsing can be extremely frustrating, or even result in useless or misleading information being found – because of the lack of an indexing process this may not be due to the information being incorrect, merely being misinterpreted by the user. "Information scent trails" have been introduced (Olston & Chi, 2003) in order to attempt to combine the benefits of search and browsing by using search techniques to highlight particular hyperlinks that may be productive to browse. This technique has been shown to enhance the browsing experience, by allowing users to find information required for a collection of seven tasks related to the Xerox website more quickly using the "Scent Browser" than without. However it does rely on the automatic analysis of the pages being browsed, which would not be possible for a general search of the WWW. It is also dependent on the user being able to formulate goals during the browsing process, and hence can be seen as more of a local guide, than a large-scale one.

This approach is similar to that of the Hyperbolic browser (Pirolli, Card, & Wege, 2003) which conceals and reveals information depending on the movement of the mouse over the area displayed. Crucially, this approach allows users to have a "local view" of the area under consideration without the problems of occlusion or loss of context associated with some other methods of information visualisation. These approaches remain general-purpose however, as the arrangement of items is performed by the author of the sites, although tools such as Inxsight's star viewer – see Figure 12- do allow passing of schema to these displays.

1 Authority

Any medical or health advice provided and hosted on this site will only be given by medically trained and qualified professionals unless a clear statement is made that a piece of advice offered is from a non-medically qualified individual or organisation.

2 Complementarity

The information provided on this site is designed to support, not replace, the relationship that exists between a patient/site visitor and his/her existing physician.

3 Confidentiality

Confidentiality of data relating to individual patients and visitors to a medical/health Web site, including their identity, is respected by this Web site. The Web site owners undertake to honour or exceed the legal requirements of medical/health information privacy that apply in the country and state where the Web site and mirror sites are located.

4 Attribution

Where appropriate, information contained on this site will be supported by clear references to source data and, where possible, have specific HTML links to that data. The date when a clinical page was last modified will be clearly displayed (e.g. at the bottom of the page).

5 Justifiability

Any claims relating to the benefits/performance of a specific treatment, commercial product or service will be supported by appropriate, balanced evidence in the manner outlined above in Principle 4.

6 Transparency of authorship

The designers of this Web site will seek to provide information in the clearest possible manner and provide contact addresses for visitors that seek further information or support. The Webmaster will display his/her E-mail address clearly throughout the Web site.

7 Transparency of sponsorship

Support for this Web site will be clearly identified, including the identities of commercial and non-

Figure 11: The health on the net code

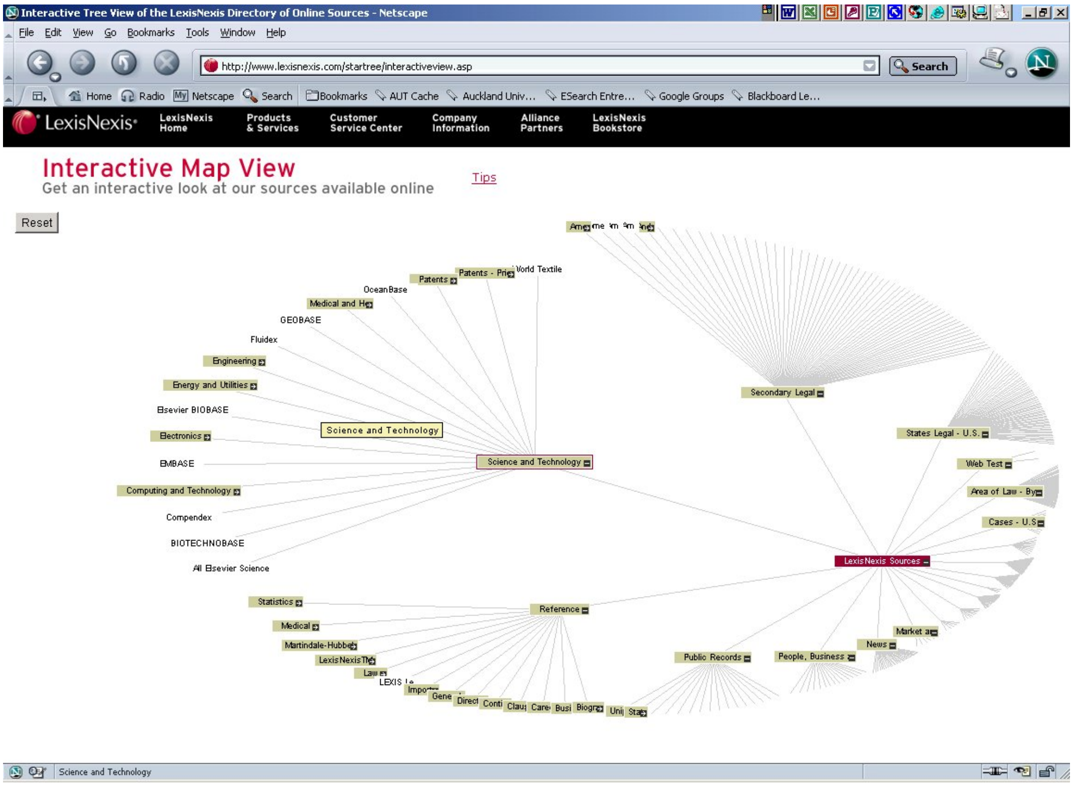
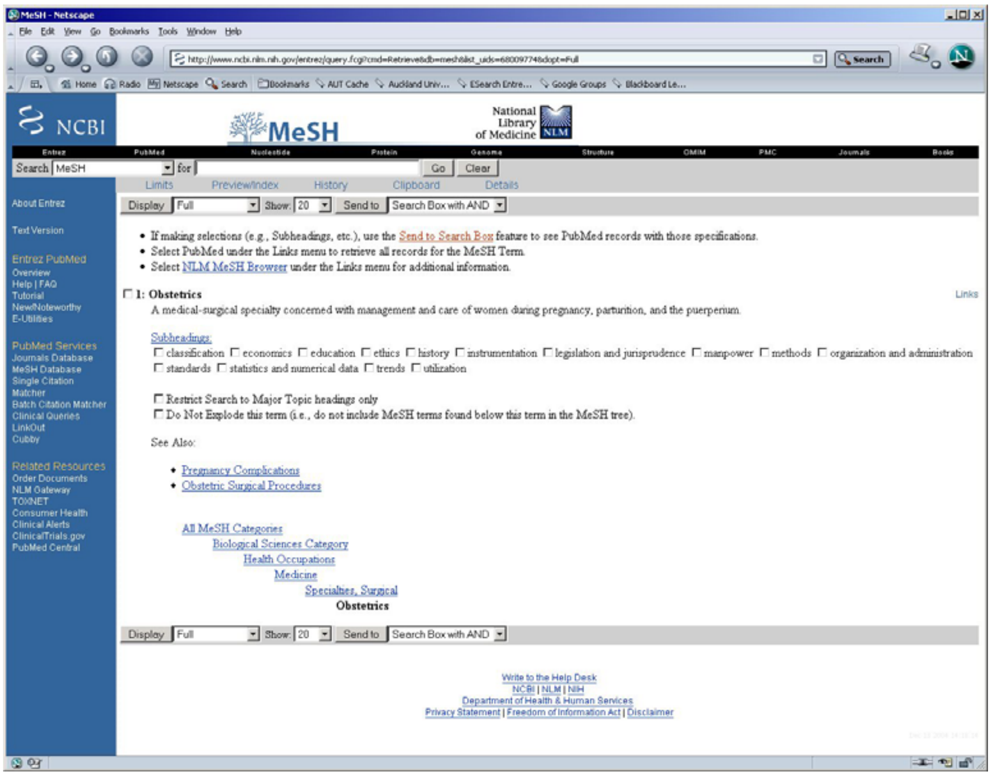


Figure 12: The hyperbolic browser

Undoubtedly the process of retrieving information from the WWW or other information sources is an absorbing task, but work reported by (Agarwal & Karahanna, 2000) suggests that enjoyment and playfulness also add to the level of cognitive absorption experienced by users when using information technology. This aspect has the consequence that if a system is frustrating or annoying to use then the attention of the user is likely to wander and performance will suffer. Another potential issue is that perceived usefulness will also increase with enjoyment. Since most clinical staff in particular are short of time generally, then a system that is difficult to use, or one that produces vast amounts of irrelevant information is likely to cause frustration, and this may be reflected in reports about the system – for example that it is slow – which reflect the frustration and loss of focus rather than merely the time elapsed in use. In the clinical environment, frustration can be particularly disruptive to the whole process of care.

2.1.5 Query expansion and refinement

Traditionally in information retrieval, query expansion relates to an increase in the results returned – often by adding similar terms with ‘OR’ statements, and refinement to a limitation in the returned document set, by use of ‘AND’, ‘XOR’ or ‘NOT’. Query refinement and expansion may use previously unknown terms that the author discovers, during the process of searching. It would be interesting to examine the effects of the approaches of the ACM in their digital library, which assigns index terms from a limited set, with a strict hierarchy, to that of the IEEE in the Xplore digital library which allows the authors to identify keywords. A first search in the IEEE library is more likely to recover documents, because of the wide range of keywords, but a consequent search using these keywords may fail to recover any new documents. In the ACM approach – also shared by MeSH, use of the index term hierarchy, will certainly recover documents, although the hierarchy itself may be obscure to users. To overcome such problems, the PubMed system provides a MeSH browser, based on parts of the UMLS system to identify preferred spellings and terms from a search-engine like interface. The MeSH browser allows the user not only to identify the preferred index term but also to view the definition and see its location(s) within the MeSH hierarchy. The system also allows terms to be exported from this process into a search of the MeSH Database, although the user interface leaves much to be desired (see Figure 13), with multiple data entry boxes. Limits are also provided in the PubMed system, which can be thought of as meta-keywords, which refer to such things as whether the research was performed on humans or the age groups involved. These limits are an inspiration for part of the framework identified in section 1.4.1, but it should again be emphasised that the effect of applying limits is not always particularly clear to the user in advance.


MeSH - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=mesh&list_uids=68009774&dopt=Full
Search
Home Radio Netscape Search Bookmarks AUT Cache Auckland Univ... ESearch Entre... Google Groups Blackboard Le...
NCBI
MeSH
National Library of Medicine NLM
Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search MeSH for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
Display Full Show: 20 Send to Search Box with AND
Links
Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

- If making selections (e.g., Subheadings, etc.), use the Send to Search Box feature to see PubMed records with those specifications.
- Select PubMed under the Links menu to retrieve all records for the MeSH Term.
- Select NLM MeSH Browser under the Links menu for additional information.

□ **1: Obstetrics**

A medical-surgical specialty concerned with management and care of women during pregnancy, parturition, and the puerperium.

Subheadings:

□ classification □ economics □ education □ ethics □ history □ instrumentation □ legislation and jurisprudence □ manpower □ methods □ organization and administration □ standards □ statistics and numerical data □ trends □ utilization

□ Restrict Search to Major Topic headings only
□ Do Not Explode this term (i.e., do not include MeSH terms found below this term in the MeSH tree).

See Also:

- Pregnancy Complications
- Obstetric Surgical Procedures

All MeSH Categories
Biological Sciences Category
Health Occupations
Medicine
Specialties, Surgical
Obstetrics

Figure 13: The PubMed MeSH database interface

It should be noted that this interface, with its multiple text entry boxes has been somewhat improved recently, however, the relative lack of indication of which database you are currently searching may confuse some users.

2.2 Metadata and ontologies

Metadata is essentially "data about data" – it describes the context and meaning of the data stored. It is important to realise that the term "Metadata" is used in this thesis to encompass more than traditional schema integration information. This can be related to a common understanding of the hierarchy of data, information and knowledge (software engineering textbook). Figure 14 displays an example of how data is converted to knowledge.

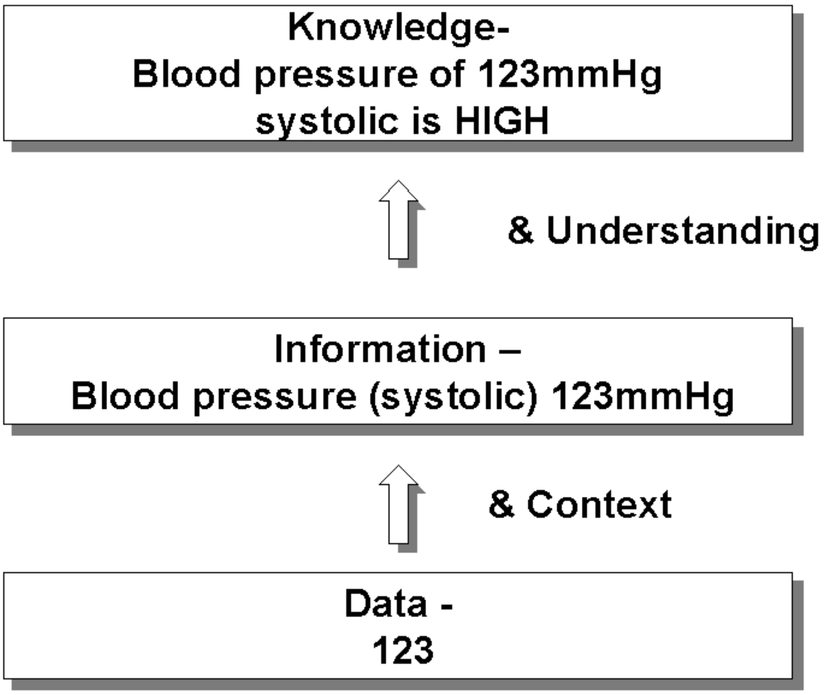
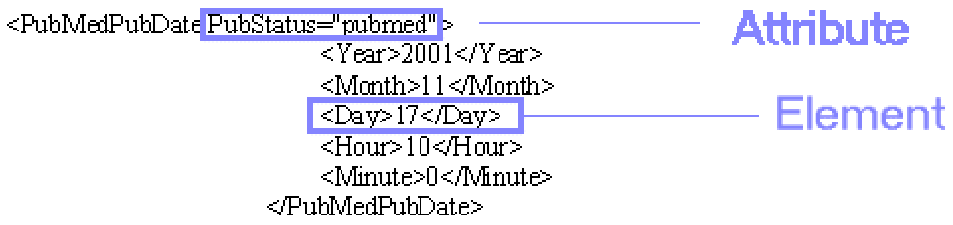


Figure 14: Knowledge from data

When dealing with numerical data as in this example, context can be very difficult to guess – in fact the issue can be made even more difficult, as the intended data format may not even be known, so that the original data could represent a date, a string or a number for example. In the worst case, of a stream of continuous data, the difference between headers, delimiters and the arrangement of data may not be known. This situation is analogous to the problem facing a person listening to a conversation in a language they do not understand, where the listener is not aware what is a question and what is an answer, let alone the content of each. The problem is generally much simpler at this level in the case of text retrieval from WWW-based information sources. Because the pages have to be viewable by a browser, some standard formatting information has to be included, for example compliance with one of the HTML standards, and in the case of HTML documents, the separation of formatting and content is usually obvious because of the use of tags (< >) for enclosing formatting or mark up instructions.

The problem at the understanding level is more difficult for text retrieval however. In schema integration problems, once the type of information has been identified, it is often the case that either the information can be understood in terms of previously available domain –specific rules or that there is a large number of cases with a limited number of fields or attributes. In the case of text retrieval, the attributes may not be known. Extensible mark-up language (XML) is one way of attempting to overcome this problem. When using XML, the author can define their own tags that contain semantic information (W3C, 2000) about the document – literally what it is about, and what each part of the document is concerned with. This is essential for the so-called "semantic web" (Berners-Lee, Hendler, & Lassila, 2001). Somewhat confusingly in the XML standard, an "element" is used to describe a part of the document identified by opening and closing tags and an "attribute" describes an part of the document which is described within a tag. In both cases, the term "attribute" can describe them generically (see Figure 15).



```
<PubMedPubDate PubStatus="pubmed">
                    <Year>2001</Year>
                    <Month>11</Month>
                    <Day>17</Day>
                    <Hour>10</Hour>
                    <Minute>0</Minute>
                </PubMedPubDate>
```

Figure 15: A simple XML fragment

Even with representations such as XML however, it is possible that the attributes assigned in different documents have different meanings, or that these meanings are not clear to the user of the document. In the example above, the date represented is not the date that the document was published, or the date received by the National library of medicine, but the date that the document was published on PubMed. This date is useful for understanding the index terms associated with the document – i.e. any terms not in use before the date of publication would not be used to categorise this document, but it only gives an broad indication of when the article was actually published in the original journal. If one wishes to search PubMed for articles that have appeared since the last time one queried PubMed then this date is the one to use, however if one wants to know what research was published in a particular field since a particular date, then this is not the most useful information. More exotically, this date also assumes use of the western calendar, rather than say an Islamic one.

There are then two different aspects to producing documents with embedded metadata. Firstly that the standard used is known to both the producer and user of the document – the standard used is often explicitly stated in the document. Secondly the meaning of the attributes has to be communicated to the user. In order to make this process efficient it is attractive to build a standard for what attributes are allowed and what they mean. This leads to the concept of ontologies.

2.2.1 Classification and the need for ontologies

Medicine has grappled for many years with the need to have efficient means of communicating data about patients, diseases and treatments. Indeed medical work can be seen as concerned primarily with information communication, storage and transformation. Unambiguous and precise communication of information is vital and a major issue in healthcare (Coiera & Tombs, 1998).

A very precise and often formalised natural language, including a large number of specialised words and compounded terms is taught to clinicians in training. This allows precise description of anatomical features, symptoms etc. and also the construction of new descriptors by combination of existing ones. Although this process is dynamic, for example the term "Gestational proteinuria and hypertension" has recently replaced "Pre-Eclampsia" it does not usually cause difficulties in communication. However, the general tendency as in many fields is to avoid ambiguity rather than encourage standardisation. Thus, a great many synonyms are used and often the preferred term is different in different specialities or countries. This issue makes data processing related to medicine difficult, and for that reason a large number of coding systems have been proposed and a number are currently in use. Ultimately, any coding system is based around an ontology – although this may not be made explicit to the user of the system.

2.3 Document Analysis

In order to classify documents, then elements of the documents must be identified. Traditional SQL-based systems may have a keyword or index based system but systems which rely on unindexed collections must use other methods to analyse the documents before classifying or clustering them.

Documents can be analysed in many ways, and in this thesis the main approaches can be represented as:

- Structural, concerning the construction of the document.
- Syntactic, which involves how the text of the document is constructed in terms of sentence construction and word forms.
- Stylistic, in terms of the use of language, for effect – both stylistic and syntactic issues are involved in reading age score calculation.
- Textual, including word frequency and association.
- Semantic, involving the meaning of the text of the document.

However, other approaches include:

- Citation and link analysis

- Semiotic or image processing approaches.

In particular, there are many methods of classifying and searching for images or multimedia which have not been pursued in this thesis. For the rest of this section, the word ‘document’ refers to text based files which include formatting information which may provide links to multimedia elements.

In order to be able to analyse documents any system must first “parse” the document. This process involves extracting relevant components from the document and then taking appropriate action. A very broad overview of many possible approaches to document filtering is given in (Paepcke, Garcia-Molina, Rodriguez-Mula, & Cho, 2000). The authors of this paper describe approaches using the content, structure and user approval of documents. Essentially they split the approaches into two- action based filtering based on implicit or explicit feedback and content based approaches including explicit content tagging, document analysis, information context and collection analysis.

When analysing documents it is useful to have a publicly available “standard” of documents. This allows comparison and testing of analysis techniques in the same way that well known datasets such as those held at the university of California Irvine (Newman, Hettich, Blake, & Merz, 1998) can be used for the evaluation of machine learning tools. Traditional text corpora have been available for many years containing extremely large numbers of documents and transcriptions of speech – for example the British National Corpus (Oxford University Computing Services, 2001) which contains over 100 Million words. Work has also taken place to build corpora for specifically medical documents, classified into genres (Zweigenbaum, Jacquemart, Grabar, & Habert, 2001). The advantage of this sort of work is that it identifies candidate classifications for new documents, and also gives an insight into what an ontology that represents a group’s view of a set of documents would look like.

Action-based filtering is also used in the work described in (Paepcke et al., 2000). Both explicit and implicit judgements are used and the reasons for this approach are given below. One of the most important aspects of this work is the recognition that the sharing of judgements between users is vital in order to reduce the workload of making judgements to manageable proportions. This approach also includes the use of action-based filtering in order to identify candidates for content-based filtering. For example a particular user or view may prefer documents of type “Patient information sheet” in some circumstances, and may be interested to retrieve documents that can be classified like this. The aspects of the document that identify it as a “Patient Information Sheet” may include length, reading age and number of illustrations for example. Creating a set of genres, or reusing ones already created allow meaningful information about the type of document to be attached to the document description, and may be included in the query construction process.

Many other web page classification techniques utilize link – following methods to identify communities and associated pages, for example PageRank(Page, Brin, Motwani, & Winograd, 1999).

PageRank works by using in-links (hyperlinks from other documents that refer to the target document). The more in-links a document has, the more highly rated it is, and in-links from documents that are themselves highly rated are assigned a higher weight. Relevance scores are calculated based on a vector space model modified by a weighting allowing from both the rating of the document and the occurrence of query terms in the "link anchor" text that is displayed as the clickable link in the referring document. However, in terms of a system that is not a search engine, that is one that does not attempt to represent and perhaps store an entire domain, such an approach has large computational drawbacks. Obviously, in order to use the link information, the nature of the destination of the links needs to be investigated. At the very least such an system will need to download the documents being linked to perform some sort of analysis. Much more demanding is the prospect of downloading the referring page, as without reasonable coverage of an area in terms of analysis there is no way of ensuring that the referring pages that you do have access to represent a reasonable sample of those which can be cited as authorities, for example in (Farahat et al., 2002).

Another issue of concern in medicine is the occurrence of revolutionary or unexpected results. One example of relevance to obstetrics was the publication of results showing that hormone replacement therapy was not as beneficial for preventing cardiac disease as previously thought. In a recent study (Thunell, Milsom, Schmidt, & Mattsson, 2006), have shown that a number of key publications have changed specialist's attitudes, but the web still contains a mass of contradictory advice, even sometimes on the same site – see http://www.menopause-online.com/benefits.htm versus http://www.menopause-online.com/update.htm .

2.3.1 Classification and Clustering

Classification essentially involves constructing rules that can be applied to a set of attribute-value pairs so that some or all of them can be placed into different group. A general description of the steps involved in learning classification rules is given by. If O is a set of data objects, each o ∈ O can be labelled with a class value. This classification is performed by having classes that are described by a set of feature attributes, which can be formalised as rules. These rules F are combinations of values of feature attributes such that $F \rightarrow C$ where C is the class value, so that the set of all objects matching F is contained in the set of objects matching C. In some cases, where only one rule is used to classify the objects, and it is complete, then the set of all objects matching F is the same as the set of all objects matching C. For learning classifications by machine, it is usual to expose the system to a set of training examples, where the classification is revealed somehow to the system, and then test its ability against a training set of examples where the classification is known. The process is outlined in Figure 16

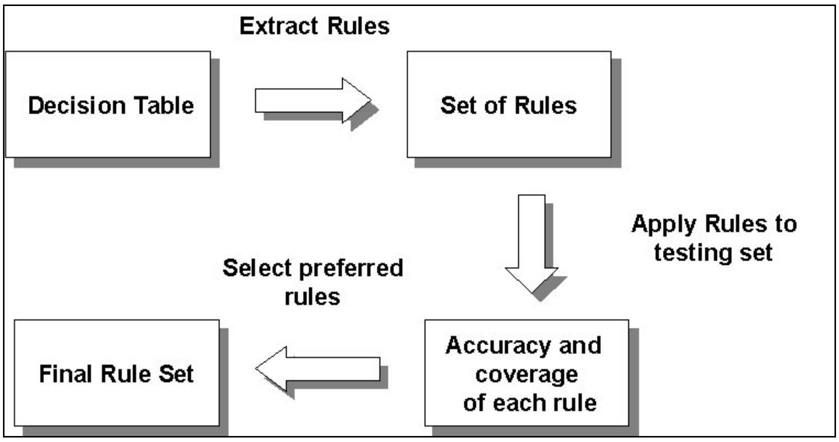


Figure 16: Construction of a set of rules from structured data

Quinlan (R. J. Quinlan, 1991) points out that machine learning techniques produce a subset of all the rules that could be produced. Most of these schemes attempt to produce a set of rules that correctly classifies a testing set of objects - where correctly in this context means the same as has happened in the past. There are a number of measures that can indicate the quality of these rules but they can broadly be described as -

- **Number of Examples classified (coverage) -**

A rule that classifies large numbers of examples is better than one that only classifies a few, and should lead to fewer rules being needed.

- **Accuracy of the rules -**

How many objects are incorrectly classified by the rules?

- **Explanatory capacity of the rules -**

The epistemological value of the rule - that is whether the rule is meaningful to those that may wish to apply it, as well as its relation to the theory of the domain. It should be noted that a rule that directly contradicts a model of the system being studied might be especially valuable, as it may cause changes in the accepted model. It is obviously hard to objectively quantify this aspect of a rule set.

- **Compactness or simplicity -**

For rules that are to be applied to large or complex datasets, rules that involve the fewest attributes are preferred, because this will make applying the rules a great deal quicker.

There can be conflict between these aims – machine-learning systems often have the disadvantage that they produce rules with very poor explanatory capacity because they cannot incorporate knowledge of the domain. Of course the selection of attributes, and the way their values are measured can make a large difference to the rules that will be learned. Simply put, if a data item is not included in the training set or there are no objects with a particular value it may be that the system cannot learn a valuable rule. Accuracy and coverage will also tend to conflict. Xiang (Xiang, Wong, & Cercone, 1995) attempted to resolve this by using measures that define acceptable levels of both.

As large databases become more common, it is increasingly important to discover the knowledge that is contained within them. Traditional statistical methods are difficult to apply when there may not be a hypothesis to test, and the relationship between the data items is unclear. Indeed the multiple application of statistical tests will be expected to discover some significance at the 95% confidence level in one of every 20 tests performed even if there is no actual difference. This sort of ad-hoc multiple testing has been called "Data torture". "Data mining" is the systematic use of techniques to analyse these large database to find the nuggets of knowledge that may be contained within them.

The aim of many KDD systems that perform classification is to be able to classify objects within a database, using the attributes recorded in that database, which would be similarly classified using other attributes that are not contained in that database. However if the attributes in the database are the same as those used for the original classification, then it seems possible that rules learned from a database using KDD will be the same as those used in the expert classification. At the same time, if there are rules learned from the database that appear to classify test examples correctly but involve attributes that are not declared by the experts then this allows for the discovery of unexpected or interesting rules. These rules would help to explain otherwise inexplicable decisions. It is important to realise that the act of selecting the attributes to record in a database involves a great deal of expert knowledge, and this applies to information tables, or any input used for a KDD system. The usefulness of any classification scheme is vitally dependent on the proper selection of attributes and classifications - KDD tools can only work within these limitations. This is widely recognised in commercial applications where “Data Warehousing” the act of selecting, cleaning and combination of data into a suitable repository is an essential precursor to the use of KDD tools (Chen, Han, & Yu, 1996). The process is shown in Figure 17.

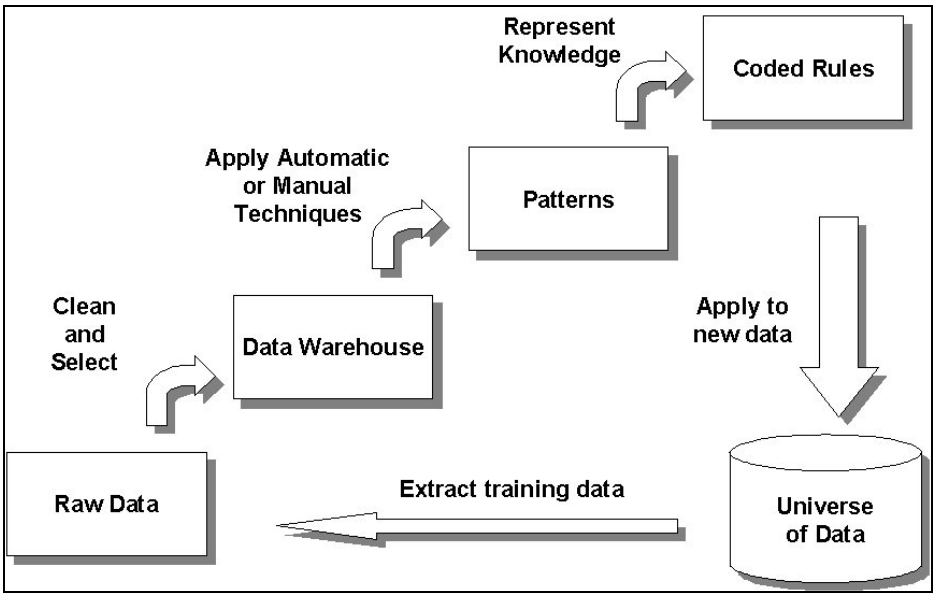


Figure 17: The process of KDD

The field of KDD gives some useful guidance as to the correct preparation and selection of data, the choice of knowledge discovery tool, and the interpretation of the results, an overview is given by (Fayyad, 1998).

Whatever the method of obtaining or arranging information, at some point if an information retrieval system is to get any better then it needs to learn from past experience. One way of doing this of course is to make sure the user continues to put effort into using the system efficiently by means of usable interfaces and training. However it is very attractive to try to get the system itself to learn in one way or another, machine learning, and this process can happen either before the user uses it via means of training corpora which can be done via offline learning – or in response to user queries and decisions about the results which are generally online learning processes. Learning systems operate in a number of different ways, and may be used in a number of different roles in any system to support information retrieval.

Kasabov in (Kasabov, 2002) page 14, describes a classification scheme of different connectionist learning methods. By using some of his criteria a description of a learning system for information retrieval can be formed.

Unsupervised learning is a form of machine learning that does not require classifications to be known in advance. One of the earliest forms of this is the “Self Organising Map” (SOM), which were applied to information retrieval from an early stage (Ritter & Kohonen, 1989). SOM’s are described in more detail in section 6.4.2, but in brief, they allow categories to be developed by a learning process, the user selects the number of categories they want, and the SOM (using a number of potential methods) adjusts the input weights of the categorical number of neurons so that each input is categorised by one neuron. After the network has been trained, the network can be tested with a group of unseen cases. In particular, this paper postulates the relation between a self-organising map, as a spatial arrangement of neurons, where objects that are similar are categorised by the degree to which they fire a particular set of neurons, to the situation in the brain, where a similar spatial arrangement applies.

There are some broad categories of knowledge discovery (Fayyad & Piatetsky-Shapiro, 1996); classification, regression, clustering, summarisation, dependency modelling and change and deviation.

In their very extensive review of data clustering techniques (Jain, Murty, & Flynn, 1999), the authors assert that clustering is the unsupervised classification of patterns into groups. As they point out, the enormous numbers of clustering algorithms available can make selection of the appropriate tool very difficult. Importantly, the choice of clustering algorithm depends not only on its performance on standard, test data sets, but crucially on the nature of the data being clustered, and the nature of the demand for the cluster. Three main approaches to assessing the validity of clustering techniques have been identified (Halkidi, Batistakis, & Vazirgiannis, 2002 -a, 2002 -b):

- External validity tests include comparisons of the clusters formed with a known existing structure that is appropriate for that data. An example of this is the situation where a testing set exists such that each object being clustered has a classification that maps somehow to a preferred cluster identified in advance. Because clustering involves unsupervised learning such a classification is not used in the clustering process.
- Internal validity involves the calculation of parameters that are present in the dataset. Examples include the Cophenetic Correlation Coefficient and the purity measures – purity measures are described in section 4.5.1. Other measures are available and some of those specifically designed for information retrieval were discussed in section 2.1.2

- Relative criteria aim to discover the best clustering scheme that can be produced on a particular dataset. This may include the Hubert statistic, scattering, entropy or other particular measures. These measures can be used as part of the scheme itself in order to decide whether new clusters are formed, or control the expansion or merging of existing clusters.

Clustering differs from classification in that in a clustering process there may not be any classes to discover, and the clusters may have no more meaning than that they are clusters – i.e. there is no a priori cause. There arises the issue of what, exactly, the clusters mean and this is often more problematic. One way to confirm that the clusters have meaning is to use the approach described below of having a training and a testing set of examples that can be used.

Extraction of classification rules from data can be done using two main methods. Classification rules can be directly generated by machine learning algorithms – such as C4.5 (J. R. Quinlan, 1993), or they can be inferred from characteristics of the data – for example from clustering techniques such as K-means. Somewhat confusingly, methods such as hierarchical clustering use techniques that are very similar to rule classification methods in order to produce their clusters. When dealing with new datasets there is a tendency for researchers to use the techniques that they are most familiar with. However tools such as WEKA and Neucom and even packages such as SPSS allow multiple methods of analysis to be performed on the same data, without conversion between formats. This can allow pilot analyses of the data, and candidate attributes that are relevant for classification and clustering to be determined.

The aim of any clustering technique is to identify objects that are similar to one another, without the need for pre-existing classification. In the context of an information retrieval system, this may be used internally as part of a search engine, - to identify for example “more like this”, or for identifying common areas.

Such an approach goes back to early work in IR for example (Jardine & van Rijsbergen, 1971). Systems also exist that cluster documents in an explicit way for the user to see- for example Grouper (Oren Zamir & Etzioni, 1999). These approaches use a number of different clustering tools, for example k-means or hierarchical clustering or variants of them, for example the partitioning method described in (Stienbach, Karypis, & Kumar, 2000). Most clustering methods in IR use a word-frequency vector approach, sometimes modified by using the inverse document frequency value of text words in order to identify important words.

A very large survey of web metrics (Dhyani, Ng, & Bhowmick, 2002) mentions similarity measures using (text) content, links and usage, but no structural similarity work is mentioned. Techniques using the structure of documents to identify clusters are rare. However this approach may allow documents that are about the same topic, but have different styles to be identified. HTML or XML documents that have their structure coded by particular tags lend themselves to this sort of analysis. This approach was used in a crude form in the paper from JAMA (Berland et al., 2001), which identified HTML pages with 50% or more of the space devoted to content as 'content' pages, rather than links pages. A method of clustering using structure and text is given by (Doucet & Ahonen-Myka, 2002). Although this approach does not produce such good results on a test corpus as text-based clustering it does appear to be computationally simpler, and offers an orthogonal approach from text-based approaches if these are not sufficient. In particular, this approach may be more applicable when specialist terms are used. In these cases the number of appropriate terms may be quite small, and the frequency of these terms in the corpus quite low.

In the medical domain there are a number of issues that can cause problems with this approach. Firstly there may simply be a lack of coverage of a particular problem, or it may be a variant of a larger problem, so that the useful material is hard to find, because the number of references about the desired subject are small compared to the number of references about a similar one. To explain this it is useful to consider the Zipf Curve Figure 18. Zipf curves use the observation that if the frequency of a term in a corpus is plotted against the rank of the frequency of that term the distribution conforms to a power law (Zipf, 1949). If log(frequency) is plotted against log(rank) then the relation is roughly linear over the centre of the range. In searching processes this is important because it means that there are three regions. Following the work of (Luhn, 1958), there are three regions.

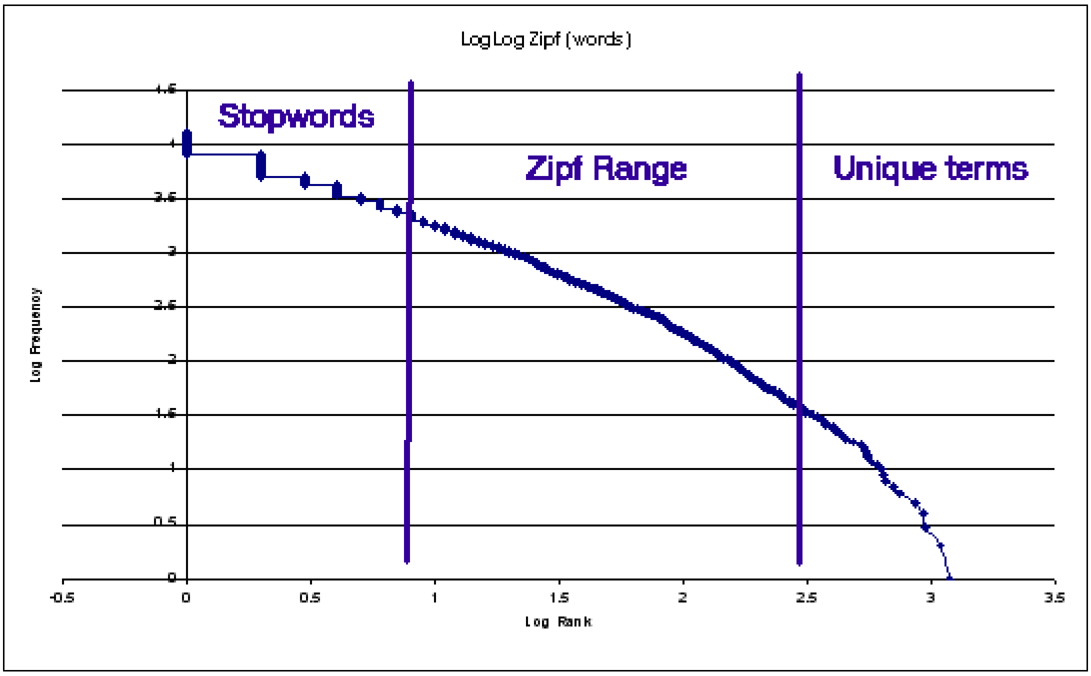


Figure 18: Example of a zipf curve

In the stopword region, the terms are so frequent that they are useless for indexing as every document has them. In the unique terms region, each term occurs in a very small number of documents, so the term may not give appropriate coverage of the concepts requested, as there are probably alternative terms. In the Zipf range, the vector-space model should work well. Work from 1975 (Sparck Jones, 1972), showed that assuming a random distribution of terms in documents following Zipf's law a weighting scheme for index terms based on Zipf's law is efficient. However, this model can still fail to detect the difference between associated but not required terms. In this case, without a great deal of extraneous downloading, the danger of using this approach is that by using incoming references as a form of authority, they are actually only being detected for articles that are part of a series from a particular institution or web ring, and even more potentially catastrophic, the nature of the referral – i.e. whether positive or negative becomes much more important with a relatively small base. Conventional content–based techniques would not distinguish with much resolution between the documents containing the search words, because we are operating at the far end of the Zipf curve. Similarly, such documents are likely to be rarely used, so usage statistics are not helpful – we are effectively imposing the constraint that the type of user AND the topic have to be the same for each instance of searching to learn any useful similarities. Also, the fact that queries may not contain a Zipf–type distribution of terms, complicates the weighting issue.

Link –based approaches have the advantage that they include unvisited sites, but again when a document set has a relatively small number of visits such approaches become less powerful, as the link map is constrained, and may become computationally difficult (and limited by bandwidth) if these destination pages also have to be retrieved.

This work argues that by including existing or future text and link based searching techniques, present in both digital libraries and search engines as the data source for a system, other classification or clustering techniques become useful for discriminating results from these systems. In particular the use of structural and other information based on an individual page is an appropriate and computationally realistic approach in a dynamic and bandwidth limited environment.

The use and comparison of clustering approaches leads to the need for measures of clustering success. Simple mathematical tools such as numbers of items in each cluster, and relative "volume" of each cluster may not be particularly fruitful. Another approach, as is common in information retrieval is the use of already classified objects and the analysis of the distribution of classes over clusters. Such an approach is difficult in this case because the classifications are postulated, rather than confirmed. However, for a test corpus some work has been done to attempt to predict the quality of clusters. In addition the use of information-entropy comparison tools is introduced to attempt to generate similarity measures.

2.4 Collective searching

2.4.1 Collective Intelligence

Collective intelligence is a term that has been applied to collections of intelligent agents that cooperate with each other (Wolpert & Tumer, 2000). In particular, collective intelligence (COIN) agents do not attempt to maximise the utility values for themselves rather than others, rather, the *overall* utility is maximised. Examples are given of for example packet routing problems, where one agent's activities may increase the efficiency of packets guided by that particular agent, but decrease the overall efficiency. A crude analogy in information retrieval may be the thought of a particular agent blocking access to all possibly useful sites in order to gain slightly quicker access. However, the basis of COIN, that independent agents should cooperate without hand crafted rules can be applied to information retrieval.

Quite sophisticated behaviour appears to be generated by relatively unintelligent actors by obeying simple rules, and being able to perform simple and appropriate communication. The importance of “Collective wisdom” has been emphasised in recent popular works such as (Surowiecki, 2004). The use of prediction markets, where numbers of people make bets on the outcome of elections for example, (S.-C. Wang, Yu, Liu, & Li, 2004) is becoming popular. These approaches appear to suggest that large numbers of diverse, interested but not particularly expert people can make more accurate predictions that small numbers of expert individuals, assuming that the right conditions are satisfied. The explanation given is that the effect of biases and missing information are cancelled out by use of larger numbers of people. The search for an appropriate information source can be seen as a prediction task.

By building a system that allows the searching agents to collaborate with each other then their overall efficiency can increase. By minimising the deleterious effect of such cooperation on individuals agent’s activities, and not requiring an overall plan in advance, the overhead of such a scheme is reduced.

The Internet has allowed very many examples of collective decision making, for example simple rating schemes as the reviews found on Amazon.com, or the trust measures implemented on eBay.com. The low cost of publishing and the universal nature of access, or at least the web users indifference to the location of material, means that such approaches are becoming more feasible and more effective with the increasing connectivity of the world population. Obviously malicious or mischievous attacks, in terms of “link spam” or other ways to try and raise ratings on such systems can occur, but there is potential for making such schemes more trusted.

The ease of use of such systems and the perceived worth of them are particularly important, the removal of the US DOD “prediction market” (Caterinicchia, 2003) which included invitations to speculate on the likely mode of attack of terrorist groups shows how unworthy or distasteful systems are unlikely to be accepted. Interestingly, some other potentially disturbing web-based tools such as “longevity” calculators (e.g. (Northwestern Mutual Life Insurance Company, 2004)) have not been withdrawn, possibly because there is less emphasis on the human source of such speculation. Collective intelligence may require a certain degree of anonymisation and obscurity to work effectively in anything but the most philanthropic of cases.

2.4.2 Humans as intelligent agents

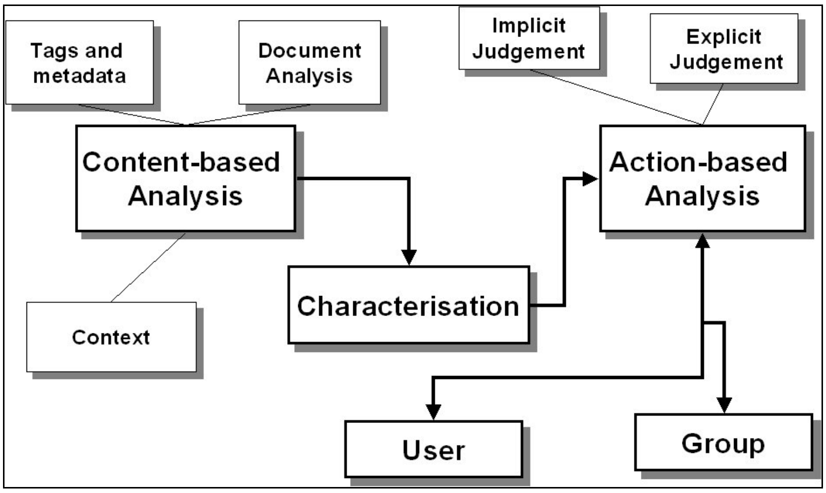


Figure 19: Diagram adapted from (Paepcke et al., 2000).

Figure 19 shows an overall model of the methods by which searching is supported by an intelligent system. The light lines represent the methods described by (Paepcke et al., 2000), and may be termed "traditional" agent based approaches when leading to the content analysis, and traditional preference based systems when leading to action based analysis. Content-based analysis is very common in search engines, for example the PageRank algorithm used by Google (Page et al., 1999). Action based analysis is commonly performed by commercial sites, such as Amazon and digital libraries such as the ACM digital library.

The important difference in this work is the linking of preferences expressed by users or groups both from the contents of the documents selected AND by the structure of those documents – the content based analysis is linked to the data available from the preferences expressed.

The search for information does seem similar to the search for food in some ways, indeed the concept of "satisfying" a query, or "slaking a thirst for knowledge" uses the metaphor of food consumption. Obviously, it appears easier to write software that obeys simple rules rather than directly modelling higher-level cognition, and the thought of using very large numbers of "software agents" that act analogously to insects for example, is especially attractive – especially in the field of information retrieval. This approach seems especially attractive in the context of the WWW, where very large numbers of pages (currently in the billions) need to be examined or at least indexed and parsed if a full link-based system is used. This approach has been the basis of a great deal of interest in "agent technology" that is the use of software agents that act autonomously, but communicate their results to each other, for example the “retriever” system (Fragoudis & Likothanassis, 1999)

.

Another approach has been to “personalise” information retrieval, by recording details of a particular people, for example (Mobasher, Cooley, & Srivastava, 2000). This approach has been very widely used and can include the recording of both the degree of satisfaction with the results obtained, and the methods by which those results were obtained, in terms of search strategy etc. This approach attempts to maximise the value of a relatively small number of searches, by an individual, to create often quite complex, specific rules that are assumed to reflect an individuals’ information needs and preferences. A Zero-input Interface for Leveraging Group Experience (Sharon, Lieberman, & Selker, 2003) describes a system for extending such a preference based-system, around a band of trusted colleagues. Essentially this is an automatic recommender system that allows browsers in the group to share history and preference information. Presumably this could be linked to the use of various measures- such as those found on eBay where users can rate each other for reliability in their transactions, -in order to extend the pool of collaborators who have common interests. However such an approach is likely to require a large number of “transactions” to take place where users explicitly rate each others judgements. This process can of course happen explicitly by communication between users and by sharing of bookmarks files or the construction of “links” pages. The analysis of links and citations is common in the mechanics of search engines and is of course one of the key elements of the “PageRank” algorithm (Page et al., 1999). However there are a number of issues with such systems – in particular the “cold start” problem. This occurs when the system is first used, and there are a large number of searches that have not been previously performed. In this situation, the system does not offer useful recommendations in most cases, but there is an overhead in the use of such systems in order to create the recommendations for future users. Such a situation of “working for nothing” undermines the rationale of mutual beneficence for such systems and has been identified as a major barrier to the adoption of such systems (Middleton, Shadbolt, & Roure, 2004). Middleton’s group describe a system that attempts to overcome this problem, as well as recording the use of their system in practice. Their approach is similar in some ways to that described in this thesis – especially in the aspects of the use of a preconfigured ontology rather than one developed entirely from user behaviour. In this thesis the automatic assignment of fuzzy ontology membership values and the use of an existing ontology are similar. Their feedback mechanism is extremely simple for the user, although the judgements are very general. Other systems have attempted to use zero-input interfaces, (Sharon et al., 2003), to reduce the apparent load on the users. This obviously reduces the effort part of the effort/reward equation, but results from this work, on the usability of the system appear to suggest that users are prepared to put some effort into improving their search results.

However, there is another alternative approach that uses the "collective wisdom" of Internet searchers in order to identify useful sites. By viewing a collection of searchers as a collective intelligence, the following aspects can be identified as important:

- The individual searchers must be autonomous if any improvement is to occur.
- Precise communication between searchers is important to allow the overall system to identify not only particular documents, but also particularly successful searches and criteria for identifying useful documents.
- A memory of some kind is required in the system, in order to identify previous strategies, and the learning process, although this memory may not be the "complete" memory of every action that has taken place.
- The criterion for success should be simple and easily calculated by the agent.
- The differing nature of each agent should be identified and recorded within the system.

One of the greatest incentives for using agent technologies is the prospect of very wide scale searching and in particular very rapid examination of large numbers of potential pages.

Agent based solutions may use approaches such as genetic algorithms or other learning systems, but these suffer from the problem of identifying when they have been successful or not. Effectively every such system relies on similarity measures between retrieved documents and "ideal" documents. This problem cannot be avoided completely, unless every document is viewed by a human, but agent based systems suffer from having to both work autonomously, and hence differently from each other in order to cover the solution space, while at the same time chasing the same targets in terms of final documents. By using humans as the agents, and dividing the potential document space between them, the users can be expected to perform better than pre-programmed systems.

Algorithms such as “PageRank” and others that work using link frequency (Arasu et al., 2001) can be thought of as being based on the actions of authors of web pages. This approach has the advantage that it can be analysed off-line, with no need for interaction with the authors themselves. However even though this system is the basis for the very successful “Google” search engine, it still relies on the opinions of producers, rather than consumers of information. At the simplest level such an approach will tend to reinforce the effect of “small world” networks (Bjorneborn, 2001). This may be desired at the refinement stage of a query, but at the beginning it is possible that the user may be “trapped” in such a small world, especially one based on terms that are familiar to experts in the field, but may not be useful for others. A similar difficulty may occur in such systems as those described in (H. D. White, Lin, Buzydlowski, & Chen, 2004), where the users are invited to identify “significant” literature. This will tend to reinforce the tendency for isolated networks to occur – based on the ‘conventional wisdom’. This may be desirable for someone wishing to study a particular domain, to check that they have not excluded significant authors – for example a postgraduate student! However this approach tends to be restrictive and unhelpful for those who are not sure that such a domain exists – just as a keyword or index-term based system relies on a user making an informed choice as to the keywords they will find helpful. For example a search on the term “nirvana”, mostly obtains links related to the band’s small world rather than a religious one. An outline of some of the roles traditionally associated with documents is shown in Figure 20.

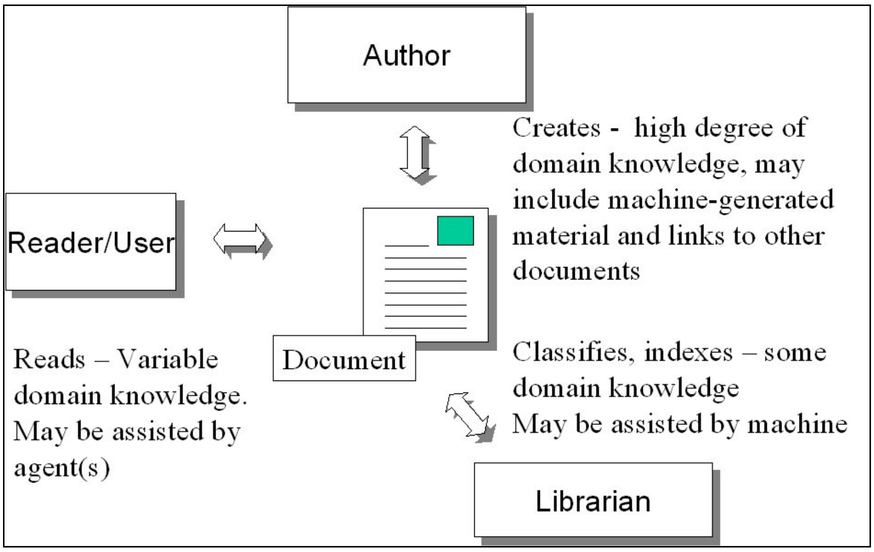

Author
Creates - high degree of domain knowledge, may include machine-generated material and links to other documents
Reader/User
Document
Reads – Variable domain knowledge. May be assisted by agent(s)
Classifies, indexes – some domain knowledge
May be assisted by machine
Librarian

Figure 20: Roles in document production, storage and use.

It should be noted that automatic systems can support all three major roles, and that individuals may in fact be members of two or more groups at the same time. Each group has different skills and views of the process. In this thesis I concentrate on the view of the user or reader, and tools to support their work. The reader has particular characteristics:

- They may have very variable knowledge of the domain.
- They are unlikely to have detailed knowledge of any indexing scheme or query language.
- The importance or relevance of a document is dependent on their prior knowledge, their current information needs and their ability to assimilate any knowledge contained in the document.
- The reader may not know the accepted boundaries of the domain.
- Knowledge assumed as a prerequisite may also be unknown to the user.

In this environment, it is vital for the user to be given as much assistance as possible in the understanding of the meaning of his or her query, in terms of the set of documents that are likely to be needed. This leads to the need for a translation process to allow dialogue between the authors, librarians and readers. By allowing users and groups of users to respond to not only particular documents, but the classifications of documents the process can begin.

Such approaches include the use of collaborative filtering (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994). In a collaborative filtering approach, the decisions of large numbers of users with respect to the relevance of documents are recorded and subsequent queries weighted in response, the idea being that the more people prefer a particular document in response to a particular query, the more relevant that document is, and the more likely that document should be returned when that query is performed.

There are a number of issues arising with using collaborative filtering. One is deciding on the method of deciding how users are similar to one another, and recent work has concentrated on this area (Jin, Chai, & Si, 2004). This work assumes that users who have similar ratings same documents are likely to have similar interests, but extends this to include the fact that some users' rating preferences are different. By concentrating on some measure of rating of documents by users, systems hope to improve relevance of the returned set. However, differences between users may mean that the relevance of a particular document to a particular query is different between users. In order for collaborative filtering to work at all, there must be some assurance that the users concerned would share a common understanding of the required relationship between the query and suitably relevant documents. In other words, users should mean the same thing when they ask the same question. Systems that work from similarity measures in terms of log files or documents selected have the potential drawback that they do not include the characteristics of the user except as demonstrated in their actions.

At the same time, systems with large numbers of users modifying the behaviour of query engines are likely to be very powerful. Collective intelligence (Wolpert & Tumer, 2000), allows emergent sophisticated behaviour to be produced by relatively unsophisticated means. It is important to recognise that the group producing the collective intelligence should be based on the readers of the documents, because of the different roles associated with document use – see Figure 20

Table 8: Relevance schemes based on user groups

Dominant group	Source of data	Examples
Users	Response to recovered data, query formulation	Examination of logs, Explicit rating schemes
Authors	Hyperlinks, Text words	PageRank, vector space model
Librarians	Keywords, Index terms	Traditional Boolean searching.

.

2.5 Ontologies for information retrieval

*“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).*

Ontologies are often used for the hierarchical classification of a knowledge domain. In effect they allow the relationships between objects to be codified and represented in a way that can be used for decision making. However there is a major problem in the way that traditional ontologies deal with two aspects of knowledge. Firstly there are differences between people’s perception of the arrangement of knowledge, and secondly the fact that the same object may legitimately live in multiple locations within an ontology. The use of a fuzzy ontology attempts to assist in the resolution of these issues, or at least build a suitable framework for the assessment of them.

There is also an issue that a word may be ambiguous, that is to say have multiple meanings with the same spelling (for example the word 'rose'), however this is a problem that is specific to the language being used and mode of transmission (whether via speech or writing) and has been the subject of much research by Artificial intelligence researchers. The MeSH System for example uses particular concept codes, which may have multiple human language synonyms. This system can be extended to other languages but the ontological problem is more profound than this.

To give an example of the problem of different perceptions - consider a "car". It may be agreed that it has four wheels and moves people about. A pure classification system would be concerned only with successfully identifying a car, and, for example it is possible to think of a car with six wheels or one that has no driver or passengers etc. However there is another issue here concerning ontologies. A car salesman would think of a car, as a product, and it would have characteristics such as price, manufacturer etc. In contrast a child crossing the road would see a car has having the characteristics of speed and direction in order to assess whether it represents a risk in crossing the road. Obviously the car as an object possesses both sets of attributes, but the importance of each is particularly dependent on the person whose knowledge is being represented, and the context that person is in.

In his fascinating textbook “Linguistic Categorization” (Taylor, 2003), Taylor discusses another complicating factor, that of polysemy. According to WordNet polysemy, also known as lexical ambiguity, refers to “the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings” (D.Slomin & Tengi, 2003). Taylor gives a number of examples which show the subtlety of differences of meaning when polysemous terms are used in everyday sentences. This is mostly the domain of natural language processing and as such largely out of the domain of this thesis, but a number of issues are relevant. Firstly there are many common polysemous terms, and they have considerable impact on the meaning of sentences. Secondly, dictionary definitions do not always distinguish effectively between these meanings, and the same may apply to ontological use. For example using the term “pain”, the sentences “Using a computer can be a real pain”, and “If I use my computer without a break, I end up in pain” have different meanings as sentences. However, the appropriate location of the term “pain” in the ontology –(see Table 9) – is probably G11.561.796.444 in both cases. The ontology does not reveal whether the term is being used metaphorically or not.

This makes searching difficult as any broadening of an initial particular term may expand an inappropriate set of related terms, which occur in a different part of the ontology. WordNet (D.Slomin & Tengi, 2003) is a good example of a system that has attempted to overcome this issues by including definitions, synonyms, hypernyms ("x is a kind of", i.e. objects higher on the ontology), hyponyms ("is a kind of x" i.e. objects lower in the ontology), Holonyms ("x is a part of" i.e. dependent child of another term) Meronyms ("parts of x" i.e. dependent children of this term). Along with examples of the way the term is used – contextualisation – WordNet allows users to identify the meaning of a particular word. For terms in a language with an accepted syntax this is fine. The problem occurs when the term has a number of meanings for different people, this is especially prevalent in the case of people searching for new information, where their own understanding, and knowledge of key terms in a field is limited, or coloured by their understanding of previously encountered terms.

Table 9: Multiple Occurrence Example –"Pain"

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 9 demonstrates how these issues arise in existing ontologies. "Pain" occurs in five locations in the MeSH Ontology. 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. A graphical approach can illustrate this more clearly- consider the obstetrical and surgical forceps in Figure 21.

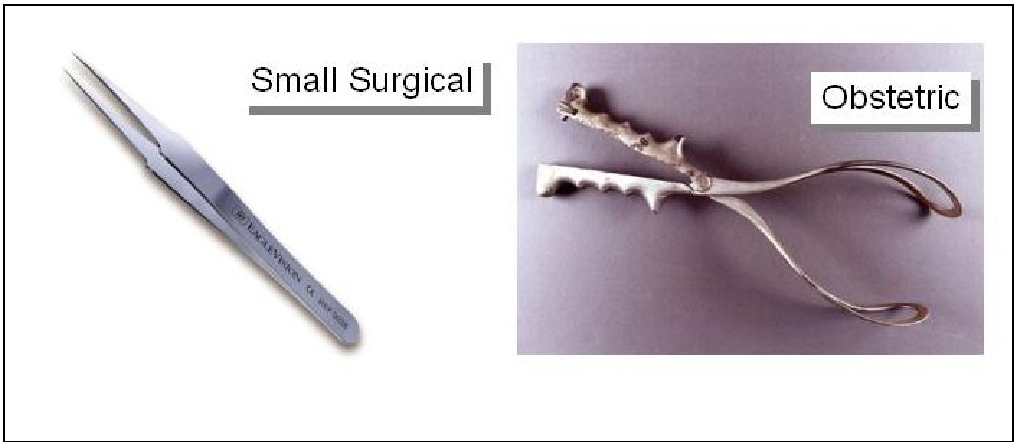


Figure 21: Forceps

Currently ontologies are seen as one of the key technologies involved in the “semantic web” (Berners-Lee 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 (N. F. Noy & Musen, 1999), which is related to the PROTÉGÉ ontology development and validation tool (M. A. Musen, Gennari, Eriksson, Tu, & Puerta, 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.

Fuzzy set theory has been extensively used in the context of information retrieval (Bordogna & 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.

The UMLS MeSH ontology is a particularly useful one for the medical domain. However, of the 21,836 terms within it, 10,072 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.

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, Henzinger, & Moricz, 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. The consequence of this is that general-purpose query expansion tools may cause difficulties because they will identify irrelevant results which the users will be unwilling or unable to refine effectively and may abandon searches before they are successful. A further difficulty is that users may misinterpret the use of the terms in documents where the author has one meaning in mind; an example of this would be an apocryphal complaint by a new manager in a hospital over the cost of employing “consultants”. The managers background included a meaning of consultant linked to management consultants, but the hospital document was relating to consultants as an alternative term for medical specialists. By clarifying multiple meanings where possible ontologies can help to remove such ambiguities, but the ontologies must be “fine-grained” enough to allow this process for people from different backgrounds, where people from common backgrounds would not recognise that an ambiguity exists.

Ontologies have been devised for many domains and a number of systems have been designed to allow communication between ontologies – for example XOL, an otology exchange language implemented in XML and developed from Ontolingua (Karp, Chaudhri, & Thomere).   The Ontolingua server (Stanford University Knowledge Systems Laboratory, 2001), allows online collaborative editing of ontologies, and the Knowledge system laboratory is also the home of a library of ontologies – for example the enterprise ontology and others. Ontologies or hierarchies have also been constructed from sets of documents (Kruschwitz, 2003), and used for recommender systems (Akkermans et al., 2004; Middleton et al., 2004).

2.6 The Semantic Web

One of the major problems facing users of the WWW is the lack of information about the content and meaning of WebPages. Currently it is very difficult for users to automatically select pages, or parts of pages that have particular meanings without reading them. For example pages that deal with travel do not have standard ways of identifying destinations as opposed to starting points, there are a wide variety of methods of displaying authorship, and even elements such as abstracts can be difficult to identify. When humans actually read a document they may have little difficulty identifying particular items, but it is extremely difficult for machine systems to reliably extract the context and/or meaning of a particular piece of information. In part this is due to the original flexibility of the WWW in terms of publishing and presentation, but with the enormous number of web pages now present, this now makes finding the appropriate information harder to support.

Tim Berners-Lee has coined the term “The Semantic Web” (Berners-Lee et al., 2001). This refers to the prospect of constructing systems that allow searches for meaning rather than matching. In essence a semantic web will allow a user to express what they mean to find and find objects that satisfy that request without regard to the language or syntax of the request. The semantic web implies ways of representing the meaning of documents and constructing queries to discover that meaning. If possible, such a process will be as automatic as possible and incorporate the advances in computing power available to users on the desktop and at the server.

Systems for implementing and exploring the “semantic web” have gained increasing prominence recently. It should be emphasised that with the increasing use of document management and knowledge management systems, the “semantic web” implicitly includes intranet as well as Internet applications. The problems of identification of the meaning of documents and other objects exists within systems as well as between systems and this explains some of the interest in tools such as XML for system development.

An extension to this approach includes including document tagging for reliability, source etc. as described by (Grutter, Eikemeier, & Steurer, 2001). This extends the idea of the semantic web to include not only the meaning of the text contained in a particular document but also the status of the documents themselves. Grutter etc. acknowledge the difficulty of generalising this approach for all documents, but suggest that this approach could be used for a set of documents which are to be searched in a bounded database. Their experimental work suggests that such an approach can be successful in such an environment. In this case the documents themselves do not need to be part of the database, but the document type definitions can be added to the standard bibliographic details such as keywords and abstracts in order to increase the likely hood that the documents retrieved are useful. However, this tagging is still done by hand and so involves a great deal of work.

2.6.1 Ontology and the Semantic Web

In Computer Science terms an ontology useful to "express formally a shared understanding of information" (N. Noy et al., 2001) page 60. Another definition has been given by Gruber "An Ontology is a specification of a conceptualization" (Gruber, 1993) – this paper also first described "Ontoligua", one of the first approaches to the specifications of ontologies.

Web based ontologies can be seen as a part of implementing the "semantic web". The Stanford medical informatics group suggest their language – PROTÉGÉ-2000 (N. Noy et al., 2001) as a means of representing ontologies on the web. The central role of ontologies in knowledge management is emphasised by (Staab & Maedche, 2001). In their work they describe a knowledge portal underpinned by an inference engine that uses domain specific ontologies. This enables appropriate index terms to be assigned to queries.

An ontology acts as guide to the relationship between concepts in a domain, which may or may not themselves represent physical objects. By the definitions above, it may be a little uncertain as to what an ontology is useful for. There seem to be a number of main uses:

- By understanding the relationships between objects, then the interaction between objects, that is the operations that can be performed on them, and the appropriate position of new objects becomes easier. For example, if an ontology contains the class "apple" as a type of fruit, then operations that are performed on fruit can be performed on apples (inheritance).

- A very diverse range of relations can be represented, for example anatomical ontologies can use relations such as “is attached to”, “The finger is attached to the hand”.
- Knowledge sharing and abstraction between groups can be supported. For example- is a tomato a fruit or a vegetable? The use of ontologies can make assumptions more visible, to encourage knowledge sharing and discussion.
- Computer systems can use the ontology to discover the semantics of phrases, and construct valid sentences and operations. In a similar way to syntactical analysis, agents can then parse unstructured material to produce material that is more structured.
- Information retrieval systems can use ontologies to expand and refine queries.

Tacit ontologies seem to exist in everyday life, as people generally appear able to share knowledge of relations in normal activities. Programming computer based systems to understand and share these relations in a flexible and efficient manner appears to be more difficult. However, the problem of whether different sources or domains of knowledge should have their own ontology, or whether all knowledge can be fitted into one ontology is still debated. In practice, multiple ontologies exist, and much current work is devoted to combining them, for example the work of (N. F. Noy & Musen, 1999). Many ontologies have extremely rich sets of relations available e.g. Unified Medical Language System (UMLS), WordNet, PROTÉGÉ. The relations for UMLS are described in Chapter 2. There has been a great deal of interest recently (M. Musen, 2001) in the construction of ontologies for representing medical knowledge. In many ways an ontology is similar to an XML document, or a class in an object-orientated programming language (Ensing, Paton, Speel, & Rada, 1994). The representation encodes a hierarchical structure, with inheritance of properties from root to branch, with additional attributes at each level. The advantage over the fixed hierarchy of say ICD10 is that such ontologies can be modified with the inclusion of new branches or leaves, and the location of such a new item can give information about it even to users that are not aware of the new terms used. This approach has been encoded in XML, for use in an electronic health record (Spidlen, Hanzlicek, Riha, & Zvarova, 2005). The authors report that their system is effective for storage and searching, but as may be expected, the need to enter data in a structured way can involve changes to working practice, so that semi-structured and unstructured data is still present in the system.

The ontology approach has obvious similarities to the concepts of object –orientated programming (OOP), both use hierarchical collections of classes for example, but as

Noy points out (N. F. Noy & McGuinness, 2001), ontologies concentrate on the structure of a class, whereas OOP concentrates on the operation of a class - OOP is about what a class does, ontology is about what it is. Both have the emphasis on constructing structures that can be reused and crucially, are understandable, translatable and limited in scope.

Systems for implementing ontologies in order to implement the “semantic web” have gained increasing prominence recently. It should be emphasised that with the increasing use of document management and knowledge management systems, the “semantic web” implicitly includes intranet as well as Internet applications. The problems of identification of the meaning of documents and other objects exists within systems as well as between systems and this explains some of the interest in tools such as XML for system development.

The UMLS system includes an approach to semantic indexing. In the UMLS concepts are identified which may be expressed as strings or words. Synonyms for these terms are stored, so for example a particular concept – “Asthma” has a concept code – Strings such as asthma Asthmas etc. are assigned to the same concept. This approach is obviously useful to identify roots, and allow systems to be used in different languages. The ontological relationships from Table 5 above then allow these concepts to be related to other concepts.

A standard for the expression of ontologies on the WWW has recently been developed. This standard – Ontology Web Language (OWL) (Smith, Welty, & Deborah L. McGuinness, 2004) allows ontology relations to be coded in a machine and human readable form within files that can be located on the WWW.

An example of an OWL statement is shown below in Figure 22– the example is taken from (Smith et al., 2004).

```
<owl:Class rdf:ID="TexasThings">

  <owl:equivalentClass>

    <owl:Restriction>

      <owl:onProperty rdf:resource="#locatedIn" />

      <owl:someValuesFrom rdf:resource="#TexasRegion" />

    </owl:Restriction>

  </owl:equivalentClass>
```

```
</owl:Class>
```

Figure 22: Owl example

This example shows the definition of an equivalent class – items that satisfy the conditions given in the equivalent class construction are regarded as equivalent. In this particular example, things from Texas are assigned to be “Texas things”. This project is still at a fairly early stage, but obviously represents a potentially beneficial approach to standardisation that may be useful for the deployment of the fuzzy ontology described later.

Systems for encoding semantic information do not have to rely on XML or other general-purpose systems. An extremely general-purpose approach to adding semantic information to documents is the WorldNet approach described in (Mihalcea & Mihalcea, 2001). In WorldNet, there are a number of keywords, each of which has a large potential set of attributes, such as synonyms, hyponyms, related terms etc. which are used for searching. Documents have their own schema that includes information about the position of e words within the document. The authors describe a successful test of this approach on the Cranfield data set. The Cranfield data set consists of around 1400 abstracts on Aeronautics collected by Cyril W. Cleverdon between 1957 and 1968. There are around 220 standard natural language queries that allow querying systems to be tested, by seeing if the results of the query via the system correspond to the human-selected results. An example of a query is “What similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft?” The set is available from ftp.cs.cornell.edu/pub/smart/cran/.

Unfortunately this approach still includes difficult problems – such as disambiguation and word root identification. As an approach focussed on one particular problem – retrieval of text documents -it may be that it is initially simpler than efforts to encode and transmit meaning such as XML and RDF.

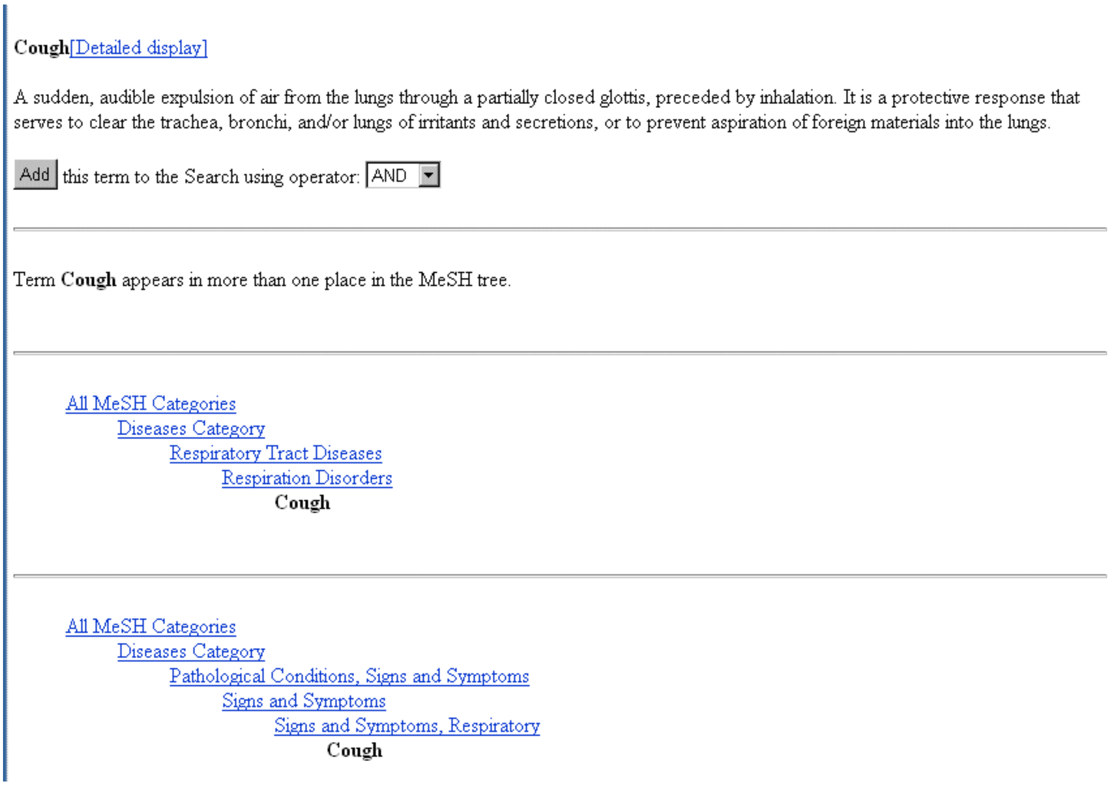
A number of systems have been designed to allow communication between ontologies – for example XOL, an ontology exchange language implemented in XML and developed from Ontolingua (Karp et al.). The Ontolingua server (Stanford University Knowledge Systems Laboratory, 2001), allows online collaborative editing of ontologies, and the Knowledge system laboratory is also the home of a library of ontologies – for example the enterprise ontology and others.

The fact that ontologies is a plural raises the major difference between the philosophical and computer science approach to the term. A philosophical ontology would encompass the whole of the universe, but computer scientists allow the existence of multiple, overlapping ontologies, each focussed on a particular domain. Indeed an understanding of the ontology of a particular domain may be crucial to any understanding of the domain. The combination of ontologies, and communication between them, is therefore, a major issue within computer science, although such issues are problematic with the philosophical use of the term. 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.

Similar issues also influence the scope and coverage of ontologies. As summarised in a recent discussion in IEEE expert (Brewster et al., 2004), ontologies can attempt to represent the maximum amount with the minimum complexity – a “Newtonian” approach, or take the “Liebenizian” route where the ontology represents the complexity inherent in the different objects covered by it. Moving to a practical level, much of the complexity of relations in systems such as Ontolingua is due to the need for machines to be able to reliably use the systems for classification or comparison. If an ontology is primarily used by humans, then some of the more complex and specific relations may not be required. Ultimately, if an ontology is dealing with human-defined or subjective terms – for example particular sensations or observations, then a great deal of precision and complexity in the ontology relations may give a misleading view of the precision of the description of the objects included in the ontology. At the same time, ontologies may be able to straddle the differences between the Newtonian and Liebenizian views by having different levels of representation for different purposes, this may involve the merging of classes and relations for a broad view and the dissection and sub-classification of the same domain when greater precision is required. This issue represents one of the main drivers for the development of the fuzzy ontology, by setting the membership values appropriately such a transformation can be made. This is explored in more detail in chapter three.

The MeSH system can also be easily represented as an ontology, although in this case an object can be allowed to be a member of more than one leaf element at the lowest level. This is because a particular document can be “about” many different and closely related subjects. It is also possible to have the same leaf element incorporated in a number of different branches. This creates problems for some representation systems and artificial distinctions have to be made between the same term depending on which branch it has come from. The figure below (Figure 23) shows the location of the term “cough”.

A keyword or index term hierarchy can be seen as a particular class of ontology. At the same time documents can be analysed in terms of such ontologies. In this case the attributes used to define a document are themselves divided by navigating the tree that represents possible classifiers. In the information retrieval case it is highly likely that documents will possess many attributes, some of which may tend to place it in differing parts of the ontology. For example, a page produced by a sufferer’s charity about a disease, may have attributes, such as brevity, low reading difficulty that indicate that it would be located in a tree associated with public information, but it may have valuable links to research teams and means of getting funding that would make it suitable for researchers, and located in that tree.



Cough[Detailed display]

A sudden, audible expulsion of air from the lungs through a partially closed glottis, preceded by inhalation. It is a protective response that serves to clear the trachea, bronchi, and/or lungs of irritants and secretions, or to prevent aspiration of foreign materials into the lungs.

Add this term to the Search using operator: AND

Term **Cough** appears in more than one place in the MeSH tree.

All MeSH Categories
    Diseases Category
        Respiratory Tract Diseases
            Respiration Disorders
                Cough

All MeSH Categories
    Diseases Category
        Pathological Conditions, Signs and Symptoms
            Signs and Symptoms
                Signs and Symptoms, Respiratory
                    Cough

Figure 23: The MeSH tree that includes “cough”

However, there is a major drawback with the use of such ontologies for representing every sort of medical knowledge. It is often the case that particular aspects of diagnosis, or aetiology are uncertain. In a traditional “crisp” ontology, an object is either a member of a particular class or not. This leads to the creation of “unspecified” classes for example in the ICD10 system.

The approach of the UMLS Metathesaurus utilises a large number of relations. The full list of potential relations in the UMLS Metathesaurus is shown in Table 6.

It should be noted that these relations are found in the source systems that feed into the metathesausus – within MeSH itself the relations are much simpler.

.

Ontologies support the semantic web by allowing machine based systems to improve the quality of relatedness searches and supporting query expansion and refinement, that is allowing the meaning and the semantic content of documents to become associated with index terms and combinations of terms. This requires mutual understanding between the implicit world-view of the actors in the environment – see section 2.4.

Ontologies have also been seen as ways of increasing the cooperation between search agents and engines on the web via ontology described services – see for example the work of (Akkermans et al., 2004; Sirin, Parsia, & Hendler, 2004). In this approach the ontology is used to allow web services to describe themselves to potential users. This would be of tangential interest to information retrieval workers except for the fact that this process is driving some of the ontology representation work such as OWL – see section 8.4. In addition some of the major applications for web service technologies are related to information retrieval – for example the Google API described in section 5.2.7.

2.7 Fuzzy logic and information retrieval

Early database systems used exact matching techniques for queries, based on Boolean queries. Issues with this approach include the difficulty of formulating vague queries, and issues with dealing with terms which are not indexed. The necessity for the user to know in advance the index terms used by the system, even in cases where “full-text” searching is possible, but not preferred, adds another hurdle. In systems which are searching an open space of discovered documents such as the WWW, rather than a constrained digital library with a limited set of index terms, such approaches are often not successful. Similarly, database queries and information retrieval queries may appear similar, but the difference often lies in the greater control database managers have over the content of their database. User queries based on boolean searches are often poorly formulated, with little understanding of the use of so called “advanced search” pages, and little desire to expend effort using them (Gerwe & Viles, 2000).

Table 10: Relationship terms used in MEDLINE/UMLS

mapped_from	clinically_similar	uses
temporally_related_to	multiply_mapped_to	analyzed_by
measures	surrounds	used_by
has_manifestation	connected_to	consists_of
property_of	carries_out	affected_by
produced_by	surrounded_by	has_tributary
equivalent_to	has_conceptual_part	default_mapped_to
has_degree	result_of	occurs_in
contained_in	exhibited_by	interconnects
affects	classified_as	traversed_by
degree_of	mapped_to	manifestation_of
causes	conceptually_related	classifies
has_branch	default_mapped_from	has_developmental_fo
uniquely_mapped_from	has_part	method_of
isa	exhibits	co-occurs_with
tributary_of	has_occurrence	measurement_of
has_evaluation	inverse_isa	part_of
functionally_related	traverses	uniquely_mapped_to
has_result	multiply_mapped_from	spatially_related_to
derivative_of	associated_with	caused_by
adjacent_to	follows	has_process
complicated_by	location_of	indicates
precedes	contains	has_measurement
has_derivative	has_location	carried_out_by
has_property	developmental_form_o	branch_of
conceptual_part_of	issue_in	process_of
indicated_by	analyzes	produces
constitutes	measured_by	complicates

has_method	interconnected_by	has_issue
evaluation_of		

Fuzzy set theory literature, beginning with (Zadeh, 1965) has been seen as an important development for effective information retrieval from databases (J. Kacprzyk & Bosc, 1995). The mathematical formulism of membership and fuzzy combinatorics are often used to avoid issues around the weaknesses of Boolean logic, in particular in avoiding "all or nothing" relevancy scores. However, the fact that fuzzy theory allows the attachment of meaningful labels to the degree of membership is equally important. Any information retrieval system is designed to be used by humans, who are almost always uncertain of the exact results that they are seeking. By acknowledging this uncertainty, and allowing its representation in the process of query construction, refinement and result evaluation, the system itself becomes more meaningful to the users.

A wide survey of different methods of application of fuzzy logic to information retrieval was outlined in (Bordogna, Carrara, & Pasi, 1995). As with Boolean searching techniques, fuzzy identifiers can be combined. Traditional fuzzy searching and probabilistic relevance measures have used a number of these approaches – for example some of those described in (Crestani, Lalmas, van Rijesbergen, & Campbell, 1998). This article is a review of probabilistic methods of information retrieval, many different approaches have been tried essentially to give some idea of the likely relevance of retrieved material.

2.7.1 Uncertainty in information retrieval

Aside from relevance scores derived from the vector space model, other probabilistic approaches have been used. A wide variety of these techniques are described in (Crestani et al., 1998). Much of this work deals with methods for calculating relevancy scores, and most use some application of Bayes theorem (Pearl, 1988). If $\vec{x}$ represents a particular document vector, and R is the relevance then:

$$P(R \mid \vec{x}) = \frac{P(R).P(\vec{x} \mid R)}{P(\vec{x})}$$

Where P(R) is the prior probability of a relevant document being recovered from a particular set, $P(\vec{x})$ is the probability of a document being represented by the document vector $\vec{x}$, and $P(\vec{x} \mid R)$ is the probability of a particular document having relevance R. Thus the probability that a particular document is relevant to a particular query is related to the closeness of the document and query in vector space, and the degree to which such document vectors commonly occur.

These models have shown a good deal of success. Other approaches have refined these models by introducing weightings related to the location of terms – for example the work of (Kruschwitz, 2003).

Because of the nature of the information retrieval process it is likely that the degree of uncertainty or vagueness in queries will alter as the process continues.

2.7.2 Fuzzy search

Fuzzy search in databases and information retrieval has been popular for some time, with extensions to the SQL command set being discussed in (Bosc, Galibourg, & Hamon, 1988; J. Kacprzyk & Zilkowski, 1986), and in more detail in later publications such as (J. Kacprzyk & Bosc, 1995). The advantages of fuzzy search include; more natural mapping of term combinations in queries ("somewhat related to", "strongly related to") and greater meaning assigned to relevance scores (" a little relevant", "very relevant"). Zadrozny and Kacprzyk have also implemented fuzzy searching in an access environment as the Fquery interface (Zadrozny & Kacprzyk, 1996). The challenge of using fuzzy querying on the web is also gaining increasing attention (J. Kacprzyk & Zadrozny, 2003).

A recent approach is described in (Widyantoro, 2001). In this paper, the authors use a fuzzy ontology to represent the relationship between terms. Effectively it uses fuzzy search techniques, as opposed to the more traditional Boolean or crisp logic, but the authors realise that such an approach can also be represented as an ontology. This can be compared to the MeSH system described above, where the basic relation is a crisp "is-a", but other relations such as contains, is a synonym etc. are also included In fact, the full UMLS relationship model is more complex and this is described below. The Widyantoro method envisages two relations, *Narrower than* (NT) and *Broader than* (BT).

The advantage of this method is that the degree of membership can be calculated by using the frequencies of the occurrences of the terms being used. The relations are simple, but they can be calculated automatically. In addition, by calculating degrees of membership automatically, this method allows automated trimming of the trees produced, by combining the membership values of the elements, calculating the overall membership value and then rounding, so that membership values greater than 0.5 become true and those 0.5 or less become false. False relations are then removed from the final tree. The choice of 0.5 is arbitrary.

More sophisticated representation than merely NT and BT are also possible. Takagi (Takagi, Kawase, Otsuka, & Yamaguchi, 2000) describes the use of associative memories to allow assessment of objects by potentially large numbers of different criteria. However this has the disadvantage that calculation of the membership function is not obvious and may become complex. In particular the practical difficulties involved in defining which aspects should have fuzzy values, and which characteristics of the objects should be chosen to contribute to these membership values remains problematic. Essentially, assigning these values may well be a difficult task, and the end-user should have a method available to modify the result by using their own criteria.

The next chapter deals with work around the implementation of some of these ideas.

Chapter Three

This chapter introduces one of the key concepts of this work, the fuzzy ontology. Fuzzy ontologies are suggested as a solution to some of the problems associated with crisp ontologies in the field of information retrieval.

Section 3.1 provides an introduction to the concept of the fuzzy ontology. Section 3.2 describes the advantages of this approach. Section 3.3 is the main part of this chapter which describes ways of learning membership values for fuzzy ontologies. Section 3.4 discusses the possible meaning of fuzzy ontology in psychological terms. Section 3.5 gives examples of the use of fuzzy ontology combinations.

3.1 Fuzzy ontology

The fuzzy ontology is based around the concept that each index term or object is related to every other term (or object) in the ontology, with a degree of membership assigned to that relationship based on fuzzy logic as introduced by (Zadeh, 1965). The fuzzy membership value μ is used for the relationship between the term or object in question where $0<\mu<1$, and μ corresponds to a fuzzy membership relation such as "strongly", "partially", "somewhat"," slightly" where for each term;

$$\sum_{i=1}^{i=n} \mu_i = 1$$

Equation 1

Where n is the number of relations a particular object has, where n= (N-1), with N representing the total number of objects in the ontology. That is, each term used in the system has the total membership value of its relations as a value of 1 summed over each dependant relation. This rule is not commutative, for the relationship between two objects, A, B:

$$A \xrightarrow{related\_to} B.....\mu_{AB}$$
$$B \xrightarrow{related\_to} A.....\mu_{BA}$$

Equation 2

$\mu_{AB} > \mu_{BA}$ or $\mu_{AB} < \mu_{BA}$ or $\mu_{AB} = \mu_{BA}$ are all possible. An example of how these values may be different is shown in section 6.2. For simplicity, it is assumed there is only one type of relation between the objects A and B. For further simplicity, for each membership value for a relation originating from A, the membership value is written as μ_B rather than μ_{AB} as the A can be assumed. The overall value of 1 is used in order to prevent objects with large numbers of relations having a larger effect in a search process than those with relatively small numbers of relations.

Imagine that such a constraint does not apply. Consider a query with two terms A and B. If A has only 2 relations, with μ_1=0.7 and μ_2=0.6 then Σ=1.3. However if B has 6 relations say, μ_1=0.7 and μ_2=0.5, μ_3=0.7, μ_4=0.6, 5=0.7 and μ_6=0.6 then Σ=3.8. This doesn't matter if A and B are the only elements in the query and both are needed, but if there is a list of terms, so that ANDing all of them is unlikely to retrieve wanted documents then there will be a tendency to select B as part of the final query rather than A and hence a bias will occur, and there will be a tendency to expand via related terms of B. Obviously this argument can be reversed, so that relatives of A could be seen as having too high a membership value, but it is likely that in a well designed ontology, the relatives of very connected terms will have sufficiently distinct meanings to be related but unwanted. Practically, in addition, if a threshold value of μ is used, large numbers of relatives will not qualify and so huge query expansion will not occur.

This does mean that there is not a consistent mapping between the strength of a relationship in words and the membership value, but for each object, the strongest relationship will continue to have the highest membership value.

An example, Figure 24, may make this clearer, each μ represents the membership value of the relationship from the apple to tree, fruit and computer company. Any relationships not shown, e.g. between IPOD and pippin, are assumed to have a μ=0, relationships directed *to* "Apple" do not have μ values shown for clarity.

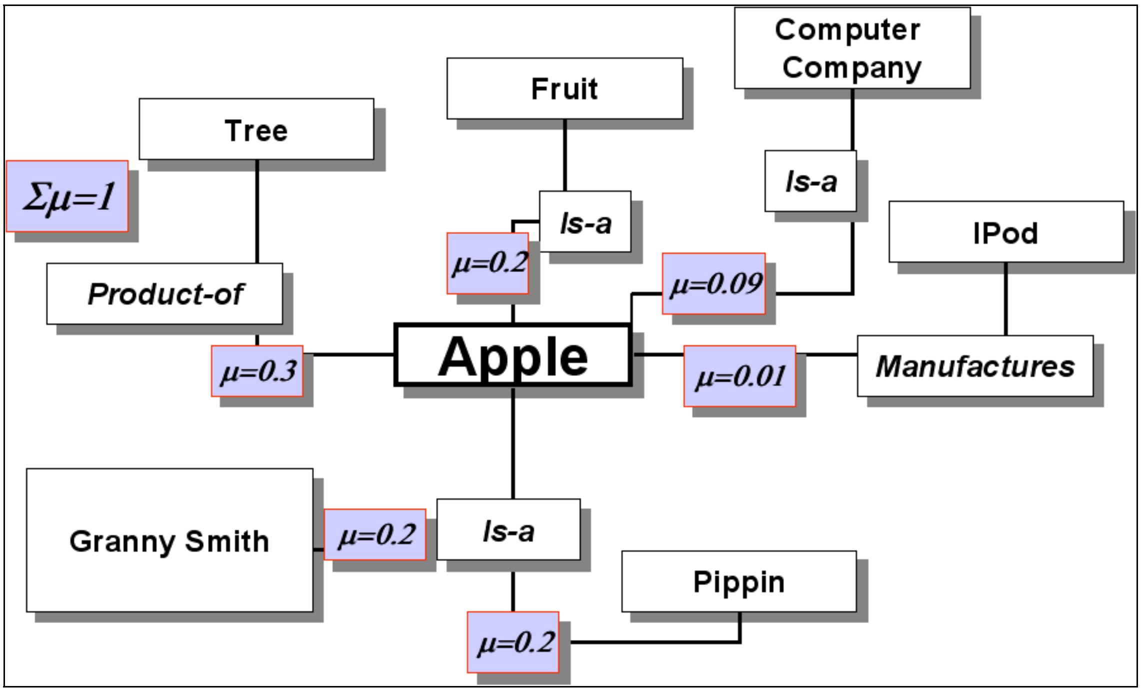


Figure 24: A fuzzy ontology scheme.

The apple can be a product of a tree (to distinguish an apple tree from another kind of tree); it can be a fruit (if the query relates to distinguishing between fruit, for example apple pie, cherry pie) or a computer company. The fuzzy ontology applies membership values to each of these possibilities; depending on how likely it is that a particular relation is required. These different relations will have different membership values depending on the context of the query, and particularly the user's view of the world. The capitalization of the word Apple/apple may also be used to assist this differentiation, but most web-based searching tools seem to ignore this information.

In conventional ontologies, particular objects may occur in multiple locations, leading to ambiguity when being used for query construction. In addition, as pointed out in (Bodenreider, 2001), unwanted cyclic relationships may occur. Unwanted relationships can be assigned zero membership values. In practice, unused terms or objects may also be ignored in the ontology, as there is no need to retain them to act as a bridge between otherwise useful terms.

3.2 Advantages of the fuzzy ontology

Many issues arise from the use of multiple ontologies as seen in section 2.6.1, including the difficulties associated with communicating between ontologies and the need for maintenance of large numbers of ontologies. The fuzzy ontology as described is partly suggested in order to allow a common framework, or base ontology, with different membership values associated with different users and groups. It should be noted that because of the learning methods involved, only “is-a” type relations are currently used, based on the currently existing MeSH hierarchy. This is because there is no difference between the learning of different types of relation.

Another advantage of this approach is completeness. Rather than impose an arbitrary standard of the importance of a particular location in the ontology, which is required in a crisp ontology to avoid too many examples of a term appearing in the ontology, the term or object can be located in all relevant locations.

Most importantly, for searching processes, the use of a fuzzy ontology for the mapping of search terms allows the relative weight of each term in the required output to be calculated. By allowing these weights to be calculated accurately, it removes the bias associated with multiply located terms being used for searching. If a term is located in multiple locations in a crisp ontology, and is used for query expansion purposes – say by including offspring, then the danger is that the large number of relatively irrelevant expansion terms outweigh those which are useful.

In particular, the use of a fuzzy ontology approach allows the convenient representation of the relationships in a domain according to a particular view, without sacrificing commonality with other views; the ontology framework is common, just the membership values are different.

Finally, this approach holds out the possibility that the representation of a potentially very large ontology can be compressed. If whole areas are not required, the relations to the core can be set to zero. Unwanted intermediate levels can also be removed, with lower-level terms only communicating directly with higher levels. This aspect removes the need to create artificial groupings to avoid orphaned terms. At the limit a fuzzy ontology, with all membership values set to 0 or 1, will have each term or object having one relation only. If each term only has one relation then a B-tree structure is possible, with each term only relating to its parent, however this arrangement is more properly called a hierarchy. Table 11 summarises some of the differences between crisp and fuzzy ontology.

Table 11: Comparison of crisp and fuzzy ontologies

Aspect	Fuzzy ontology	Crisp ontology
Multiply-located terms	Does not occur	Issue for disambiguation
Query expansion	Depends on membership value.	Depends on location only
Customisation	Simple, based on modification of membership values	Requires new ontology and/or ontology sharing.
Intermediate locations for grouping	Unnecessary	Needed for construction – may be useful
Storage required	Depends on the number of terms in the ontology and the membership values of the relations, can be smaller or larger than crisp.	Depends on number of terms in the ontology
Knowledge representation	Related to use	Related to structure.

3.3 A learning scheme for the fuzzy ontology

This method assigns a membership value for each potential relation for each term by using the fuzzy categories described above, so that the value of μ_i changes according to the number of terms related to the search term that are discovered in the documents, and the number of documents queried. As part of the data structure of a fuzzy ontology membership value the number of queries using a term is recorded along with the membership value for that term. The updating of the fuzzy ontology membership value 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 (❍$_i$), and the number of queries that have confirmed the intended meaning of this term (Q_{Hist}).

$$\mu_{New} = \mu_{Old} * \left| \mu_i - \mu_{Old} \right| / Q_{Hist})$$

Equation 3

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

L₁'s membership value is decreased to 0.5 then the extra membership values are split in three, so that the new value for μ_2 is given by:

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

Equation 4

Where μ_{Change} is the change in μ_1 in this case – 0.1? Equation 3 describes the updating of the membership value of the object of interest whereas Equation 4 describes the updating other the other synonyms.

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

Group updating is performed in a similar manner, with each individual membership value for each relation being averaged and the overall result being normalised.

A fuzzy ontology can be used to enhance information retrieval in a number of ways. Firstly, the membership value can be used as a cut-off for query term expansion, by adding query terms that can be discovered using a simple scheme such as "Use terms where the mean membership value divided by the minimum membership value is greater than x". This would allow strongly related terms from outside the immediate neighbourhood to be used preferentially. Secondly, the retrieved set could be examined for terms related to the required terms and the weighting of each document adjusted by the membership value. One particular enhancement is the use of the "unwanted" membership value, which, by being given a membership value of -1, acts as an implicit "Not".

Another view of the "fuzzy ontology" is that of a standard ontology, but with membership values attached to the target-root relationship. The terms involved in the ontology represent specific concepts, however the relationship in the hierarchy is dependent on the fuzzy modifiers which leads to a fuzzy ontology (see Figure 25).

The fuzzy ontology is based on modification of an existing crisp ontology. Currently there are ontologies with an extremely rich set of relations between members, for example the UMLS has 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", many of these come from non-MeSH sources (see Table 10 for some examples). By preserving these relations, an extremely rich set of relations can and do 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.

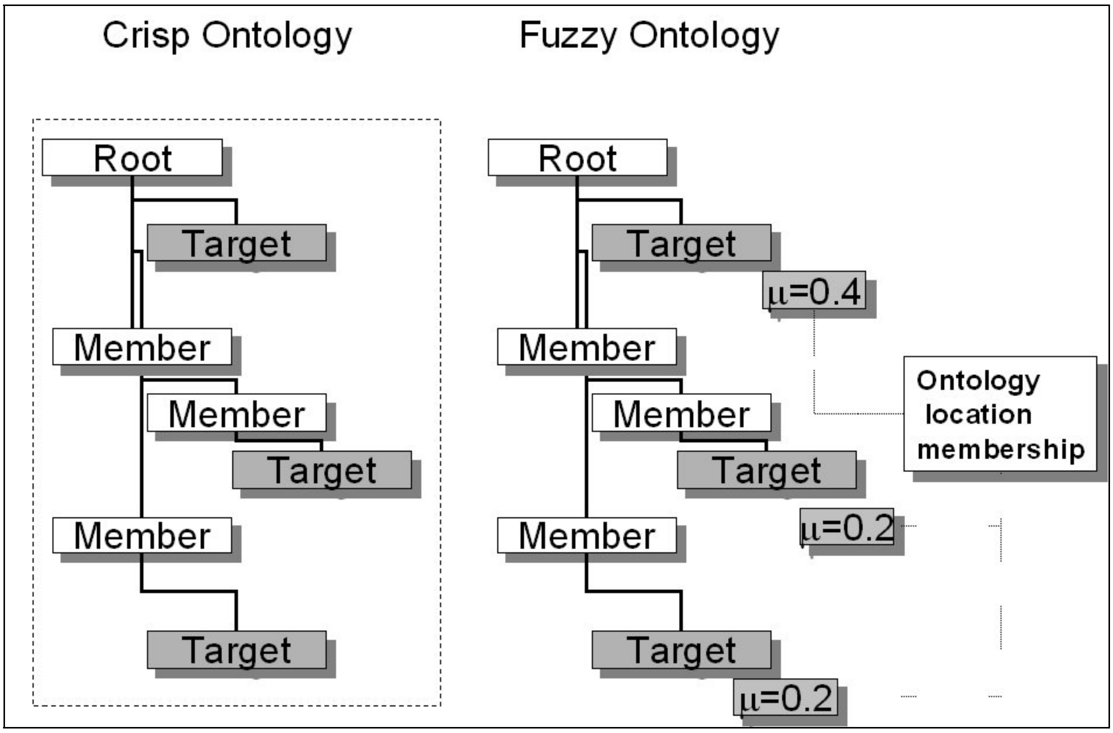


Figure 25: The Fuzzy Ontology

This is because it is primarily concerned with mapping from queries to the ontology. In the vocabulary of Noy (N. F. Noy & Musen, 1999), 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 terms used for the fuzzification are shown in Figure 26, along with the degree of membership associated with each relation. The use of these terms is described in section 3.4. These terms were chosen as fuzzy labels because they are easy to understand, and refer to relations between terms rather than a general quality of a term. This is important because the system requires relations to be investigated to construct the fuzzy ontology. The “opposite” label is used in order to allow otherwise useless terms or documents containing these terms to be excluded. Its role is similar to that of the “NOT” operator, if for example one was searching using a term with a number of different meanings. An example would be the use of “Labour” NOT “Unions” to exclude organised labour pages when searching for obstetric labour.

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.

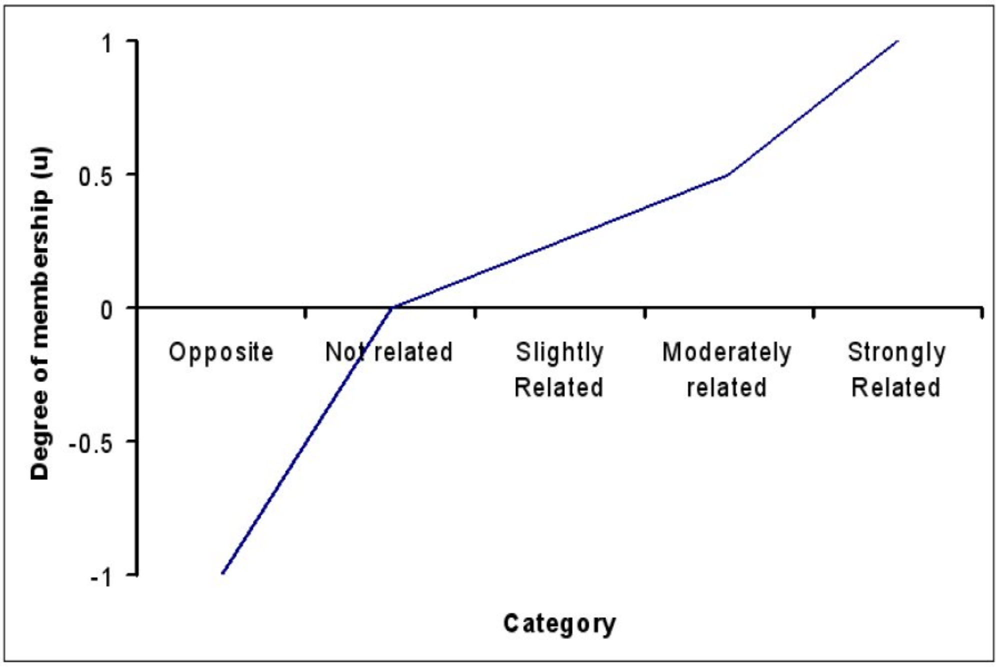


Figure 26: Fuzzy ontology membership

The key element in the creation of the collective searching systems 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 fuzzy ontology, with modifications via an incremental learning algorithm when queries take place and the relevance score noted. As part of the data structure of a fuzzy ontology the number of queries using a term is recorded along with the membership value for that term.

A similar algorithm performs the individuals' fuzzy ontology 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 fuzzy ontology and the collective one in other cases. The algorithms selected for updating are extremely simple, and other approaches may well be useful.

It is important to note that the updating process will produce different values for the individual and the collective membership values within the fuzzy ontology for the same term. Therefore a collective fuzzy ontology, and an individual fuzzy ontology both need to be maintained. In the implemented system, when the user has to assign terms to various relatedness categories, the users selects terms from a selected document and drops them into the boxes corresponding to those relations. Details of this process are described in the system design and description section (Chapter Five). The important aspects of this approach are:

- The assignment of terms is done in the context of a document.
- Terms are located into the ontology by assigning a degree of membership, using the term, rather than having a slot in an existing ontology requiring a term.
- Only terms that the user believes are important are included in the analysis.
- The ontology assignment process is voluntary, and is linked to the user's assessment of the usefulness of the document.

3.3.1 Automatic membership allocation

This section deals with learning the membership values of relations from documents rather than directly from users. The rationale for this approach includes the fact that very large collections of documents are available, the process does not require recruitment of large numbers of users and different algorithms can be tested easily. The overriding reason is that with human-based systems, the danger is that large parts of the ontology will remain unvisited or unused or a long time and that the results obtained from those few people may be biased or simply inadequate. If fuzzy ontology supported search is to be provided, it will be useful to have at least some coverage of all the likely domain areas, even if refinement is necessary as time goes by.

Initially a set of query terms was 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 (www.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 that 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 were weighted as 3, terms in the title (PubMed) or headings were weighted as 2 and terms in the abstract (PubMed) or main body were weighted as 1. The aim of this weighting is to boost the importance of terms that are likely to be central to the meaning of the document. A “local term” was defined as one that 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).

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

Equation 5

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 above. The difference in weight according to the number of queries performed is deliberate in order to use the fact that both Google and PubMed are being queried with the results delivered in descending order of relevance. Changes in weighting may need to be introduced in the future to emphasise particularly important documents, or to acknowledge revolutionary results, which may need the introduction of a “forgetting” mechanism.

This approach is in some way similar to the approach described in (Wollersheim & Rahayu, 2002). The authors of this paper use the UMLS Specialist lexicon and elements from the metathesaurus in order to build a dynamic taxonomy. A dynamic taxonomy can be used to support browsing operations by allowing only those terms that are related to the current term or document to be highlighted. In some ways this approach mirrors that of the hyperbolic browser (see Figure 12). However this approach is most useful for individual searching episodes rather than inter-searcher support over a longer timeframe.

3.4 Cognitive Psychology and Fuzzy Ontology

One question that is always important in any representation of the world, and especially ontologies is whether such a representation is reasonable in terms of the way people actually think. In this case the relationship between language, and the world language describes, is vital.

The question of whether language proceeds thought or vice versa is a key issue in developmental psychology, and has been a source of fascination for centuries (Bloom & Keil, 2001). This issue is actually very important in terms of the meaning of ontologies, as it raises the issue of whether the ontologies can be entirely based on the terms included in them – that is that language causes thought – or that they may be representations in language of existing relations. Obviously the choice of terms or objects and relations in an ontology affects its’ expressive power, but the question can be asked as to the degree to which an ontology represents existing knowledge, as opposed to the view that ontologies create knowledge.

This dichotomy is unlikely to be resolved soon, but recent work (Hespos & Spelke, 2004), appears to show that children are able to distinguish between classes of objects even when their language does not support such a difference. This seems to support the idea that language, and by extension ontologies are primarily representations of knowledge rather than their origin. Given this, then a potential objection to a fuzzy ontology can be opposed. If an ontology is seen as a source of knowledge, where the relationships and objects represented exist as things in themselves only, then fuzzification threatens to destroy the value of the ontology. However a fuzzy ontology is permissible and can be useful if the primary aim of the ontology is to represent and communicate pre-existing knowledge or objective or even subjective information.

The fuzzy ontology approach may provide a simple method of reusing ontologies, where terms occur more than once in the ontology. The ontology can be seen as an acyclic directed graph, with nodes representing index terms or objects, and edges representing the relations between them. 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 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 fuzzy ontology automatically constructed from that may have different location membership values than that discovered from MEDLINE.

In addition to the technical use of a fuzzy ontology for information retrieval, such an approach can also be used in a similar way to that of "meaning chains" (Taylor, 2003), in language understanding. This approach is similar to Markov chain analysis in that the likely meaning of a sentence or series of sentences is altered as new parts of the sentence are parsed, making certain meanings more likely. In a fuzzy ontology based analysis the particular sense of a term is indicated by the degree of membership of relations to that sense of the terms within a sentence or query. This is addressed in section 3.1 in terms of the combinations of fuzzy ontology-based terms. By specifically addressing the overloaded nature of terms in an ontology, the fuzzy ontology allows the use of ontologies for information retrieval to be extended.

In a conference presentation, (Mamdani & Bonissone, 2004), identify some of the issues that make fuzzy logic hard to reconcile with physical reality. Their conclusion, that the important information is in the knowledgebase, rather than in the tool that queries the knowledgebase, and that such tools can be used without profound understanding of the basis for their development so long as that development is sound and supports the aims of the user. One drawback of this black-box approach is that the means by which relevancy scores are calculated becomes obscure, and this can lead to the information not being trusted or the best approaches understood (Gori & Witten, 2005). Such issues may be addressed by publishing the fuzzy ontology, and allowing direct modification of the fuzzy ontology by users as well as producers

3.4.1 The MeSH Obstetric Domain

The problem of ambiguous terms applies to technical language as well as ordinary speech. This section describes the characteristics of the MeSH hierarchy and especially a subset of this hierarchy that was selected for its relevance to Obstetric information.

Table 12: MeSH term overloading

Number of Occurrences of this term in the MeSH hierarchy	**Number of terms**
18	1
15	1
14	2
13	2
12	15
11	18
10	11
9	31
8	53
7	109
6	190
5	391
4	1072
3	2288
2	5888
Total	**10072**

By including each term in each tree that included these terms and below, then 472 terms were identified. In total there are a total of 39,828 positions within the complete 2003 MeSH hierarchy. The terms selected are shown in Table 13.

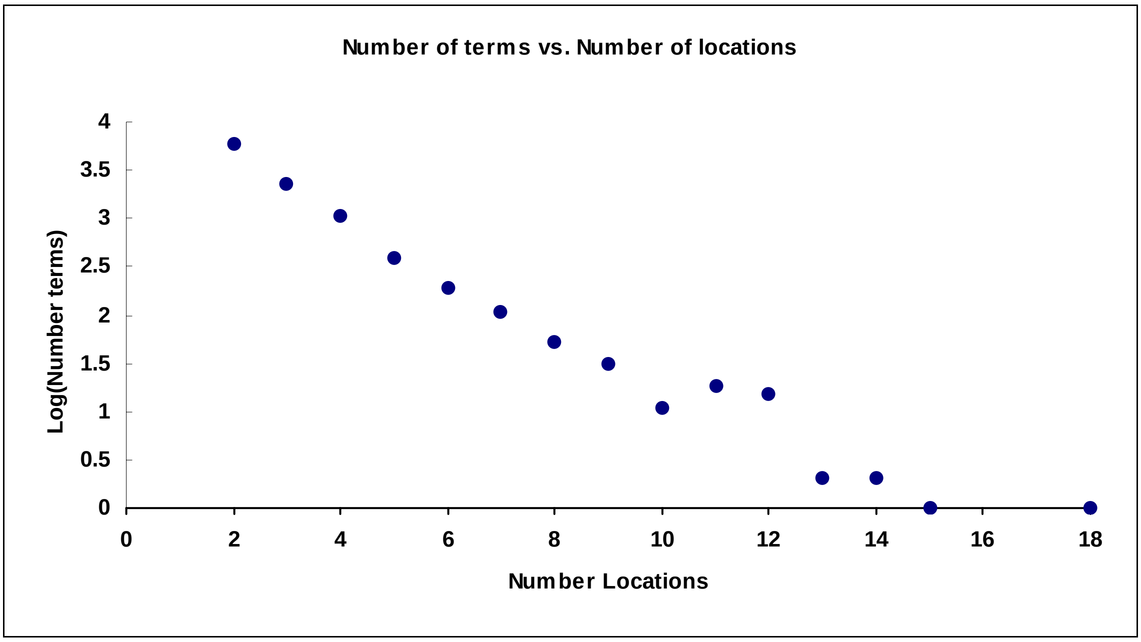


Figure 27: Frequency of multiply occurring strings in the MeSH hierarchy

Table 13: Root headings for terms taken from MeSH

Code	**Description**
C13	Female Genital Diseases and Pregnancy Complications
A16	Embryonic Structures
A01.673	Pelvis
A16.378	Fetus
A16.254	Embryo
A16.759	Placenta
A16.950	Zygote
A16.631	Ovum
C02.800	Sexually Transmitted Diseases
C13.371	Genital Diseases, Female
C13.703	Pregnancy Complications
E04.520	Obstetric Surgical Procedures
F03.600	Mood Disorders
F03.800	Sexual and Gender Disorders
G08.520	Reproduction
M01.438	Multiple Birth Offspring

Table 12 and Figure 27 demonstrate that there are over 10,072 “overloaded” entries in the MeSH Hierarchy – that is 10,072 strings appear in more than one place within the hierarchy. Plotting the number of multiply occurring strings (MOS) versus the number of times (Figure 27) these strings appear, demonstrates the exponentially decaying relationship between the number of MOS’s and the number of multiple occurrences with number of MOS’s being inversely related to the log of the number of occurrences. Within the obstetric domain 735 distinct locations were selected in the MeSH hierarchy. These had a far smaller percentage of MOS’s than in the entire hierarchy – see Table 14 and Figure 28.

Table 14: Multiply occurring terms in the Obstetric portion of the Hierarchy.

Number of terms	Number of Occurrences
5	1
4	2
3	17
2	49
Total	69

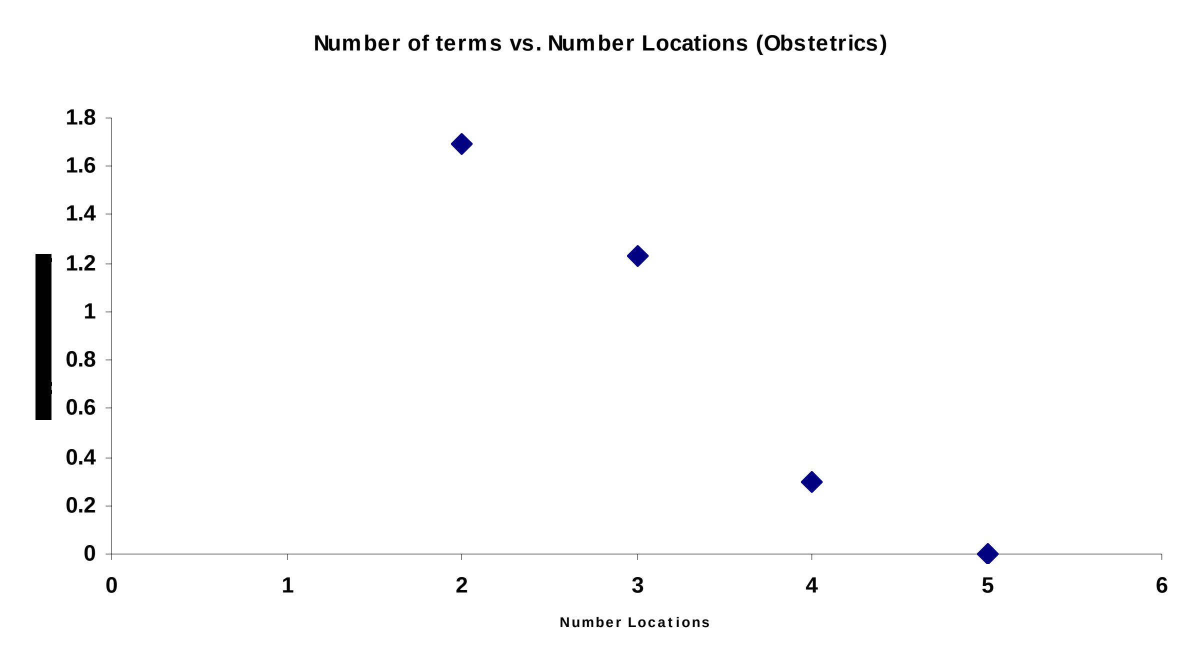


Figure 28: Frequency of multiply occurring strings in the obstetric component of the MeSH hierarchy

At first glance, this appears to reduce the problem considerably. With relatively low “overloading” of terms, and assuming other areas of medicine follow obstetrics’ example, then the possibility of confusion is reduced. However the situation is not that rosy, because the majority of the overloading of terms in a particular area occurs outside that area – see Table 15 and Figure 29.

Table 15: Number of MOS' with degree of overloading between obstetric region and elsewhere

Number of terms	**Number of Occurrences**
18	1
10	1
9	2
8	6
7	7
6	9
5	20
4	42
3	52
2	176
Total	316

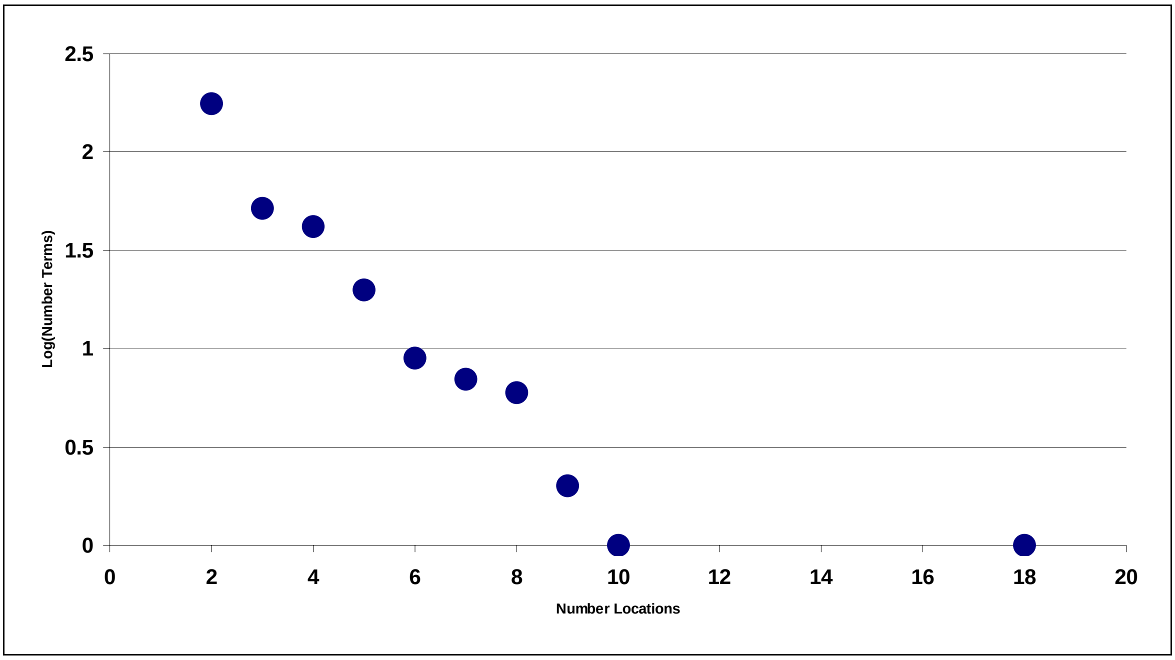


Figure 29: Frequency of Strings in the Obstetric component that also occur outside it

3.5 Example of the use of fuzzy ontology

As with Boolean searching techniques, fuzzy identifiers can be combined. Traditional fuzzy searching and probabilistic relevance measures have used a number of these approaches.

In the case of the fuzzy ontology proposed, there are two main operators used:

- AND - conjunction
- OR - disjunction

The fuzzy logic approach to these combinations is very similar to the crisp one, except that the membership value is included in the calculation. For this part of the work the fuzzy membership function refers to a particular mapping of a term to an object in the ontology, rather than as previously to a particular relation. This is done in order to simplify the examples. The two approaches can be reconciled – if there exists a "standard" set of relations in an ontology, then the degree to which the fuzzy ontologies correspond to this can be seen as an indicator of the preferred location in the crisp ontology.

The reason this is important is in the retrieval part of the process once the fuzzy ontology is generated. Similar work has been done on this already for example (Bordogna & Pasi, 2000), this section outlines some of the methods that can be used to utilize the fuzzy ontology once constructed.

3.5.1 Fuzzy AND operator

For this operator, two or more groups of results are combined in order to provide the most selective coverage of the topic.

The truth table for the AND conjunction is shown in Table 16.

Table 16: Truth table for AND

A	B	A AND B (A Λ B)	Membership value of combined terms
1	1	1	Min (μ_A, μ_B)
1	0	0	0
0	1	0	0
0	0	0	0

The fuzzy "AND" operator used in this thesis is the *min* operator. In this approach, the membership values of each piece are calculated and the minimum membership value is assigned to the combination. For example, assume that there are two terms included in a query; "complications" and "forceps". "Forceps" occurs twice in the MeSH ontology and "complications" occurs 11 times. Simplifying the situation considerably it can be imagined that there are two possible domains that may be required when searching for "Forceps AND Complications":

a) General Surgical complications – associated with the use or loss of forceps during operations.
b) Complications of labour and delivery that require the use of forceps.

It can be seen that in case a) the forceps cause the complication, but in case b) they are used in an attempt to avoid it. Assuming that the membership value of relations involving "forceps" remains at the initial value of 0.5, and the value for each of the relations involving "complications" is around .09, then the Fuzzy Min relation returns a value of 0.09 for documents containing "forceps" AND "complications" in the obstetric portion of the ontology. This is the case where the ontology membership values have not been altered since instantiation. However, if the ontology being used has been modified, so that forceps is considered to have a membership value of .8 for the obstetric portion of the hierarchy, and complications has .5 for the obstetric portion then the membership value of the conjunction is now .5, for the obstetric domain. If the fuzzy ontology approach is not used, then it is still possible that a query of "forceps AND complications" will return many obstetric results, however if the desired results are not obtained, it is not certain as to which sets of related terms should be used for query expansion and refinement. In the case of the fuzzy ontology approach, the query expansion goes to related terms – such as assisted delivery, vacuum extraction etc.

3.5.2 Fuzzy OR Operator

The fuzzy "OR" relation takes the maximum value of the membership value of its elements, as the combined membership value. In an information retrieval system, the use of fuzzy operators can be combined with a threshold and/or a weighting system in order to assign relevance scores. The effect of this is to increase the importance of "certain" i.e. high membership value terms in a compound query. This means that any automatic query expansion is in danger of being biased towards these terms if memberships are translated as weights for relevance scores. The fuzzy ontology approach uses these membership values. As an example, consider a query with 3 terms with different membership values associated with them in different locations - "obstetric" (μ=0.5,0.5), "delivery" (μ=0.3,0.3,0.4) and "theatre" (μ=0.2,0.8). Firstly, the terms are examined to discover the likely region of the ontologies. This is done by examining the various possible locations of the terms in the ontologies. For each term-location pair the parents, siblings and children of the term in each location are examined. Siblings are included in this calculation because even if siblings represent orthogonal concepts they still relate to the same area. The system then selects an area one level further out – e.g. grandparents, grandchildren aunts/uncles and nieces/nephews (i.e. siblings of parents and children). This process continues until all of the terms in the query are identified.

Table 17: Truth table for OR

A	B	A Or B (A V B)	Membership value of combined terms
1	1	1	Max (μ_A, μ_B)
1	0	1	μ_A
0	1	1	μ_A
0	0	0	0

The score for a particular location in the ontology is then given by the membership value of the original term plus the sum of the membership values of the nearest terms divided by the network distance to each of those terms. This value is given as the support for this location.

$$Support(t,i) = \mu_i \sum_{1}^{n} \mu_j \Big/ D(i,j)$$

Equation 6

Where μ_i is the membership value of the term t at that location i, μ_j is the membership value of the other terms 1..n at the closest location to the location i, and D(i,j) is the network traversal distance i.e. number of edges within the ontology that are needed to travel from i to j. The support is calculated for each term in the query, and the maximum value selected. This maximum value is known as the *centre of the query*, and the query expansion takes place on the basis that expansion terms are chosen which are adjacent to locations nearest to this centre where terms are ambiguously located. To return to our example, it may be discovered that the combination that has the greatest support is centred on the term “Theatre” which had the membership value (0.2).

In practice this means that non-ambiguous terms in queries will tend to have the greatest influence on the location of the centre of the query and that distant modifiers will not have much influence at all. Rather than this “winner-takes-all” approach another method would allow a number of centres to be used, with the relevance score – however calculated – for each recovered document to be multiplied by the support for each centre. Integration of the fuzzy ontology approach to an intelligent searching system is described in Chapter Five.

Chapter Four

This chapter deals with the use of information entropy for the characterisation of documents, and identification of attributes. The aim of this work to investigate methods of discovering similarities between documents, such as language, authorship origin and genre that are applicable to different types of documents. Section 4.1 deals with the background to this approach, including previous work. Section 4.2 describes the use of an online learning archive for testing this approach. Section 4.3 deals with the generation of corpora from the web, web domain identification and the derivation of the Kolmogorov distance. Section 4.4 deals with corpus segmentation in the context of the BMJ topics collection. Section 4.5 describes clustering using Kolmogorov distance. . The majority of the experimental results in this chapter are contained in sections 4.2, 4.3, 4.4 and 4.5.

4.1 Introduction

One intriguing approach to language identification in documents, and by extension, the identification of relationships between languages was reported in 2002 (Benedetto, Caglioti, & Loreto, 2002). This work is based on the characterisation of documents by means of the information entropy contained in them. However, the novelty of this work lies in the fact that a commonly available compression algorithm – the "zip" algorithm is used in a fairly simple scheme.

Briefly, the "zipping" method is based around the concept of the relative information entropy of a document. The concept of information entropy was introduced by Shannon (Shannon., 1948). One way of expressing this concept is to view a document as a message that is being encoded over a communication channel. A perfect encoding and compression scheme would produce the minimum length of message. In general, a document that can undergo a high degree of shortening by means of a compression algorithm has a low information entropy – that is there is a large degree of redundancy, whereas one that changes little in size has a high degree of information entropy, with little redundant information. A good compression algorithm should never increase the size of the "compressed" document. As the authors of (Puglisi, Benedetto, Caglioti, Loreto, & Vulpiani, 2003) point out, a good zipping algorithm can be considered as a sort of entropy meter. Compression approaches can be different for different requirements, for example the joint picture group .JPG format for images and the various codecs for sound. Compression algorithms can be “lossy” or “lossless”, depending on whether information is lost during the process but for text documents, a lossless approach is almost always appropriate. The only exception might be text summarisation or snippet production.

The Lempel-Ziv algorithm reduces the size of a file by replacing repeating strings with codes that represent the length and content of these strings (Ziv, 1977), and has been shown to be a very effective scheme. To work efficiently, the Lempel-Ziv algorithm “learns” effective substitutions as it examines the document sequentially to find repeating sequences that can be replaced in order to reduce the file size. For example, if the phrase “this phrase repeats” is used a number of times in a document the document will be compressed in size if it is replaced by a marker that refers to the original phrase in one location, but every other occurrence just has reference to this value. I use the term “motif” to refer to such replaceable sequences in the rest of this document. This algorithm is the basis of the popular and rapid zip software in its various incarnations including Gzip, Pkzip and WinZip. This method relies for compression on finding as many common replaceable strings within a document or file in order to achieve maximum efficiency. By adding a document of unknown characteristics to one of known properties (for example language, author, genre etc.) then is suggested that the combined relative entropy is smallest when the two documents are most similar, as in that case the number of repeating motifs in the combined document will be highest, and hence the efficiency of compression will be highest and the relative entropy will be lowest. The procedure is shown in Figure 30.

In their paper (Puglisi et al., 2003) the group made the observation that if the second document was between 1-15 KB, with a first file length of 32-64 KB performance was acceptable in terms of accuracy of classification. The authors were using Unicode formatted files, so each character was represented by 2 bytes in the file. The explanation for this result was that the algorithm was effectively being trained to identify particular replaceable strings from the first document, if the second document was too long then the algorithm was learning replacement strategies based on the second document.

Further recent work by the Benedetto group has used simulated documents to identify the optimal ratio of lengths between the first (known) and second (unknown) documents (Puglisi et al., 2003). The motivation for this work was to discover the optimal ratio between the first and second document for a maximal value of the relative entropy and hence maximum discrimination between sources. A number of simulation experiments were performed using strings generated by Bernoulli schemes, Markov Chains and Lozi Maps (a set of functions that can be used to generate strange attractors).

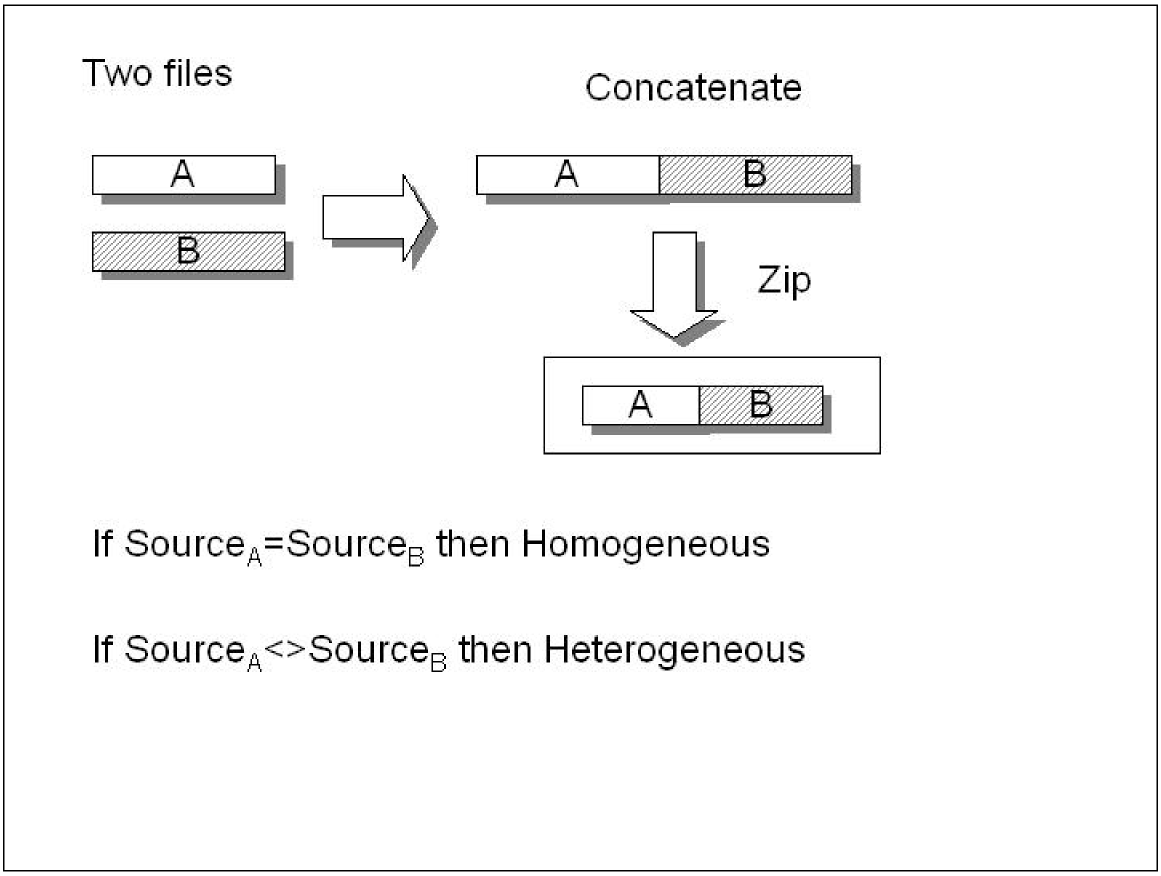

Two files
A
B
Concatenate
A
B
Zip
A
B
If Source$_A$=Source$_B$ then Homogeneous
If Source$_A$<>Source$_B$ then Heterogeneous

Figure 30: The concatenation process

These experiments supported the hypothesis that there was a "crossover point" where the increasing length of the second string causes a reduction in the relative entropy. This point was identified as occurring when the initial length (L_A) is 10,000 Bytes and (L_B) is 1300 Bytes (see Figure 31 for a graphical demonstration of these identifiers).
However, for reasons explained in section 4.2 below, such an approach was not adopted in this work, and similar length known and test strings were normally used. These would still have variances in the relative entropy because of the relative inefficiency of having to select from two groups of motifs when the documents are heterogeneous, as opposed to the situation when the documents are homogeneous.

4.2 Author Identification

Authorship identification remains an important issue for the verification of electronic documents. Computers have been used for a long time to try and verify the identity of authors in the humanities (Sedelow, 1970) and in the field of software forensics (Oman & Cook, 1989). Various techniques have been used in the past including Bayesian

Inference (Mosteller F. & Wallice D., 1964), Neural Networks (Singhe, 1995) and more sophisticated methods using Support Vector Machines (Vel, Anderson, Corney, & Mohay, 2001). However, such approaches tend to be extremely language- and context-specific although often very effective.

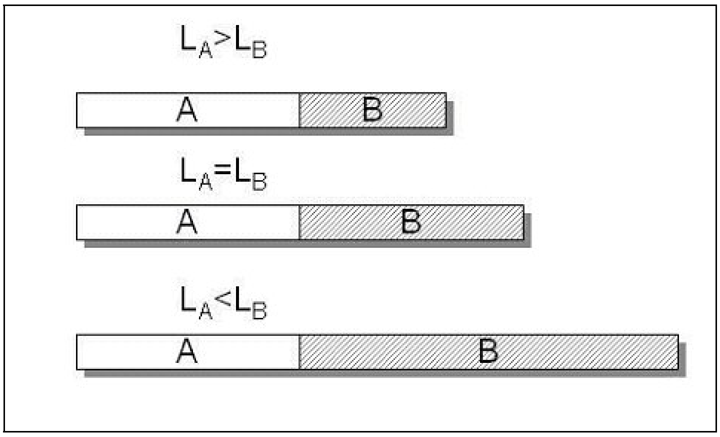


Figure 31: Length of files comparison

The aim is to characterize sections of electronic text that will allow collaborative searching via conventional search engines and other data sources. Often this data is in small fragments and without a consistent global structure with no use of <Author> tags for example.

The work of (Benedetto et al., 2002) discussed above demonstrated that it was possible to identify the language used in a document by comparison with known documents. The authors also briefly mention the use of this technique to identify authorship, but for a limited set, and do not provide much information about the characteristics of the documents used. Interestingly the authors pointed out that such analysis is not dependent on the meaning of the data stream but purely on its information content. This implies that any source of data that can be sequenced, e.g. text, video or DNA can use this sort of technique. In terms of electronic documents, this means that such an analysis not only does not require subject domain specific rule sets, but also will include stylistic features, such as punctuation and control characters that are not considered in semantic method of analysis. This method is therefore complementary to other methods that concentrate on the understanding of the document, much as handwriting or voice analysis widens the possibilities of author identification, even if the content is not distinctive (Srihari & Lee, 2002).

Various corpora of texts are used for the testing of textual analysis techniques, for example the British National Corpus (Oxford University Computing Services, 2001). However the choice of test corpus depends on two aspects – the appropriateness of the corpus, in terms of the similarity to the target data, for example in the vocabulary, style and length of the documents, and the usefulness in terms of the ability to achieve reliable results e.g. the presence of known authorship and the avoidance of artefact etc. One particularly rich source of information is archived newsgroup and list server postings that often contain particularly relevant information in a concise format. Newsgroup postings provide a rich corpus of material to study classification schemes – for example the use of readability or other scores (P. Sallis & Kasabova, 2000) to characterize discussion. There are disadvantages to this approach however – in particular the presence of quoted text from previous postings and the possibility of concealed authorship via anti-Spam devices.

In this study, data from an online teaching system discussion group was used. The reasons for using this corpus are that; it was readily available, the authorship was identifiable, the message length is variable but always below 7000 characters so that large numbers of files are manageable and the system did not add any formatting to the text. In addition because of the large number of messages available this corpus was suitable for development work. In this system, Business Online (BOL) students and staff members in the faculty of business at Auckland University of technology (AUT). BOL has been in use at the business faculty of AUT since 2000 (A. Sallis, Carran, & Bygrave, 2000). It is a WWW-based system, designed to support rather than replace the face-to-face process. BOL is primarily a communication tool following the philosophy of electronic collaborative groups (Fåhræus, Bridgeman, Rugelj, Chamberlain, & Fuller, 1999). Features include synchronous and asynchronous text-based communication, file sharing, and rostering and project management tools. Asynchronous communication occurs via a threaded discussion tool while file sharing is done via set of online folders.

Both the file system and the asynchronous discussion (known as the forum) include logs of who uploaded and downloaded the objects and when this occurred. This information is displayed by the system, allowing users to identify who has read their message or downloaded their file. Messages posted in the forum cannot be edited or deleted by students after posting and anonymous posting is not allowed. Users are identified as students or staff and only have access to those parts of the system they have been enrolled in.

For the purposes of this study, parts of the asynchronous discussion area were used. This data is an appropriate corpus for examining discussions areas such as newsgroups as it includes messages of varying length, with diverse content. The data is especially useful as a testing set as BOL does not automatically format the messages in terms of layout or spell checking and does not include previous messages automatically. Authorship is identified unambiguously by means of the login process. All work was done on data that had identifiable authorship indicators removed and replaced by a randomly assigned author identification number.

4.2.1 Methods used for validation of the technique

All the messages from semester 1, 2002, BOL were initially screened. There were a total of 10,086 messages in this data set. Messages longer than 100 characters (including spaces and punctuation characters) were selected and the rest discarded. The data was stored in an SQL Server 2000 Database (Microsoft). A database utility was used to convert the data from 16 bit Unicode formatting to 8 bit ASCII formatting to allow easier manipulation and storage. The original distribution of the message length is shown in Figure 32 - the system did not allow posting of messages more than 7,000 characters long. The message authors were then checked and those with more than 20 messages (range 20-155) in the remaining set were selected. All messages from other authors were deleted. This set was further refined during the experiment

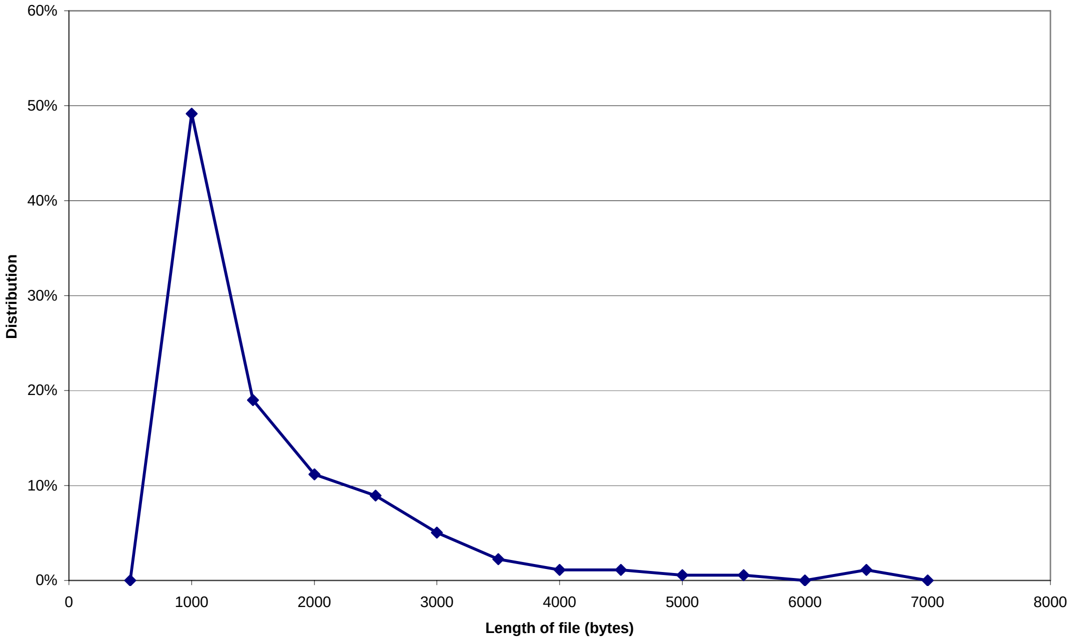


Figure 32: Distribution of file lengths

The software developed to test this approach read the message data from the database as a string (F_1) and then concatenated it with another message string (F_2) as determined by the experimental protocol to create a combined file (F_c). The length of F_c (in bytes) was calculated, giving the initial file length (L_c). This string was then zipped to give a zipped file (F_Z). The length of this file was then calculated, giving (L_z) in bytes. The efficiency of the zipping (Z_{eff}) was calculated as:

$$Z_{eff} = {L_z}/{L_c}$$

Equation 7

During the experiment, many different F_2's were combined with each F_1 and the different Z_{eff}'s were then sorted into ascending order. The combination of files that produced the lowest value of Z_{eff} was designated the closest match. The values of Z_{eff} varied between 0.207 and 0.531 with a mean of 0.431, but this value is never used in the analysis, only the relative ranking of the Z_{eff} compared to other Z_{eff}'s is used.

All the messages combined with themselves (i.e. F_c=F_1 & F_2, where F_1 =F_2 gave the lowest value for Z_{eff}, hence these combinations were excluded from the analysis.
Initially, each message was combined with 10 other messages, only one of which was by the same author. The "same author" message was chosen randomly from the collection. No allowance was made for differences in file length between the messages but the initial message (F_1) was always at least 3,000 Bytes long. The maximum length of F_1 was 7,000 Bytes, mean 4,625 Bytes. This length was chosen to reduce the number of messages studied and exclude extremely short ones. Similar figures for F_2 length in this experiment were – minimum 1,006 bytes, mean 1,837 bytes

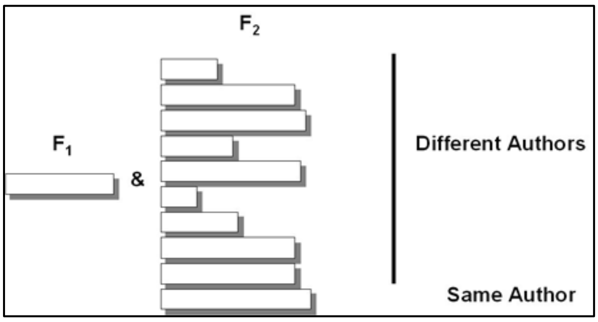


Figure 33: First experimental protocol

and maximum 7,000 bytes. A total of 149 F_1 files were selected. The protocol is shown in Figure 33.

Because the number of compared documents was deliberately limited, and this method depends on ordering of results, conventional information retrieval tools for recall cannot be used. Instead, if the Z_{eff} when both F_1 and F_2 had the same author was lowest, this was scored as a correct classification.

The overall success rate of the technique in protocol 1 was 24% over 149 groups. That is to say 24% of the time the minimum Z_{eff} was achieved when the first author was the same as the second author. By chance one would expect this to occur only 10% of the time. A Chi-squared test comparing the observed occurrences of lowest Z_{eff} where Author1=Author2 compared to the expected occurrences due to chance showed significance with $p<0.001$. It appeared however that if the second file was a comparable size to the first one then correct identification became more likely. In the 7 groups where the second file had the same author as the first one and the file length was 3,000 bytes or more, all of the minimum Z_{eff}'s were assigned the Author1=Author2 case. No correction was made as to whether the messages were in the same discussion area, but they were selected at random from all discussion areas.

In order to investigate this effect further, the rank of the Z_{eff} was plotted against the length of the second message (Figure 34), without regard to the authorship. It can be seen that the longer the second message is, the more likely it was to produce a lower ranked Z_{eff}. In this experiment there was no minimum length of F_2 imposed.

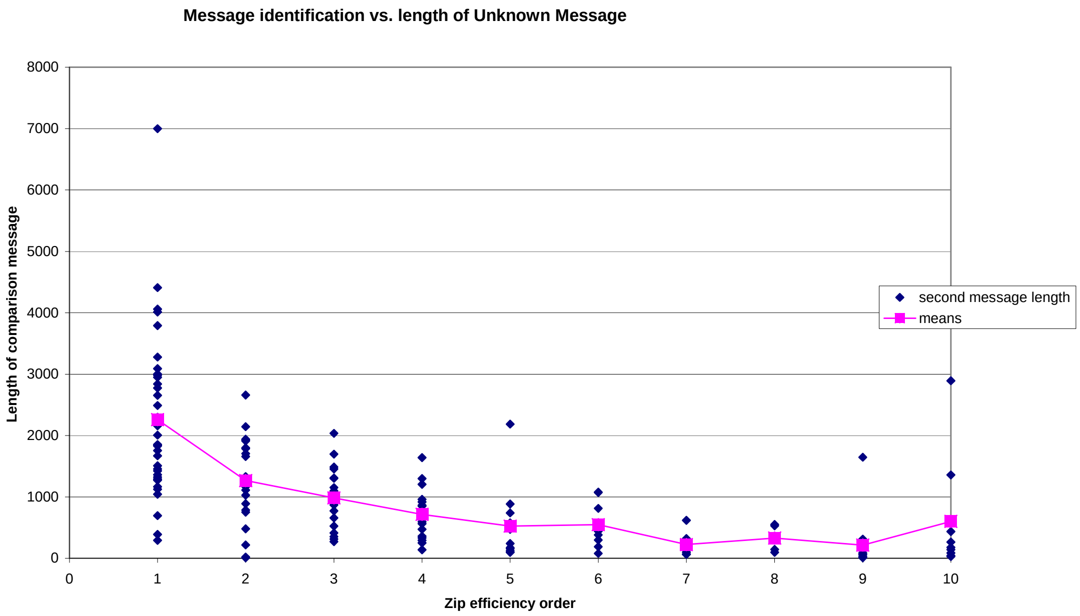


Figure 34: Distribution of order versus second length

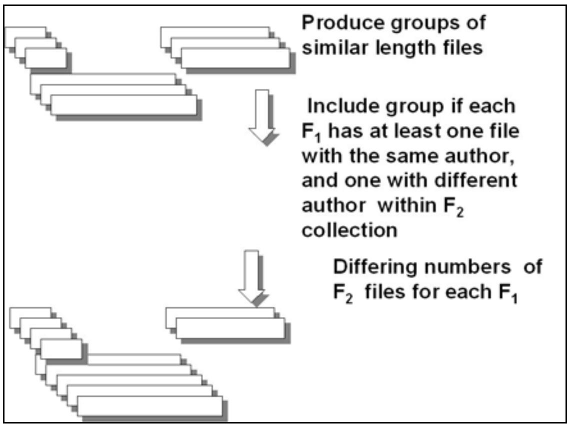

Produce groups of similar length files
Include group if each F_1 has at least one file with the same author, and one with different author within F_2 collection
Differing numbers of F_2 files for each F_1

Figure 35: Second experimental protocol

To mitigate this effect, the second experiment (see Figure 35) selected files that were of similar length. In this case files were selected that were no more that +/- 5 bytes different in length to each other. This group was then reduced by setting up a similar scheme as used in the first experiment where each file was compared with files of a similar length. Any file could be selected as long as there was at least one message from the same author of a similar length. There were 25 different F_1 Files used in this protocol and 618 combinations (Fc) were created, the minimum length was 3,003 bytes, maximum 7,000 bytes, mean 6,572 bytes.

With the second experiment in 14 of a possible 25 cases (56%) files with the same authorship had the lowest Z_{eff}. As a whole only 36 of the 618 (5%) combined files (F_c) had author1=author2. Again, Chi-squared shows that this result is significant at the $p<0.001$ level. This result is not conclusive as some messages were combined with more than one message by another author, however it assisted with the development of the length-selecting approach.

Another experiment then took place, combining the two approaches above. Each first message was matched with ten other messages of similar length, only one of which was by the same author.

This method of author identification appears to produce results that are better than chance, especially in the case where the two files are of similar length. No attempt has been made to correct for aspects of the files that may affect Z_{eff} apart from file length, so the general applicability of this result may be limited. Nevertheless this appears to be a technique that may have application in conjunction with others in improving automatic author identification. Certainly it would appear that with the correct combinations of circumstances, this approach could assist in the classification of electronic texts along with current methods. One advantage of this method is that it should work on texts in any language and be robust with respect to misspellings and stylistic quirks. Indeed such characteristics may increase the value of this method.

Table 18: Results of author analysis

Count	Status	Percentage
114	Author1<>Author 2	71.25%
46	Author1=Author2	28.75%
160	**Total**	

Using Chi-Squared, this result – shown in Table 18 - is significant at the p<0.001 level (SPSS 11) χ^2=(1,N=160)=62.5,p<0.001. The proportion of messages with common authors having the smallest distance is a great deal higher than expected by chance.

4.3 Web domain identification via zip method

This technique was applied to Web Pages in the hope of identifying both authorship (which may be institutional or individual) and in terms of style or genre. It was hoped that the addition of information from the structure of Web Pages including both tags and white space for example will allow a stylistic corpus to be developed, which in turn will allow greater efficiency of searching by allowing for stylistic information to be included in queries along with keyword or content-based systems. As the authors of “The Art of the obvious” (Nygren, 1992) point out, user interface design for electronic material often gives less information to the user than they expect for paper documents, so some sort of ‘style similarity’ measure may be helpful in this context. For example, a “Coffee Table Book”, is easily differentiated from a research paper in a paper library, by means of physical clues – size, appearance etc. This allows browsing to be directed effectively to likely useful information sources. In the non–physical world of electronic documents, such clues are not available and other approaches are needed.

4.3.1 The Web “Spider”

“Spider” is the name given to programmes that automatically download WebPages for caching or analysis. This method of retrieving pages is very useful for developing a real world, if unstructured corpus. Search engines are very large users of these devices, with many millions of pages being visited every day. The timing of visits to sites depends on the expected lifetime and stability of the site, by visiting sites regularly; the spider software (so called because it travels around, and weaves together the web) calculates the mean lifetime of pages in a particular domain, and times its visits appropriately. Some particular sites, for example weekly or daily publications can have particularly tailored visiting patterns set up. Some sites – such as the IEEE -particularly request that spiders or web bots do not visit their pages. The reasons for this may be that such visits increase the workload or costs associated with the server hosting the site particularly if the owner is charged for sending datal via the internet, that material is not suitable for general use, or that they do not wish to be indexed by search engines because the site is dynamic and better searched via internal tools. Other sites, such as PubMed may provide other means of harvesting web pages. Others again, such as the BMJ, have a minimum time between page downloads that is enforced by the server. “Well behaved” spiders are required to obey any rules and in particular, to avoid downloading pages tagged with the HTML 4.01 standard “ROBOTS” <META> tag. (W3C, 1999). The site may also have a robots.txt file that includes the same information. E.g.

```
<META name="ROBOTS" content="NOINDEX, NOFOLLOW">
```

The spider built for this work is particularly simple, initially based on the Microsoft Internet transfer control, and then on the more sophisticated web client module provided with .NET. Because of the limited error and status reporting associated with the Microsoft Internet control, the spider downloaded the target page into a web browser control at the same time. The web browser control, which provides more error and status information was used to confirm the download had completed. However it was rare that the spider downloaded more than 2000 pages without failing, and needing resetting.

The spider that downloaded the pages was instructed to download all page links from the original pages using the Document object model (DOM, described in more detail in section 4.3.2) to identify HREF type links that represented other pages, i.e. not media files or “mailto” - type statements. These recovered pages were then scoured for links and the process continued until no valid pages could be recovered, or the process had continued for more than 1 hour. If a page was not in the original domain then it was downloaded, but links from it were not. That is, pages that retained at least the organisation, country and type identifiers were downloaded. PDF files, word and PowerPoint files were not downloaded normally, but some work was done with PDF files during the RCOG/RCS work (see section 6.1).

4.3.2 Corpora generation – the web corpus

4,389 Web Pages were downloaded using the spider described in the previous section from the following root sites shown in Table 19. These sites were selected to get a range of topics and genres.

Table 19: The Web Corpus Description

Domain label	Root	Description
AUT	www.aut.ac.nz	The AUT university main website
Guardian	www.guardian.co.uk	The Guardian Newspaper
OBGYN	www.obgyn.com	OBGYN a site for professionals and patients, based in the USA
Microsoft	www.microsoft.com	The Microsoft Website
HON	www.hon.ch	Health on the Net – Information for both consumers and practitioners
Apple	www.apple.com	Apple Computer

Some pages produced errors, or represented script-only pages or small frames and these were excluded producing a total of 3,678 Pages. Table 20 shows the number of files downloaded from each site and the mean file length.

Table 20: Domain Characteristics

Domain Name	Number of Pages	Average File Length (bytes)
AUT	2192	58518
Guardian	234	38326
OBGYN	203	25937
Microsoft	442	882771
HON	19	21600
Apple	588	37319
Total	**3678**	**177411.8**

In addition to the file length, the following objects from the Document object model (DOM) were counted for each page:

- Link count
- Form count

- Paragraph count
- Image count
- Frame count – this applies to the “parent” frame only. Subframes were not identified.
- List count

The identification scheme via zipping was similar to that described above for authorship analysis. AUT, Microsoft, OBGYN and Guardian each had 20 pages randomly selected, with varying file lengths. These were then matched with 9 similar length files from other domains and one from their own. The algorithm for doing this was:

```
For each original file
 Initialise difference ( to 100 Bytes)
        Do while not selected
        Select a second file with length between (original length+difference) and (original length-difference) and not already used
 If second file found, add to final, exit loop
Otherwise, increment difference and loop
Select next file
```

The pairs were then zipped separately and then concatenated and zipped together. The relative efficiency of the zipped combined file was then calculated.

In all 800 pages were used in the analysis. In this case using the Chi-squared test this gives an odds ratio of 11.2 (3.85 <OR<35.02) with a p value of < 0.001. This indicates that files from the same domain are 11 times more likely to have the lowest Z_{eff} than those from a different domain. Note that in this case the WWW pages HTML code was used for the comparison, not only the text.

Table 21: Results Of The Web Corpus Analysis

Count	**Status**	**Percentage**
45	Website1<>Website2	56.25%
35	Website1=Website2	43.75%
80	**Total**	

Again, using Chi-Squared, this result is significant at the p<0.001 level (SPSS 11) χ^2=(1,N=80)=101,p<0.001. The proportion of websites with common authors having the smallest distance is a great deal higher than expected by chance.

4.3.3 Using the Kolmogorov distance method with the zip algorithm.

Recalling the formula for distance shown above:

$$D(A,B)=\frac{C(A|B)+C(B|A)}{C(A)+C(B)}$$

By using the Zip program, the compressed length can be calculated, but the compression dictionary is not available.

However, in the same way as described above, concatenating files and then zipping then allows the zip algorithm to develop its dictionary on the first file and then apply it to the second. The algorithm used is given by:

- Obtain the two files – file_1 and file_2
- Concatenate them in two ways, file_1+ file_2 = (file_{12}) and file_2+ file_1 =(file_{21})
- Calculate the zip length of:
- file_1 as zip_1
- file_2 as zip_2
- file_{12} as zip_{12}
- file_{21} as zip_{21}

The distance (D) is then given by:

$$D(file1,file2)=\frac{(zip12-zip1)+(zip21-zip2)}{zip1+zip2}$$

Equation 8

where {zip_{12}, zip_{21}}> {zip_1, zip_2 }so that the value of D is always positive. This approach assumes that the dictionary is stable after the compression of the first document, which in turn is affected by the relative sizes of the dictionary and document.

4.3.4 Message comparison

For this experiment, the business online data was used again. 16 non - staff users were selected who had posted more than 10 messages on the system within semester 1 2003. Previous experimental work suggested that messages longer than 1,000 Bytes were required for reasonable comparison, and that messages should be concatenated with messages of similar length. This resulted in a potential pool of 1,253 messages. For each user 10 messages were selected and combined with 1 message from the same user – which was not the same as the initial message - and 9 from other users. The concatenated files were then zipped, and the relative size of the zipped files and the original files were compared.

Compared with the results above, the Kolmogorov method performed even better. With a total of 160 comparisons the results are shown in Table 22.

Table 22: Kolmogorov Distance for Messages

Count	Status	Percentage
83	Author1<>Author2	51.88%
77	Author1=Author2	48.13%
160	**Total**	

Again, using Chi-Squared, this result is significant at the p<0.001 level (SPSS 11) χ^2=(1,N=160)=258,p<0.001. The proportion of messages with common authors having the smallest distance is a great deal higher than expected by chance.

4.3.5 Website comparison

The same comparison was done for the website domain-based group shown in Table 23. Again, using Chi-Squared, this result is significant at the p<0.001 level (SPSS 11) χ^2=(1,N=80)=451,p<0.001. The proportion of websites from common domains having the smallest distance is a great deal higher than expected by chance.

Table 23: Kolmogorov Distance for Domains

Count	Status	Percentage
15	Different Websites	18.75%
65	Same Websites	81.25%
80	**Total**	

4.4 Corpus segmentation by Kolmogorov distance

This section deals with attempts to validate the Kolmogorov distance in terms of a fairly homogenous corpus, where only topic, and not author, language or genre is different. In the initial experiment, described in section 4.4.2, the web pages being examined are used “whole” that is as HTML downloaded from the WWW, in the second described in 4.4.3, the text is extracted from the pages. The “whole page” technique has a great advantage in simplicity, in that local processing is minimised as no page parsing or analysis needs to take place. In addition, should the corpus be composed of pages that differ more in structure than in text, such an approach may be preferred, in particular if there is, as seems likely a lower limit to the file size which can use the Kolmogorov distance, then the HTML code provides a useful degree of expansion to the files in question. There are three major disadvantages of this approach:

- The data contained in the HTML structures is sufficiently dissimilar to that in the text elements to allow different motifs to be identified. As HTML is written using the same tags no matter which readable text language is being used, this may have the effect of “washing out” language differences. Common tags such as <p> or <br> and their equivalent closures may well be deemed replaceable by the algorithm, but these occur in virtually all HTML documents.
- Stylistic similarities between documents originating from the same author, HTML editor or conversion program may overwhelm the differences between them.
- Because an HTML document has a symmetrical structure, at least in terms of opening <> and closing </> tags, the documents cannot usefully be truncated in order to achieve a ratio between the first and second document which may be desirable in terms of the experimental work described in section 2.8. However, it can also be argued that real documents have characteristic structures that are not amenable to arbitrary truncation.

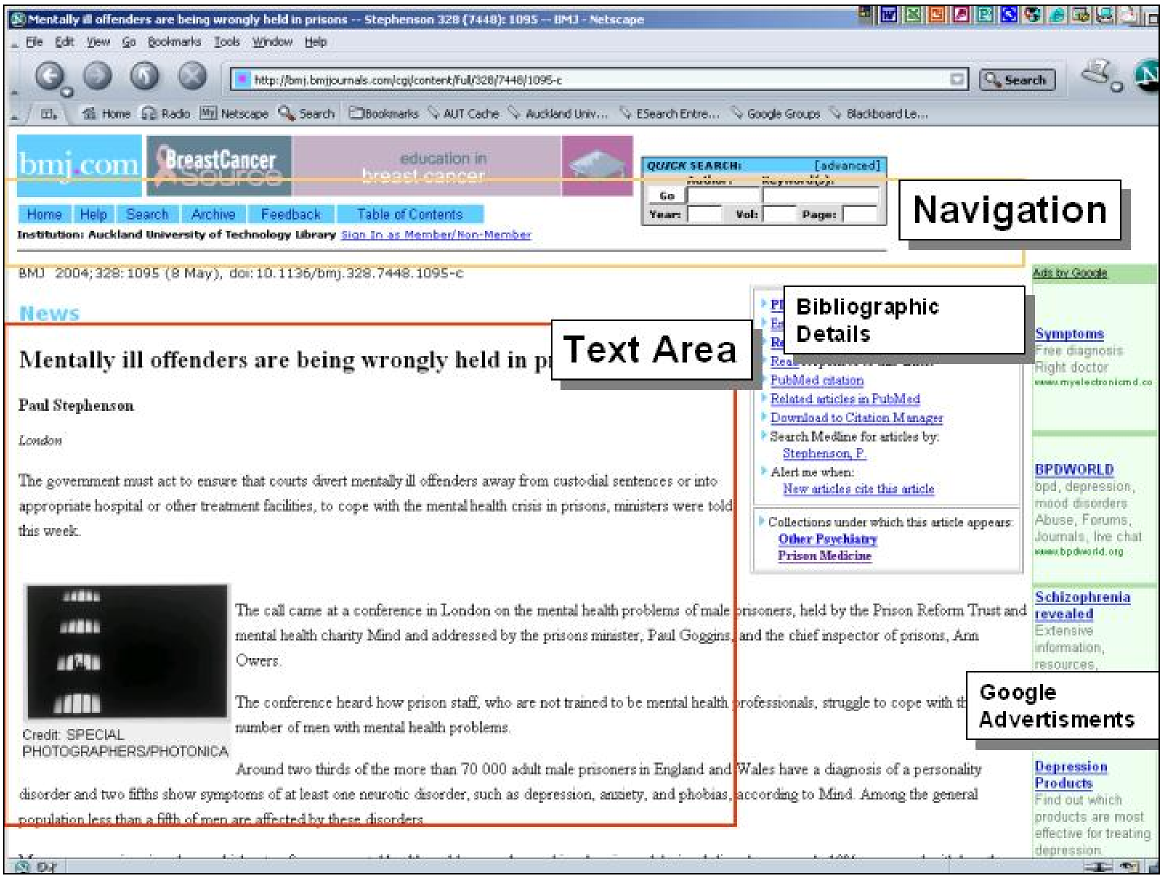
4.4.1 The British Medical Journal Corpus

The British medical journal has an extensive corpus of material located on its website, under the title of “Topic Collections” at http://bmj.bmjjournals.com/collections/. These collections comprise Web Pages representing peer-reviewed papers, letters, news items, editorials and other types of item relating to a particular area of interest. A spider was created to recover a maximum of 20 HTML pages from a number of these sites, for further analysis. Details of the numbers of terms found in each file title that was recovered are shown in the appendix in Table 65. The tiles were generated from the root URL of each topic segment e.g.

bmj.bmjjournals.com/cgi/collection/10_minute_consultations

=10_Minute_Consultations.

Example pages are shown in Figure 36. When the pages were downloaded, only the Base URL of each page was downloaded. Although the HTML code of the page is quite complex, each page is similar in some elements of structure. There is a fairly consistent writing style in the BMJ, as can be seen in the instructions for authors. Selection of articles as being worthy of inclusion in these specialist topic collections would also tend to enforce a consistency of writing style, and UK English is used throughout. As the BMJ itself produces the HTML versions of the articles, there is likely to be consistent HTML style as well as text.



Mentally ill offenders are being wrongly held in prisons -- Stephenson 328 (7448): 1095 -- BMJ - Netscape

http://bmj.bmjjournals.com/cgi/content/full/328/7448/1095-c

Institution: Auckland University of Technology Library Sign In as Member/Non-Member

BMJ 2004;328:1095 (8 May), doi:10.1136/bmj.328.7448.1095-c

News

Mentally ill offenders are being wrongly held in pr

Paul Stephenson

London

The government must act to ensure that courts divert mentally ill offenders away from custodial sentences or into appropriate hospital or other treatment facilities, to cope with the mental health crisis in prisons, ministers were told this week.

Credit: SPECIAL PHOTOGRAPHERS/PHOTONICA

The call came at a conference in London on the mental health problems of male prisoners, held by the Prison Reform Trust and mental health charity Mind and addressed by the prisons minister, Paul Goggins, and the chief inspector of prisons, Ann Owers.

The conference heard how prison staff, who are not trained to be mental health professionals, struggle to cope with th number of men with mental health problems.

Around two thirds of the more than 70 000 adult male prisoners in England and Wales have a diagnosis of a personality disorder and two fifths show symptoms of at least one neurotic disorder, such as depression, anxiety, and phobias, according to Mind. Among the general population less than a fifth of men are affected by these disorders.

Collections under which this article appears:
Other Psychiatry
Prison Medicine

File Edit View Go Bookmarks Tools Window Help
Search
Home Radio Netscape Search Bookmarks AUT Cache Auckland Univ... ESearch Entre... Google Groups Blackboard Le...
bmj.com
BreastCancer Source
education in breast cancer
QUICK SEARCH: [advanced]
Go
Year: Vol: Page:
Home Help Search Archive Feedback Table of Contents
Navigation
Text Area
Bibliographic Details
PubMed citation
Related articles in PubMed
Download to Citation Manager
Search Medline for articles by:
Stephenson, P.
Alert me when:
New articles cite this article
Ads by Google
Symptoms
Free diagnosis Right doctor
www.myelectronicmd.co
BPDWORLD
bpd, depression, mood disorders Abuse, Forums, Journals, live chat
www.bpdworld.org
Schizophrenia revealed
Extensive information, resources,
Google Advertisments
Depression Products
Find out which products are most effective for treating depression.

Figure 36: A BMJ web page

4.4.2 Kolmogorov distance calculation for whole pages

The calculation of Kolmogorov distance by zipping method as described in section 2.10.3 was used in order to attempt to identify the home domain of each page from the BMJ corpus. In this experiment, the whole page structure was used using pages in HTML format. This analysis includes the formatting tags as well as the various <INCLUDE> and <HREF> statements, although the linked or included files were not themselves incorporated.

The process began by selecting those home domains i.e. topic areas, as shown in appendix that had at least 5 valid pages downloaded. For each of these valid domains (n=133), 5 initial pages were chosen. One page from the same domain, and nine different pages from other domains were then selected, in a similar manner to that described above. Again, the files were selected to be of similar length, and the pages zipped together, using the Kolmogorov distance by zipping algorithm. Self-comparison i.e. where file1=file2 was not permitted. The results are shown in Table 24.

Table 24: Kol–distance BMJ same length

Source		**Number of occurrences with shortest distance**
Different topic domain		119
Same topic domain		546
	Total	665

Using CHI-Squared implemented in SPSS version 11 the results show that the minimal distance is significantly more likely to occur using files from the same domain, rather than ones of similar length from other domains. χ^2=(1,N=665)=3839,p<0.001

4.4.3 Topic domain identification using text only

Previously (in section 4.2) I have discussed the simulation experiments that appear to show that a ratio of file 1 to file 2 in the range 10:1 to 10:2 appears to give the best resolution of file source. However, the problems of heterogeneity and common structure between the raw HTML documents preclude simple truncation of the second file to achieve this ratio. The truncated file 2 may just contain header and style HTML code, which is almost identical between documents in this corpus, which would not allow differentiation between different topic groups because there is literally no actual topic content in the comparison. Merely choosing shorter documents for use as file2 is also problematic, because it can be seen from observing the structure of the HTML files that shorter documents have more HTML and less text. In order to check the hypothesis that file 2 should be considerably shorter than file 1 for maximum resolution, the displayed text parts of the HTML documents were extracted into other files, and these files were used for the comparison.

The text extraction process used created files representing the text that would be visible to a user viewing the file downloaded in a browser. This is referred to as “visible text”. Text contained in images, in ‘include’ files that are displayed, or from such sources as RSS are not able to be analysed in this way. The process, built using a function written in VB.net treats the source HTML as a file and includes all material not contained in <> brackets. For later analysis some of the formatting information is included, for example special characters like &nbsp (non breaking space) are interpreted and represented in the file. Large runs of blank space are reduced to single spaces, which although it may be a stylistic feature may also be a quirk of the HTML editor.

The first experiment using this approach attempted to compare "visible text" files of similar length. In order to do this the similar length algorithm of 4.3.3 was used. 655 comparison sets were used, i.e. 655 different instances of file1. For each of these comparisons, one file from the same topic area as file 1 was used and nine which were from other topic areas. Rather than always use different files for the second file the algorithm selected second files so that:

- If Topic(File 1)=Topic(File 2) – the homogenous case - then file1 must not be the same as file 2
- Generally, the file 2 that was nearest in length to file 1 was selected, even if this means reusing the files, up to a limit of 10 occurrences.
- If Topic(File 1)<>Topic(File 2) – the heterogeneous case- then each file 2 must be different and come from a different topic group.

The results are shown in Table 25. This appeared to support the idea that the text extraction process actually influenced the information entropy of the resulting document sufficiently to remove detectable differences. There are 656 results rather than 655 because in one case there were two equal minimum Kolmogorov distances.

Table 25: Topic Kol-Distance

Source		**Number of occurrences with shortest distance**
Different topic domain		423
Same topic domain		233
	Total	656

Again using SPSS χ^2=(1,N=656)=490,p<0.001

At first sight, this result seems paradoxical at best – the visible text can be thought of as the reason for having the document in the first place, and the text–only messages studied in section 4.2 have a better performance at least in terms of author identification. However, it should be remembered that the text-only messages are a product of human thought and were intended to be viewed in the format studied, whereas the visible text files have been parsed by machine, from originals that included non-text formatting controls that included emphasis or highlighting elements such as italics, headings etc. The topic-based corpora Although the Kolmogorov system does not understand such elements, the link between a key phrase and the tags that identify it as such may be significant, i.e. “diabetes” may be common, but “<h1>diabetes</h1>” may be rare.

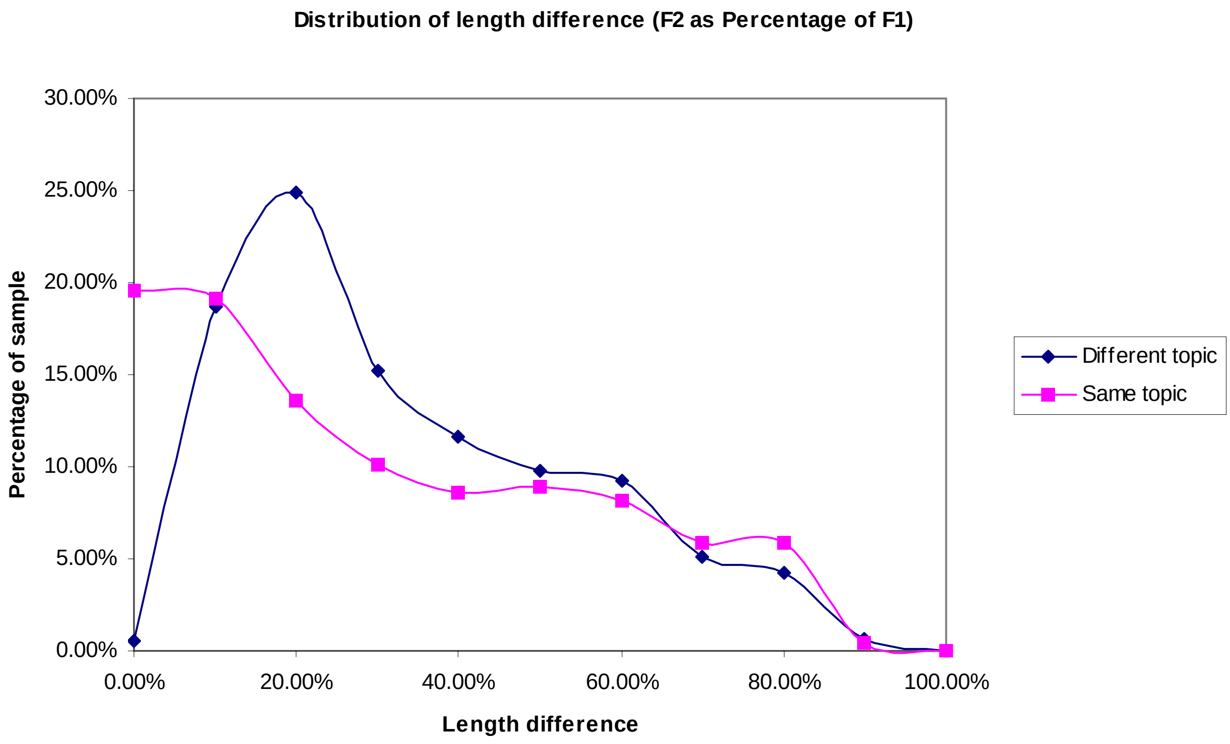


Figure 37: File lengths

Table 26: Distance where File1 >File 2

Source		**Number of occurrences with shortest distance**
Different topic domain		374
Same topic domain		298
	Total	672

χ^2=(1,N=672)=880,p<0.001

The distribution of file lengths after this process has occurred is shown in Figure 37. Thus the majority of both sets are within 40% of one another in length and there is no consistent difference between the two groups.

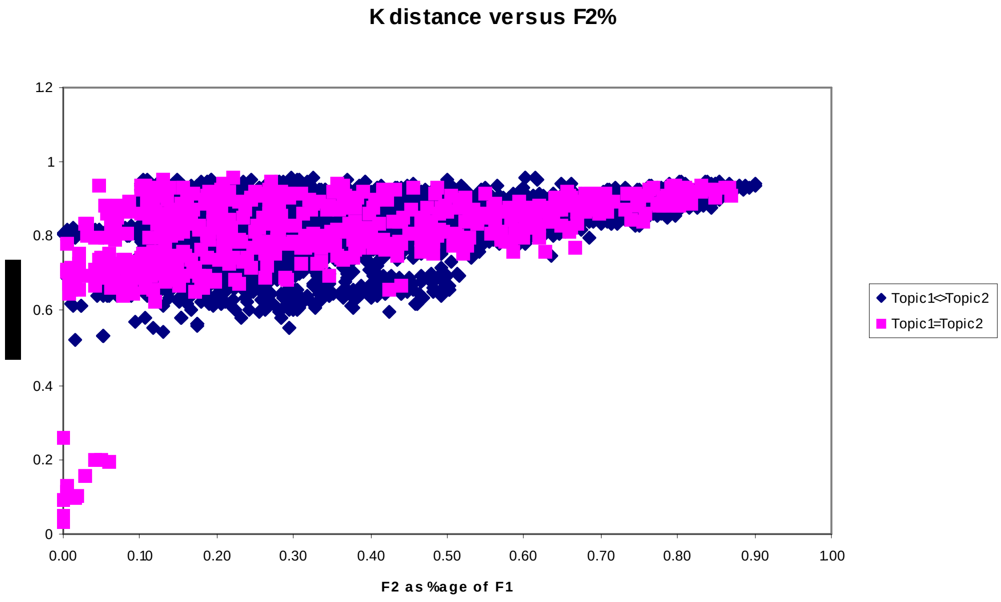


Figure 38: Kol-distance comparison.

This graph (Figure 38) shows the trend for the Kolmogorov distance (Kol-distance) to decrease as the length of file 2 as a percentage of file 1 lengths decreases. This occurs for the situation where the files are heterogeneous or homogenous. However, it is not a very strong association, so that the hypothesis that large differences in file length are required for Kol-distance to be a useful measure, is not completely supported.

4.5 Clustering via Kolmogorov distance

Clustering is a well-known approach to unsupervised learning, and was introduced in section 2.3.1. Tools such as k-means and self-organising maps are described later in this thesis (section 6.4). Essentially, clustering involves the identification of similar objects by means of some sort of distance measure. If the collection of objects being clustered is characterised by a two-dimensional space, for example age and height, then the clustering process is easy to visualise.

The Kolmogorov distance can be used as a distance measure for clustering or classification purposes. The concept of a centroid, as a position in real space, that may not be occupied by one of the clustered objects will not apply to this approach.

One potential approach would be to seed the cluster with a number of known objects. Because the Kolmogorov approach does not imply an absolute location, only relative distances, then the resulting cluster arrangement can be thought of as a set of nodes of a network, rather than the mapping of a solution space. In effect this is a one-dimensional clustering technique. The clustering approach could be performed by a semi-supervised method:

Firstly, a number of initial documents are selected as initial centroids. If these documents represent exemplars for some reason, then as new documents are clustered, the distance from each centroid is calculated and the documents are assigned to whichever they are closest to. For the simplest case this process continues until all documents have been clustered. This approach is computationally very simple, as each new document only needs to be compared to a number of other documents, which are cluster seeds.

In order to use this technique, the cluster seeds need to be identified. For this, an hierarchical clustering approach was used. Firstly, the distance between each document in K-space was calculated and a distance matrix constructed. This is computationally expensive as (n)(n-1)/2 Kol-distance calculations were performed. Then the two closest objects form the first cluster, and the distance matrix recalculated, with the distance to each other object calculated by:

$$D(C,a) = \frac{\sum_{i=0}^{i=n} D(m_i, a)}{n}$$

Equation 9

Where D(C,a) is the Kol-distance from the cluster to another object, and $m_i..m_n$ are the n members of the cluster and D(m,a) is the Kol-distance from each member to the object a.

This process is repeated with the closest objects, whether individuals or clusters being merged into clusters or new clusters depending on their nature:

- If both objects are individuals, then a new cluster is created, and this cluster substituted for the occurrences of each individual in the distance matrix, and the new distances calculated
- If both objects are clusters then a new cluster is created, and all members of the parent clusters substituted in the distance matrix and new distances calculated
- If one object is a cluster, and one an individual, the individual is added to cluster, and the cluster replaces the individual throughout the distance matrix and the new distances are calculated

This process continues with the two new nearest neighbours until the last two clusters remain. A tree is created with the branches representing clusters, with nodes at the point of combination.

The number of clusters used in the final analysis is partly based on preference, for example if it is considered that higher level combinations of clusters have meaning as common ancestors for example. There are three sorts of clusters in a hierarchical system:

a) Clusters made up of original objects
b) Clusters made up of the combination of clusters
c) Cluster made up of both of the above

Clusters of type b, represent ancestral clusters, whereas those of type a) represent primary clusters. See figure 38

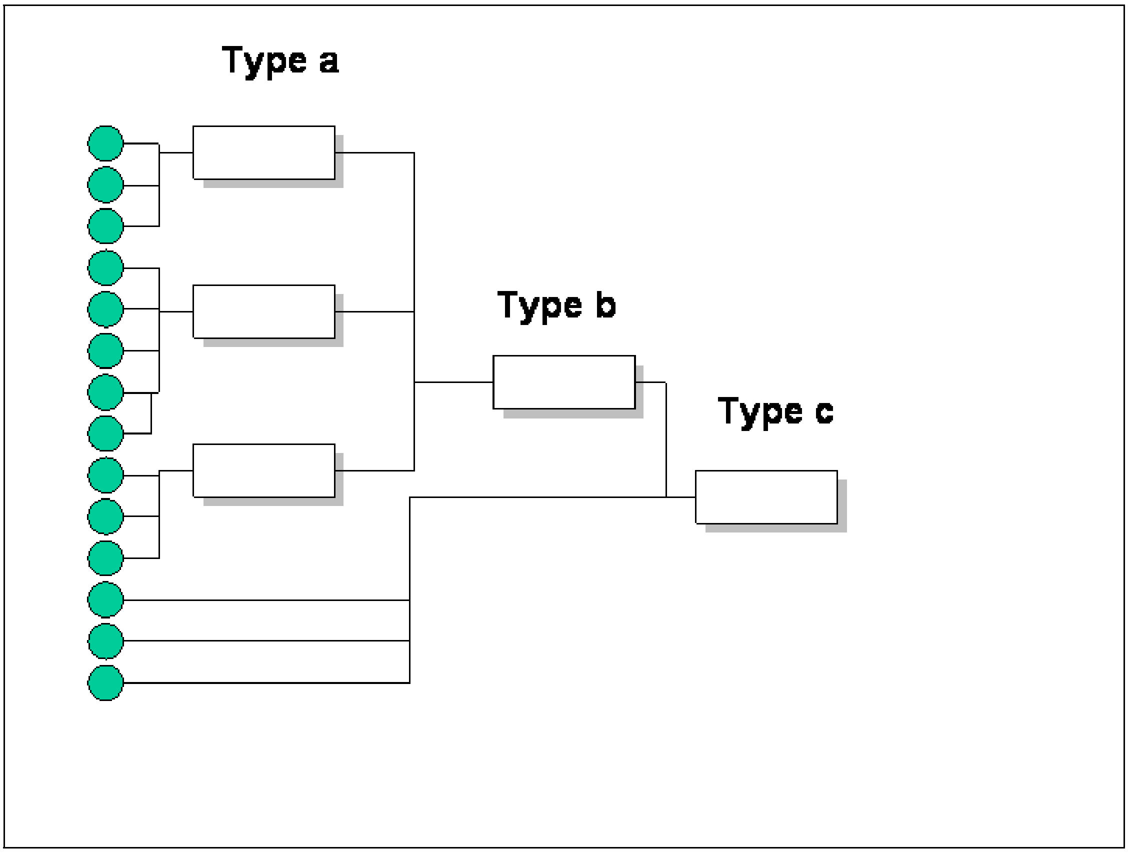


Figure 39: Hierarchical Cluster types

4.5.1 Results of the clustering

The clustering process involved 92 files taken from the BMJ topic collections. These were selected as members of a group of 8 topics. The Kol-distance was calculated between each pair, and a distance matrix created with (92)(91)/2=4186 elements in the matrix. Previous work with 10, 20 and 40 pairs only produced one cluster. The clustering process continued until there was only one cluster remaining. The process was stopped when each Kol-distance pair had one member that was a member of the final cluster.

The coverage of each cluster at each iteration is shown in Figure 40. Most of the clusters show a rise and then fall in membership implying that they are not particularly effective at identifying particular groups.

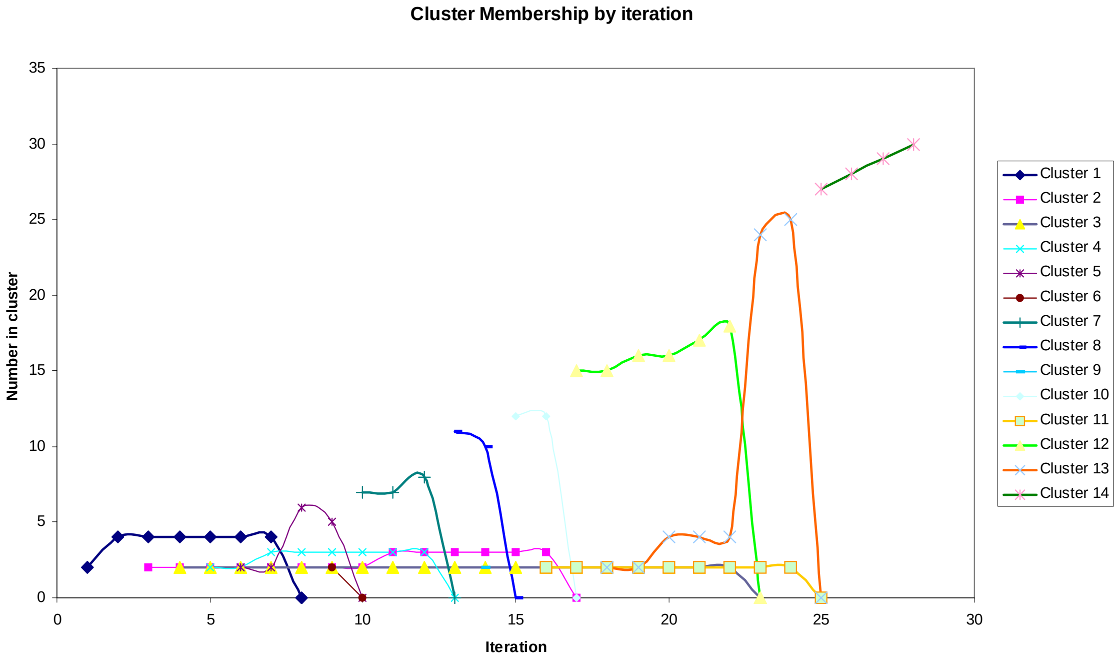


Figure 40: Clusters coverage by iteration and cluster

The degree of overall coverage, by iteration is shown in Figure 41. It can be seen that the rise in coverage becomes linear after around 25 iterations, with one new object being added to a cluster at each iteration.

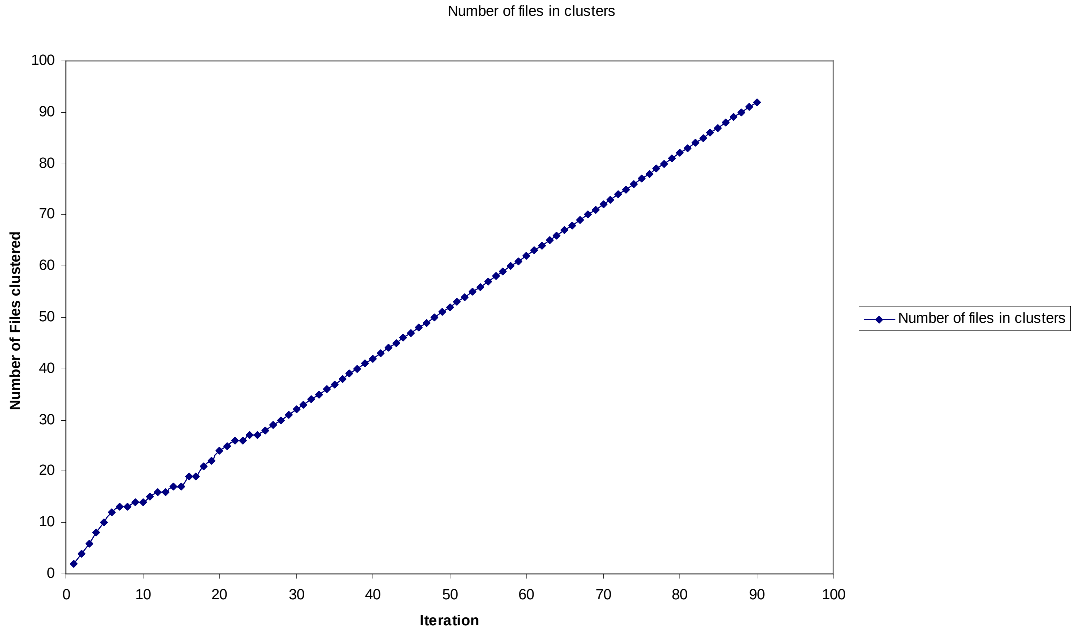


Figure 41: Overall cluster coverage

14 clusters were produced labelled 1-14 and the dendragram produced is shown in Figure 42. The dendragram shows that clustering occurs both by combination of existing clusters and creation of new ones.

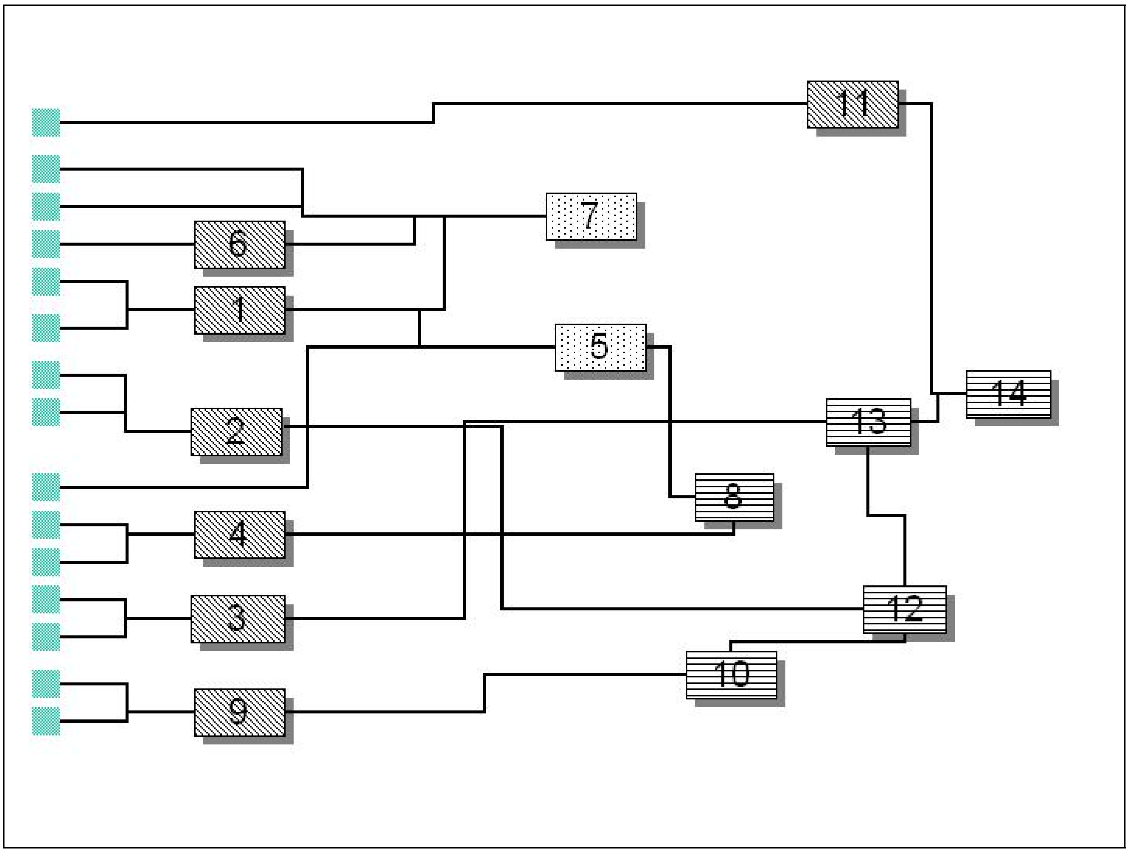


Figure 42: The BMJ Dendragram

4.5.2 Analysis of the clustering centroids

For the purposes of classification, only clusters which include original objects (type a and type c) are included in the analysis of the quality of the clusters. Cluster classification quality was analysed in terms of purity. This is calculated using the method of (Doucet & Ahonen-Myka, 2002). This approach calculated the purity for each cluster by dividing the largest of the topics in each cluster (N_t) by the number of items in that cluster (N_c). The overall purity of the clustering approach is calculated by weighting the Purity of each cluster (P_i) by the size of that cluster (N_c).

$$P_i = \frac{N_t}{N_c}$$

Equation 10

$$P_{Tot} = \sum_i^n \frac{P_i}{N_c}$$

Equation 11

These clusters were then used to cluster the remaining items; this task used the following algorithm:

- Find the “most central” member of the centroid by finding the one with the minimum distance to the others of the centroid identified in section 4.5.1.
- For each unclustered object, calculated the Kol-distance from it to each of these points, representing all a and c type clusters.
- Assign the object to the appropriate cluster.
- Check to see the quality of the classification using purity.

The weighting is important as otherwise the overall purity of a universe with one cluster containing one object would be 100% - the individual purity of that cluster only should be 100% but the overall purity should be less.

Figure 43 shows the result of this analysis. The maximal purity is around 15. This implies that the purity of clusters produced is low. Although this does not mean that the clusters are useless it implies that they are not particularly effective in allocating the pages to particular groupings.

4.5.3 Other approaches to clustering

Although this method is computationally intensive, and somewhat artificial, the alternative of using concatenated files of each member of the cluster is more so. In this approach as members are added to the cluster they would be added to a cluster file, and this would then be the File 1 for the Kol-distance calculation. This approach was not considered as it is hard to see how such an approach would allow the preservation of the identification of the documents – that is to say “end” or “start” sections would appear in the middle of the document.

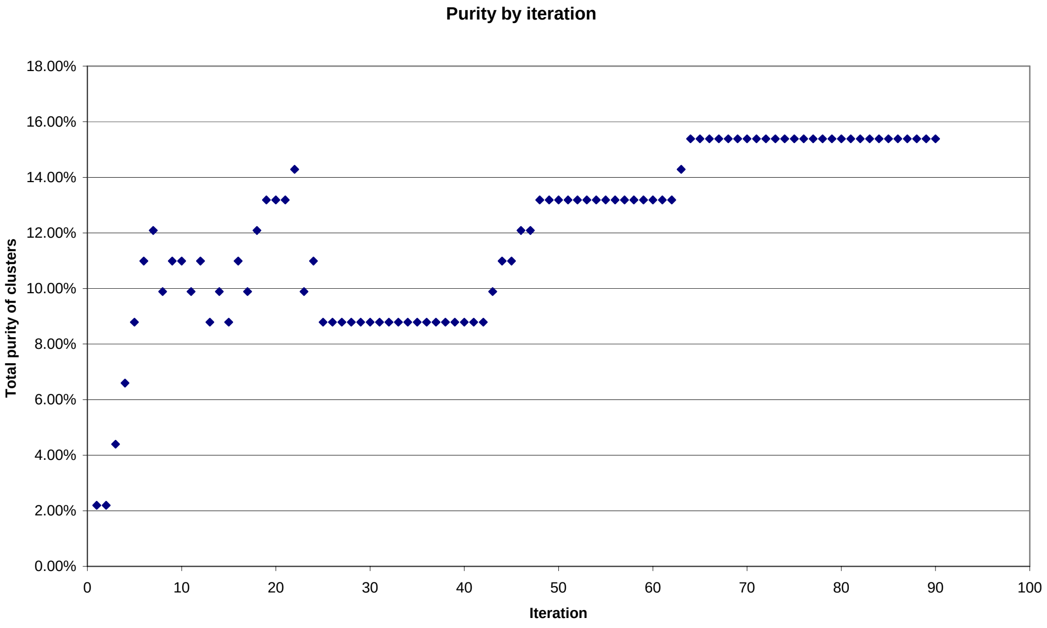


Figure 43: Purity of Kol-distance clustering

With similarity measures and fuzzy ontology theory investigated the next chapters go into details of the implementation of these ideas in software and validation of the approaches.

Chapter Five

This section deals with the components required to construct a system (fSearch) to support improved searching in the medical domain. Section 5.1 describes relevant features of the unified medical language system that have been used in this work. Section 5.2 deals with the actual programming involved, the tools used and some of the issues encountered. Sections 5.3, 5.4 and 5.5 deal with some of the interface and XML issues encountered. Sections 5.6 and 5.7 deal with particular parsing and result presentation issues especially the use of pen-based computer devices.

FSearch comprises three conceptual modules:

1. Knowledge Acquisition.
2. Knowledge Recording.
3. Knowledge interpretation and presentation.

These conceptual modules are implemented by a number of functional modules including:

1. User Interface
2. Internet Interface
3. Database Connectivity
4. Database logical Structure
5. XML document production and interpretation
6. Automated Document analysis

An idealised view of the arrangement of the system processes is shown in Figure 44. Not all the parts of the system have been implemented in the fSearch application, the main parts are the parser, query broker, the fuzzy ontology learning and the ratings recording. Other elements such as similarity measures and the query expansion via fuzzy ontologies were investigated in an off-line or theoretical way.

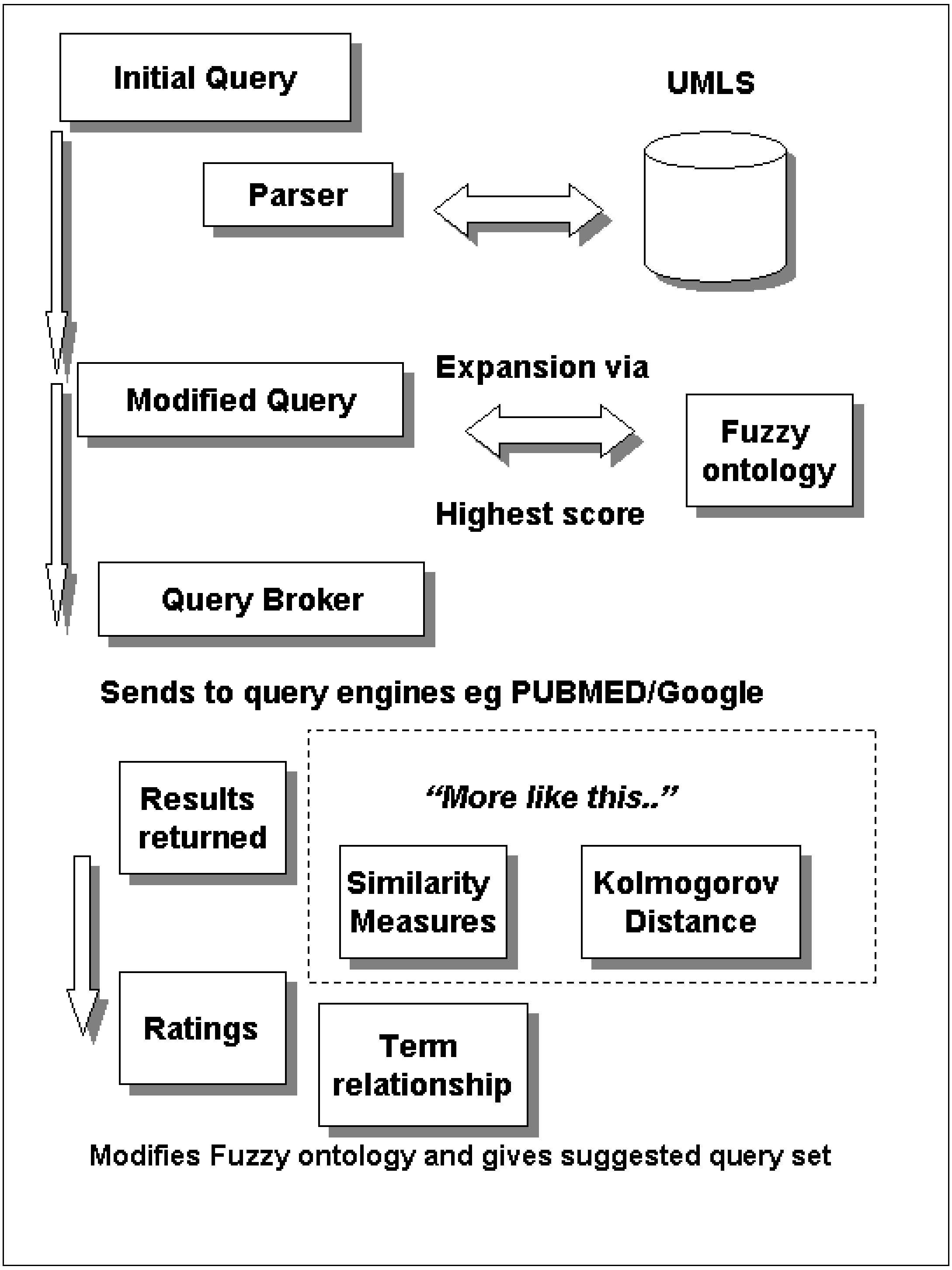


Figure 44: System Processes

5.1 Background

UMLS is an extremely large repository of concepts and related terms for use in biomedical science. Developed by the National Library of Medicine (NLM), it includes not only its own subject hierarchy (MeSH), but also the metathesaurus, which incorporates ontologies from other sources (Nelson, Schopen, J., & N., 2001). These have lead to the identification of currently over 800,000 strings, with about 330,000 unique concepts. For this work I have used the 2000 version that has a smaller number of strings – just below 700,000. The NLM is to be congratulated for not only providing an extremely large lookup table for these concepts (MRCON), but also providing a list of the relations between them (MRREL). Unfortunately because of the mixed parentage of these concepts, the metathesaurus has a number of problems in being used as an ontology.

Firstly there circular relations between terms (Bodenreider, 2001) . Although potential solutions for this problem have been outlined, the multiple origin of the concepts will always lead to a certain heterogeneity.

Secondly there are differences in the types of relation between concepts. The relations in MRREL are quite simple, but they refer to the relationship within the metathesaurus, rather than in the original ontology. Table 27 shows the relations in the metathesaurus system (adapted from (Boxwala, Ogunyemi, & Zeng, 2003)). The representation is expressed in sentences of the form <Concept1> <Relation Code> <Concept 2>. This means that the metathesaurus is useful for comparison between sets of relations, but that translation of relations between systems or ontologies is not exact.

Table 27: Metathesaurus Relations

Code	Type
RB	Broader
RN	Narrower
RO	Not synonymous, narrower or Broader
RL	Similar
PAR	Parent
CHD	Child
SIB	Sibling
AQ	Allowed Qualifier

5.1.1 Database design

UMLS components are available from the NLM on a non-commercial basis after signing of a licence agreement. The UMLS Components are the Metathesaurus, Semantic network, Information sources Map, Specialist Lexicon and Lexical processing tools. The Metathesaurus contains the MeSH hierarchy within it. The UMLS knowledge source allows online use of the components, as well as downloading of particular files or the whole set. One particular complication of the UMLS project is the need to abide by various agreements regarding confidentiality and commercial protection of particular terms or term relations. This especially applies to items in the metathesaurus coming from commercial products such as SNOMED.

The files associated with UMLS can be extremely large. For example the 2003 MRCXT file – which associates concepts and strings with particular placements in hierarchies included in the metathesaurus has over 26 million rows.

Importing such files into database systems can challenging for a number of reasons:

1. Indexing issues can become overwhelming, because some import approaches attempt to index while importing.
2. Automatic import systems often fail, especially if there are any inconsistencies in the file. so "home built" systems must be used.
3. Traditional text editors cannot be used on the source file in order to remove source lines that can confuse import systems, disk file editing systems may be more successful.

In order to overcome these problems a Visual basic.net import utility was written. This reads each line of the import file, checks that it has the correct number of delimiters, and that each calculated field is the correct size and data type and then writes validated data to the database. By use of the try ..finally structure, errors in this process can be ignored if necessary. Thus, failures on particular lines – for example the use of forbidden characters do not cause a failure of the whole import procedure. Where possible indexing, stored procedures and data cubes were utilised to speed up the searching process within for example the MeSH terms. Large numbers of ad-hoc queries were written and temporary tables created in order to effectively clean and convert the data. This all resulted in a very large database – around 1 GB of data in the operational system.

5.1.2 Manual assignment of membership values for the fuzzy ontology

This method uses the membership function shown in Figure 26. 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 5.2). 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 select these terms 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. 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:

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

Equation 12

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.

5.2 Programming overview

Because this represents a research project, the coding and nature of the modules has changed over the course of the project. In addition, there are a number of major programs that do not form part of the final project, but were constructed for particular testing or data formatting roles. This section describes the constraints involved in the choice of programming approach, the software tools used, external programs that are incorporated, common approaches and programming clichés and a general overview of the system. A functional model of fSearch is shown in Figure 45.

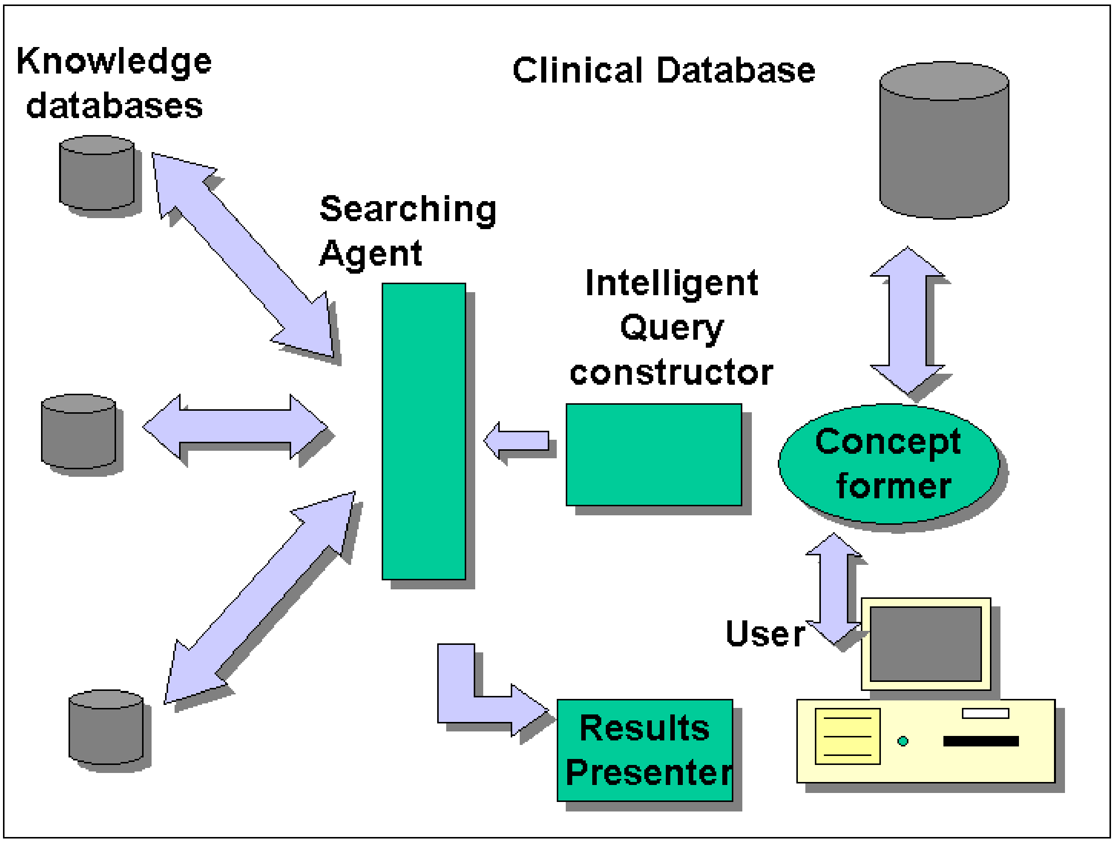


Figure 45: The overall model of fSearch

5.2.1 Constraints and issues

One of the most important aspects of successful system development is the identification and overcoming of constraints. In this case a number were readily apparent:

a) Access to the Internet, especially in institutions is often controlled by strict regulations and rigid firewall policies. In many cases these policies cannot be selectively relaxed, so that particular users or applications cannot claim privileged status.

b) Access to the Internet in many institutions is limited by both speed and cost considerations.

c) Deployment of any application is much easier if royalty-free versions can be compiled.

d) The system is a research project, and usability is critical, so that easy modification is vital.

e) The developer has access to specific tools and experience in their use.

Each of these points is expanded in the following paragraphs.

Firewall policies have a number of effects. In particular they can limit the use of web services such as the “Google API” described below. There are a number of legitimate security fears in regard to web services applications, especially ones that use ports other than port 80 and/or applications other than browsers. The usefulness of web services and their growing acceptance has begun to reduce the constraints associated with them. By modifications to the cache access this constraint has recently been removed in AUT, and web services are now accessible from within the AUT firewall.

Cost of access to the Internet, and limited speed of connection are also common within the target institutions. Within the AUT context, access to the Internet is controlled by use of a proxy, which requires logging in by the user, who is then assigned any costs associated with data transfer. This approach mitigates against use of a server-based page retrieval system, as costs could not be accurately assigned using the current arrangement. In practice, servers at AUT are not able to access Internet pages in order to act as an intermediary. Limited speed access increases the attractiveness of a system that tends to use the institutions' cache, rather than fetch new pages every time. It also suggests that identification of slowly loading pages in advance is useful.

The choice of the major software development tools is largely determined by the final 3 constraints. In this project Visual Basic.net was selected as the main development environment because of the chance to deploy the system without royalty fees, its ease of modification, and the integration with browser and other components and the experience of the developer. This approach did have its difficulties as the Microsoft Visual studio products are tightly integrated with the Windows operating system and require installation of the runtime environment. Other software used for specialist tasks included MATLAB for the network building and SQL server 2000 as the primary database.

5.2.2 Software tools

A number of different software tools were used in the development of the system. As is often the case, different versions of the software and development environment were released and distributed during the lifetime of the project. The final version of the software used is described in each case. Porting software between operating system platforms was generally not excessively difficult, although issues arose from time to time. The priority for the choice of tools was to allow algorithm development and implementation to be as simple as possible. Porting to other languages should be reasonably simple in most cases. In particular putting the user interface on the web was always considered in the process.

5.2.3 Visual Basic.net

Visual Basic.net was released by Microsoft in 2000 and the version used in the project was version 2003. An integrated development environment ("Visual Studio") is provided with the software. The language has been heavily influenced by object-oriented approaches, so that in contrast to previous versions, each object in the application is defined as a class. However, traditional structures such as arrays, and even 'GOTO's are still supported. Form construction is easy, and additional components from other Microsoft or third party products can be incorporated. Deployment is generally easy in the windows environment. VB.net can be compiled in a "debug" or "release" form. The debug form gives the interactive debugging capabilities of an interpreted language, while "release" produces a smaller executable. VB.net also uses code that is similar in structure and form to the ASP.net language, which is a scripting language designed to be used on the WWW. This can simplify any upgrade to a web based system. One of the outstanding benefits of using Microsoft products is the vast amount of support information available both within and outside Microsoft – and especially the large number of newsgroups and bulletin boards available. The relatively low cost of the product, and its wide availability also encourage the formation of a virtual support network. The degree to which support is provided free by enthusiasts and developers seems to confirm the observations of (Kidd, 1994) by presenting the developer community as an important source of both information and esteem for the knowledge workers in that community.

Weaknesses of the language include very limited support for text manipulation and analysis along with a plethora of legacy interfaces to databases etc. Many of the Internet handling components are actually located in the Internet Explorer browser, and hence need it to be installed. Missing or incorrect versions of different components remain an issue with deployment of VB programs because of the number of ways that they can be integrated into the development process. Version control remains an issue for VB, although with a single developer this problem can be minimised. Issues with proprietary software can also cause problems, with deployment into a Microsoft Operating System (OS) environment by far the easiest course. Upgrades and bugs represent a continuing problem for development software and Microsoft products often appear less finished than products from other manufacturers, partly at least because of the intersecting cycles of operating system and Visual Studio releases. As is discussed in (Weiss, 2001), the concept of writing code once and deploying it on multiple devices in the .net environment is very attractive to those who wish to standardise development processes but the tight coupling between the development environment, operating system and other components increases the risks involved.

In order to download material from the Internet, VB.net uses objects supplied with Internet explorer (IE). The WebBrowser object is essentially a new instance of whichever version of the IE browser is installed on the current system. This allows the preferences of the user to be set within IE, and unless modified by the VB.net application continue to be used within the application. In particular cache, proxy and formatting information allows the user to see similar WebPages to those that are viewed directly in the browser, including images and other media and frames etc. Any “popup–killer” software attached to the browser will also continue to function. In this case “Another IE popup killer” (Fast software) was used. The WebBrowser object also exposes the DOM classes, which allow analysis of the displayed page in terms of components such as links etc. This functionality is available in both HTML and XML forms. The Internet connection object represents another means of downloading via the Internet. This object allows setting of particular variables (such as proxy server etc.) directly and does not require a browser to be present. File downloads via this method are particularly suitable for text analysis as only the page requested is fetched, and not the images etc. associated with it. Although it may seem wasteful, because pages are generally cached by the proxy, downloading a document by both methods is not much slower than using one only.

Although VB.net has poor text analysis functions, Microsoft Word exposes its spelling and grammar functions via a Common Object Model (COM) interface. Because of the widespread use of Word, this is not a great limitation on the deployment of the system. In particular the word counting and reading age calculations provided by Word in the check spelling grammar class are efficient and useful. Again, care must be taken to align the preferences stored in the original copy of word, with those needed in the application, although they can be somewhat laboriously set within the VB.net application.

5.2.4 MATLAB

MATLAB is used for much of the learning processing. There are numerous ways in which MATLAB can interact with VB.net, but in this case the interaction is kept as simple as possible by using the database toolbox to load, process and store data in a common database to that used by the main application. The interaction is then limited to the calling of pre-written file functions using the COM from VB.net. Because of licensing issues, no local online processing is done on the client machine.

5.2.5 Database - SQL server 2000

The SQL server database uses ODBC to communicate with MATLAB (using the MATLAB database toolbox) and initially used Active Data Objects (ADO) version 4.6 to communicate with VB. During the course of the project, ADO.net became available and sufficiently stable to use. This increased speed and control of database interaction considerably. ADO.net uses XML messaging and is able to penetrate firewalls more easily than the messaging used in previous versions. The database uses TCP/IP to communicate with the application, which allows a central database to be easily maintained. SQL statements run via ADO are used extensively within the application. The Java programming language is used for accessing the Google API.

Use of a database, raises the issue of whether the database should be stored on a local machine or whether it should be on a remote server. There are advantages and disadvantages to both approaches;

1. Local Servers may allow for quicker access to the data – especially in the increasingly common situation where the local machine has comparable computing power to a server.
2. There are no firewall issues with a local database. Many conventional database connection tools require particular open TCP/IP ports (for example port 1433). ADO.net may be more flexible in this regard than ADO. This is due to the fact that ADO.net is an XML-based system and is less dependent on direct TCP/IP open port connections than ADO.
3. Having all the data at a central server allows immediate sharing and updating of the overall picture. However, at the same time, some of this processing is extremely resource intensive (for example the automatic membership value calculations). This may cause the response to queries necessary for the running of the system to become too slow. Having distributed data requires the use of

some sort of replication mechanism, and admits that instant transmission of information between users is not possible.

4. Licensing issues. This can work in both ways –licensing of server databases may basically be calculated based on the server – for example a “per processor” or a “per server” licence, or the number of connections to the server, for example per connection or per user. There are, however many variations and combinations. The situation is further complicated by the fact that academic institutions may be able to purchase or use software at a much lower cost than commercial ones, however the licence may include restrictions on whether people not employed or taught by the purchasing institution are allowed to use the software. This boils down to a number of potential approaches:
    - Using software with a developers licence – so that there is no restriction on the end users.
    - Using relatively cheap and ubiquitous software – for example access or MYSQL and storing the data locally
    - Using server –based software, but negotiating a licence that allows use by non- staff or students.
    - Using locally- based software and expecting the end-user institution to pay for it.

Some of the same considerations apply to the use of expensive analysis software (such as MATLAB). Although this software does have a “compiler” option that allows redistributable programs to be redistributed, this does not work with the database toolkit (personal communication from Marc Brienne at HRS). After all these considerations, the current system supports two approaches:

1. Centralised Database for machines running on the AUT network
2. Portable Hard drive-based database for machines running outside this network.

In the case that option 2 is used, the static data – for example the material from UMLS etc. is copied onto the portable system, and the dynamic data, such as the user responses etc. are copied back into the main database when that is available. This scheme does not allow for the immediate updating of the fuzzy ontologies, but still allows them to be used.

Any discussion of the costs associated with a particular architecture is open to reinterpretation at a later date as both absolute and relative costs change. It seems likely that broadband Internet costs will continue to decrease, and that processing power and storage capacity on standard personal computers will continue to increase, the portable hard drive system being replaced by flash memory keys, or Internet file transfer for example. However software costs do not follow such a predictable pattern, and the use of open source applications or those with minimal costs per client may well continue to be important.

5.2.6 Other software tools used

As described earlier, the Zip algorithm was used extensively in this work. Initially the BW zip Compress OCX component (Binary Works) was used for this project. In later implementations the PKZIP program used as a shell process, as this was considerably quicker. The WORDNET program, version 2.0 (D.Slomin & Tengi, 2003) was also used in a shell process. Both of these programmes have a command line interface (CLI) mode that allows simple shells to be set up. Limitations on the CLI include the limitation of filenames to “traditional DOS” 8.3 format, with only alphabetical characters acceptable as the first character in output filenames. The Wordnet application was called using a batch file created by the visual basic programme including the various CLI parameters, saved, then called using the shell command. The Wordnet output was redirected to a file on the hard drive and then this file was read by the visual basic application. Although it sounds cumbersome this approach is reliable and efficient, although formatting can be problematic.

In order to analyse PDF documents, a number of approaches were used, but finally the PDF2TXT application was used, again with a shelling system using the CLI. This system allows formatted text documents to be produced from PDF’s, with images ignored.

5.2.7 Web Services

Along with the traditional query string approach to requesting data from WebPages – as used by the ENTREZ utilities – Web Service technology was used for the GOOGLE interface. Web services represent a “family” of protocols and methods including Web services Description Language (WSDL), Simple Object Application Protocol (SOAP), Universal Description, Discovery and Integration (UDDI) and XML. Standards for Web services continue to be developed by the WWW consortium (W3C, 2004b) and this effort brought together some of the work being done on the component protocols.

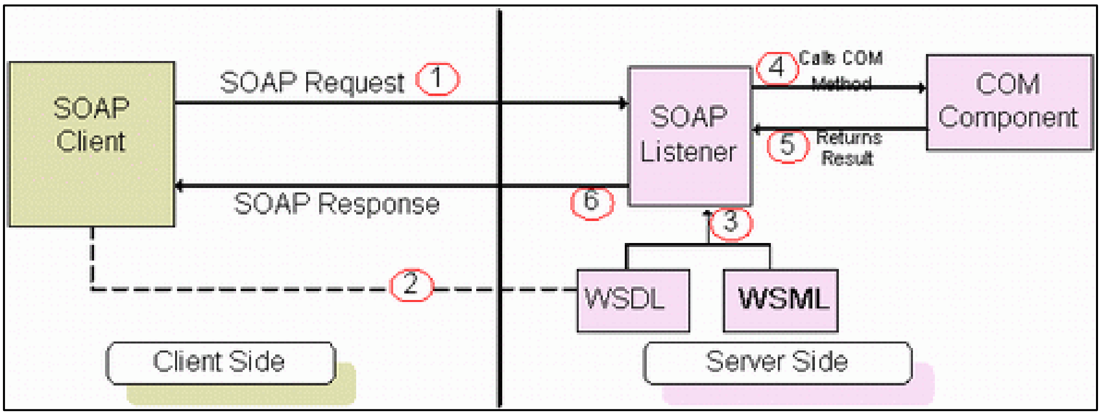


Figure 46: SOAP process from (Ammasai, 2004)

The use of web services is described in (Hogg, Chilcott, Nolan, & Srinivasan, 2004), but the outline is quite simple. Web service providers publish a description of their web service, written in WSDL, which is represented in XML. Publicly available services may be made available via an UDDI site. SOAP is then used running over HTTP to send messages from the client to the web service server as shown in the diagram from (Ammasai, 2004) (Figure 46). Key issues with this technology are:

- All communication occurs using XML documents, running over HTTP. This approach allows transmission through most firewalls.
- The UDDI-based service discovery approach means that services can be easily located and used. The process is particularly simple in the Visual Studio.net environment, where the WSDL specification is downloaded, so that the development environment treats such services as components that are supported by syntax checking and context sensitive help.
- Because of the simplicity of the SOAP approach, interfaces to services can easily remain the same while the internal workings of such services can be updated.

In the case of this work, the PubMed ENTREZ SOAP utilities were evaluated but not used, and the query string approach continued to be used, because there was no discernable benefit at present. However the Google API was integrated into the main application with relative ease using the SOAP/WSDL approach. The other uses of the Google API used the java version of the interface. Google allows the use of the API for non-commercial use, with a limit of 1000 calls per day, and a limit of 10 records returned per search. The API requires the use of a licence key that is freely available from www.Google.com, allocated using an email address as the unique identifier. It would therefore be possible to allocate unique licence keys to individuals for interactive searching, linked via the email address that is also used as an unique identifier in this system. Calls to the API can return alternative spellings of terms as well as the well-known web search.

The web search call will return the estimated number of results, any transformation (i.e. expansion or refinement) of the query string sent, and the top ten results. Each result includes the URL, the title or description of the resource located, the format (html, PDF etc.), whether it is cached and a snippet of the page that is identified. When used as a web service, a Google search object is created which has the attributes described. When used as an offline search for preparation of results, via the Java programme run as a batch application, the results were piped into a text file and this file then imported into SQL server. One point to bear in mind in the use of the offline approach is that a "day" is defined by the Google location on the west coast of the USA, so that bulk queries need to be timed appropriately. "Good behaviour" constraints also commonly apply to other systems – for example no more than one retrieval a second, with a maximal limit on the number of pages or amount of data retrieved, and in the case of ENTREZ a preferred time of access, outside US working hours.

Apart from Google and PubMed, other information sources were used at various times in the project. The BBC search engine (www.bbc.co.uk), Metacrawler (www.metacrawler.com), Excite (www.excite.com) and AltaVista (www.altavista.com) were used at various times. The BBC search engine was particularly good at identifying patient information services and UK-based sources. Metacrawler performs searches on a number of other search engines. Excite is another well-indexed searcher. AltaVista was the original leader in the field but has since fallen back. All of these search engines use a query string approach to sending queries. However, the interfaces for query string searches change, and restrictions are placed on their use. During 2004 for example the BBC, AltaVista and Excite interfaces and search results pages all changed. The greatest difficulty occurs with the returned result set, where formatting changes regularly, and it is optimised for browser presentation, rather than parsing. The maintenance effort associated with supporting these interfaces lead to their eventual abandonment. However, other API's are no doubt going to become available and there may well be web services associated with new entrants into the search engine field, whether free or charged-for.

5.3 Neural network integration

Neural network software can be developed in a standard programming language such as “C” or even VB.net. However tools such as MATLAB considerably simplify the process. MATLAB includes within the neural network toolbox high level commands that allow the programmer to implement a simple neural network in a few lines of code. However, communication between MATLAB and other languages is not trivial.

MATLAB does support a number of interfaces between itself and other languages. A legacy system is the use of Dynamic Data Exchange (DDE), which was commonly used for transmission of data to and from spreadsheets and other office products based on Microsoft technology. More recently, object linking and embedding (OLE) and other tools have become available, which allow for greater communication between running programmes. However, each of these processes suffer from an overhead associated with asynchronous operation – essentially a great deal of effort, and potential failure modes, is required to ensure successful and reliable communication between the different programmes. An additional consideration is the high price of the MATLAB software, which makes it unlikely to be usable on many client machines.

In order to overcome these hurdles, the neural network and fuzzy ontology software can work in two modes, online and offline. When access to the server database and licensed software is available, the online mode is used. When at a remote location, the offline mode is used. This also has the advantage of allowing data to be “cleaned” before use in the calculation process, and reducing response time in-vivo.

Data transfer between the different components of the system is via the Database. MATLAB includes a database toolkit, which allows access to ODBC data sources. This approach also allows a relatively large amount of pre-processed data to be retained on the server that may be useful for re-examining the performance of different learning systems.

For the development of system the “NEUCOM” system was used. This brings together a number of different analysis techniques including SOM and evolving Fuzzy Neural network technology. Much like “WEKA” (Witten & Frank, 1999) such an approach allows the whole cycle of data preparation, visualisation, analysis and result formulation to take place in a common environment with a common data format and minimal programming. By using these tools initial off-line pilot approaches were able to be tested.

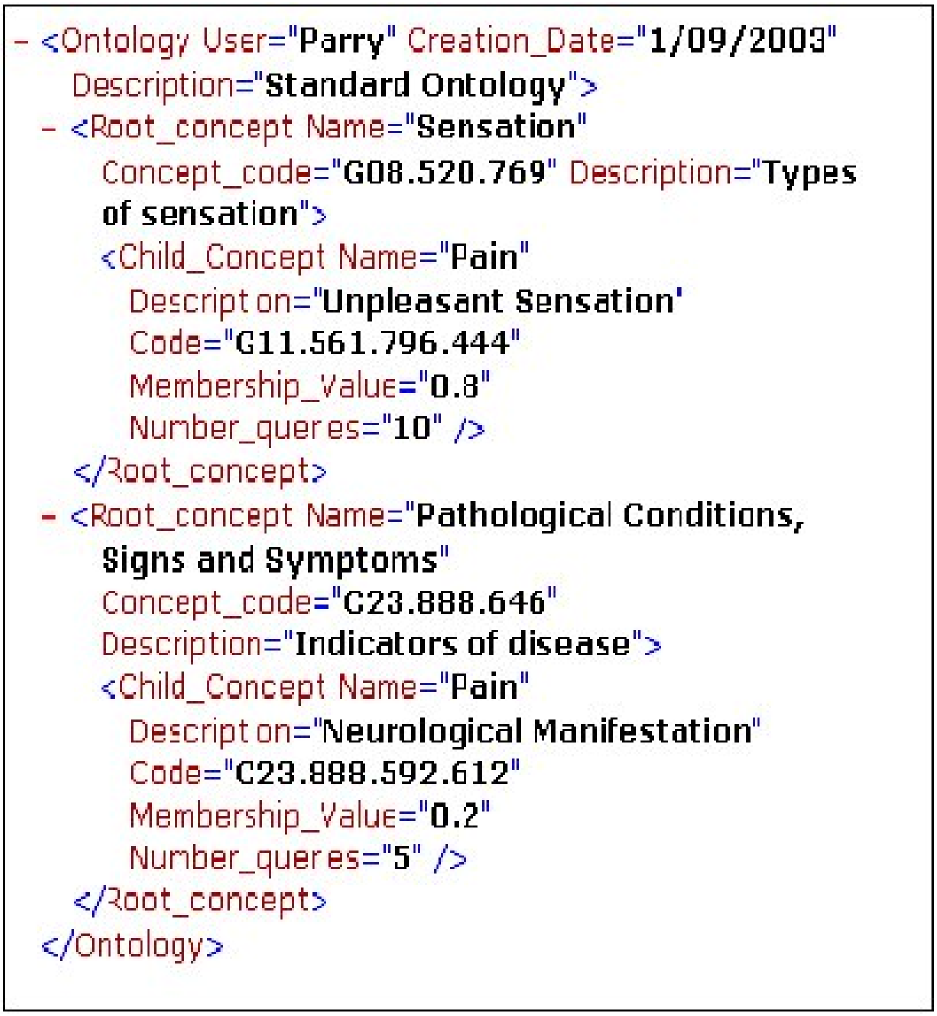
5.4 XML creation

SQL server allows the creation of XML directly from the tables in the database. However VB.net makes the XML handling elements of the .net framework visible directly and in general it is easier to generate documents in this way – especially when more complex processing takes place. Interestingly when ADO.net is used, the communication between application and database is via messages contained in XML format anyway, so there are potentials for simplifying this process.

The process of producing the XML documents follows the following sequence:

- The user details are extracted and turned into the “user element”.
- Terms that are included in the ontology for this user are identified.
- Root terms are identified – those with no “parents”
- Other terms are added to the ontology, with child terms attached to their various parents with appropriate membership functions and relations included.

An example of a fuzzy ontology record is shown in Figure 47. This is an extremely simple representation, but it aims to keep the fuzzy ontology document as simple as possible – in particular avoiding the need for a separate Document Type Definition (DTD). In effect there is an implicit DTD that is understood by the parser. Although very powerful, the use of a DTD can cause problems if the parser does not have access to the file – for example if the DTD is identified by means of a hyperlink and the parser is not connected to the WWW.



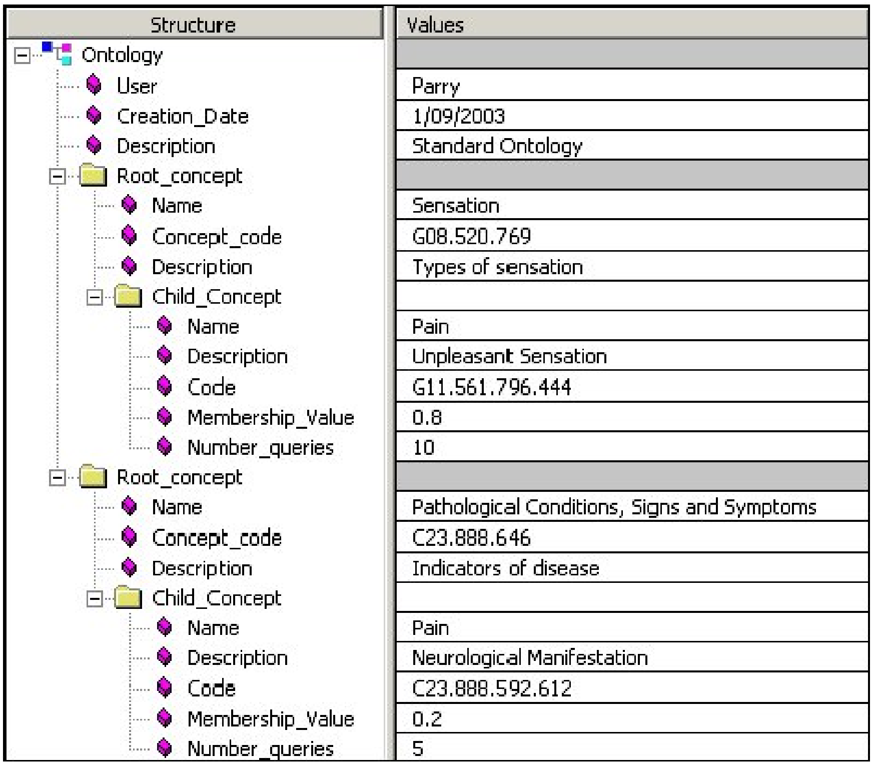
```
- <Ontology User="Parry" Creation_Date="1/09/2003"
    Description="Standard Ontology">
  - <Root_concept Name="Sensation"
      Concept_code="G08.520.769" Description="Types
      of sensation">
      <Child_Concept Name="Pain"
        Descript on="Unpleasant Sensation'
        Code="G11.561.796.444"
        Membership_Value="0.8"
        Number_quer es="10" />
    </Root_concept>
  - <Root_concept Name="Pathological Conditions,
      Signs and Symptoms"
      Concept_code="C23.888.646"
      Description="Indicators of disease">
      <Child_Concept Name="Pain"
        Descript on="Neurological Manifestation"
        Code="C23.888.592.612"
        Membership_Value="0.2"
        Number_quer es="5" />
    </Root_concept>
  </Ontology>
```

Figure 47: XML representation of the fuzzy ontology

This can be more easily understood via an XML editor – in this case XML notepad (Figure 48).

5.5 User interface design

The interface is particularly important for any information retrieval system. ‘Learnability’ is particularly important as it is envisaged that this system will be used by a large number of users who will not be prepared to undergo training. A good introduction to the concept of learnability is found in (Haramundanis, 2001), and the author emphasises the fact that it is a continuum from the unknowable to the trivial. Centralising information retrieval for a particular person in one application with an interface that does not change regularly may make it easier to learn through familiarity, or at least easier to justify the learning time, than an array of different interfaces.



Structure	Values
Ontology	
User	Parry
Creation_Date	1/09/2003
Description	Standard Ontology
Root_concept	
Name	Sensation
Concept_code	G08.520.769
Description	Types of sensation
Child_Concept	
Name	Pain
Description	Unpleasant Sensation
Code	G11.561.796.444
Membership_Value	0.8
Number_queries	10
Root_concept	
Name	Pathological Conditions, Signs and Symptoms
Concept_code	C23.888.646
Description	Indicators of disease
Child_Concept	
Name	Pain
Description	Neurological Manifestation
Code	C23.888.592.612
Membership_Value	0.2
Number_queries	5

Figure 48: Fuzzy ontology as seen in XML notepad

5.5.1 Overall design

The user interface is based around a conventional windows environment. A number of heuristics were used in the construction of the interface, in the spirit of three of the "golden rules" of Shneiderman (Shneiderman, 1998) – Reversal, internal locus of control, and reduction of cognitive load. The concept of cognitive load is associated with the usability domain, where cognitive load is reduced in particular by the use of a graphical interface where the requirement to memorise commands is eliminated, with a limited number of commands available. Reversal implies the ability of the user to reverse their actions, for example use a "back" button, without catastrophic consequences. Internal locus of control refers to the user being in charge of the process. The route through the system was based on a model of the user wishing to answer a clinical question about a particular patient in a session with limited time. For this reason the usability evaluations could occur in less than half an hour each. The overall goal of the user was perceived to be to obtain useful information, in terms of diagnosis, treatment or general information for themselves or others from electronic sources. It is considered acceptable to run multiple searches in the course of each session, with information from the original search and the users own knowledge being utilized. See Figure 49 for an outline.

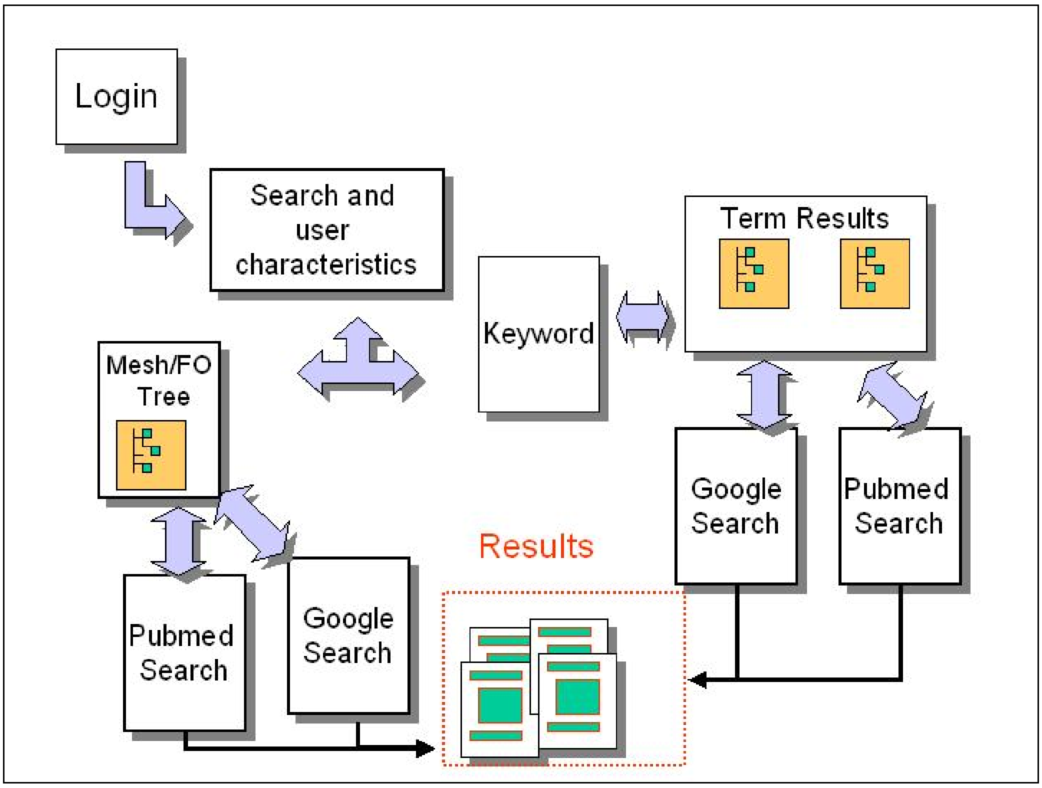


Figure 49: Overall interface design of fSearch

The environment was identified as a clinical area or office type of setting. In this environment it was assumed that there was reasonably rapid access to the Internet, a powerful enough machine to access the various databases, and that there were no particularly restrictive aspects to any firewall or minor-protection software. In this last case, some software designed for filtering objectionable material may use keywords to block material. In the case of obstetrics and gynaecology, it is likely that many of these keywords would occur in clinically interesting and legitimate documents, and these filters may block legitimate health information sites, especially if they are set at a less permissive level (Richardson, Resnick, Hansen, & Rideout, 2002). In addition, the Kaiser Family Foundation, has pointed out that even a "perfect" filter may involve the deliberate blocking of potentially useful sites (Rideout, 2003). The user was seen as anyone who needed clinical information, although the particular area of specialty (see Chapter one) was recorded. The user interface for the system was generally the same for all users; the information presented would be different for different groups. All users were presumed to be interested in the domain of obstetrics and gynaecology. All users were also expected to be able to use the WIMP interface, and read English. Other interfaces, for different languages could be provided quite easily, however there may be limitations on the types of characters that could be displayed, and the selection of MeSH terms is not comprehensive for all languages. A greater objection is the relative paucity of WWW-based information in languages other than English. The following tasks were identified as necessary for the satisfaction of the overall goal:

a. User Registration
b. Logon and authentication
c. User and search characterisation
d. Selection of search approach
e. Term selection
f. Database/search engine selection
g. Document retrieval and analysis

Reversal is particularly well supported by the ability of the user to move back a stage at any point.

5.5.2 Results presentation

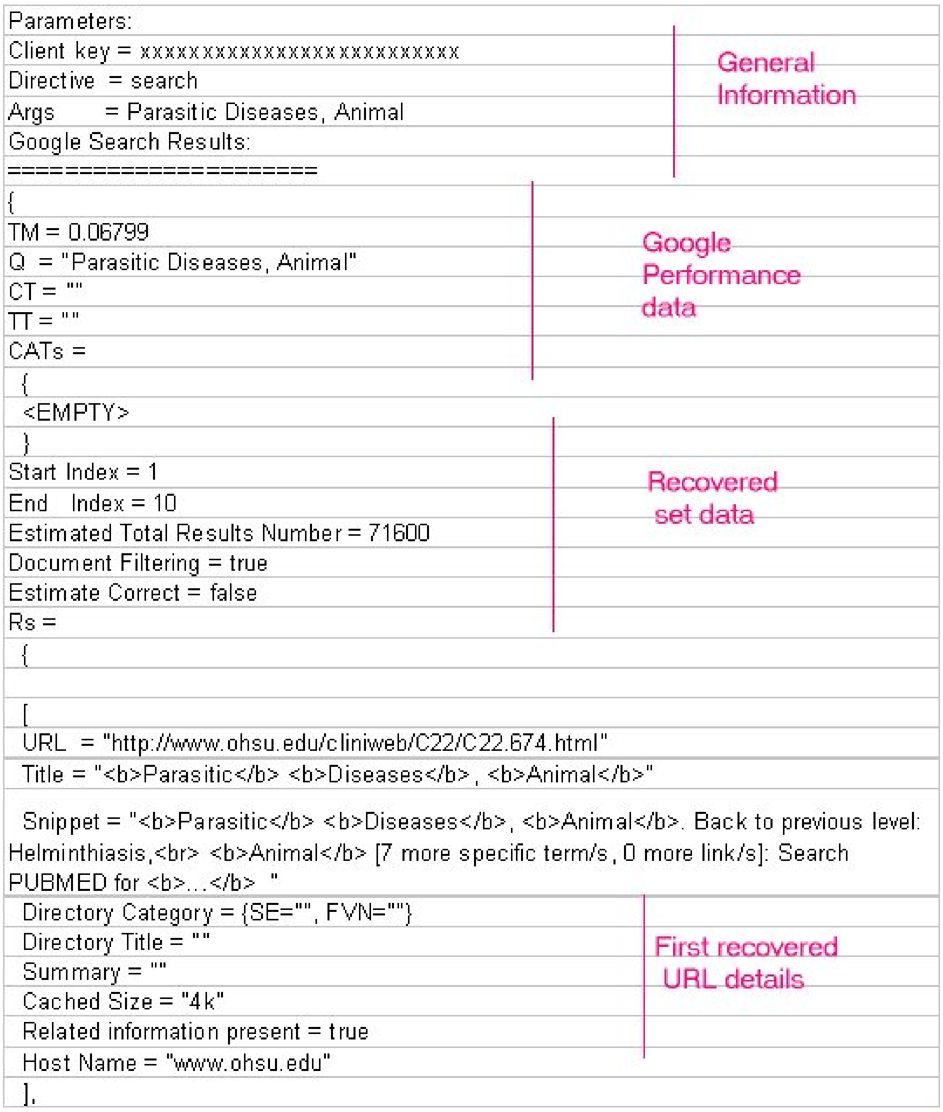
The Google results system used a customised browser to present the website's discovered. This is shown in Figure 78.

The website URL had previously been discovered in most cases using an offline Google API search described in section 5.2.7. The data is recovered in the format shown in Figure 50.

This format was used for the off-line processing. The overall estimated number of records, the URL, the title and the first 200 Characters of the snippet were stored for each of the up to 10 recovered items. When the Google API is used in the online environment a Google API object is created which can be directly read by VB.net.

5.6 PubMed XML parsing

The ENTREZ esearch and efetch utilities were used to recover XML formatted PubMed records. The PubMed XML structure is shown in section 8.8 in the appendix. A large proportion of these documents are essentially metadata, dealing with the status of the document and the date various milestones in the PubMed process occurred. For the purpose of the application, index terms, title, institution, authors and abstract (where available) were required. The parsing took place using the .net XML framework. The data was presented in different ways depending on whether the process is offline or online. The offline process was used to create a MEDLINE corpus, and loaded the data into a text file that was then imported into SQL server. The "good behaviour" constraints outlined in section 5.2.7 were followed. The online process involves users doing a search using the application search interface and then running the *esearch* and *efetch* utilities in real time, and in this case the “good behaviour” criteria are not so important as it is unlikely that the users will be performing hundreds of searches in a single session, and the need to read the retrieved information is taken into account. By default 10 records are retrieved but this can be changed by the user to between 1-10. The retrieved results are ordered by internal PubMed relevancy scores, with overlapping windows showing the most relevant on top. Currently the search uses the modifiers “has review” and “has abstract” and “English Language” in addition to any terms in the search. The results of this process are displayed via the interface shown in Figure 51. Keywords (index terms), title, abstract if available, author names, date of publication and author institutions are displayed.

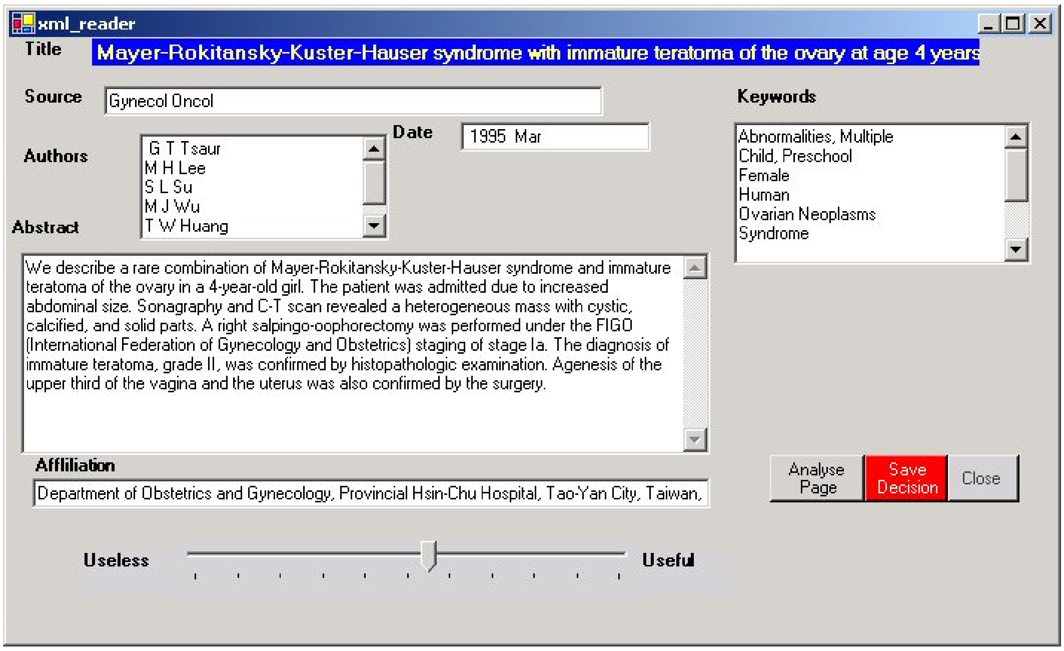


```
Parameters:
Client key = xxxxxxxxxxxxxxxxxxxxxxxxxxx
Directive  = search
Args       = Parasitic Diseases, Animal
Google Search Results:
=====================
{
TM = 0.06799
Q  = "Parasitic Diseases, Animal"
CT = ""
TT = ""
CATs =
  {
  <EMPTY>
  }
Start Index = 1
End   Index = 10
Estimated Total Results Number = 71600
Document Filtering = true
Estimate Correct = false
Rs =
  {

  [
  URL  = "http://www.ohsu.edu/cliniweb/C22/C22.674.html"
  Title = "<b>Parasitic</b> <b>Diseases</b>, <b>Animal</b>"

  Snippet = "<b>Parasitic</b> <b>Diseases</b>, <b>Animal</b>. Back to previous level:
Helminthiasis,<br> <b>Animal</b> [7 more specific term/s, 0 more link/s]: Search
PUBMED for <b>...</b>  "
  Directory Category = {SE="", FVN=""}
  Directory Title = ""
  Summary = ""
  Cached Size = "4k"
  Related information present = true
  Host Name = "www.ohsu.edu"
  ],
```

Figure 50: The recovered data from the Google API search



Title Mayer-Rokitansky-Kuster-Hauser syndrome with immature teratoma of the ovary at age 4 years

Source Gynecol Oncol

Date 1995 Mar

Authors
G T Tsaur
M H Lee
S L Su
M J Wu
T W Huang

Keywords
Abnormalities, Multiple
Child, Preschool
Female
Human
Ovarian Neoplasms
Syndrome

Abstract
We describe a rare combination of Mayer-Rokitansky-Kuster-Hauser syndrome and immature teratoma of the ovary in a 4-year-old girl. The patient was admitted due to increased abdominal size. Sonagraphy and C-T scan revealed a heterogeneous mass with cystic, calcified, and solid parts. A right salpingo-oophorectomy was performed under the FIGO (International Federation of Gynecology and Obstetrics) staging of stage Ia. The diagnosis of immature teratoma, grade II, was confirmed by histopathologic examination. Agenesis of the upper third of the vagina and the uterus was also confirmed by the surgery.

Affiliation
Department of Obstetrics and Gynecology, Provincial Hsin-Chu Hospital, Tao-Yan City, Taiwan,

Figure 51: The PubMed Result display

One of the advantages of using a VB.NET implementation is that the application is easily portable between Microsoft Operating Systems, so one of the aspects of implementation was to consider the effect of using novel devices.

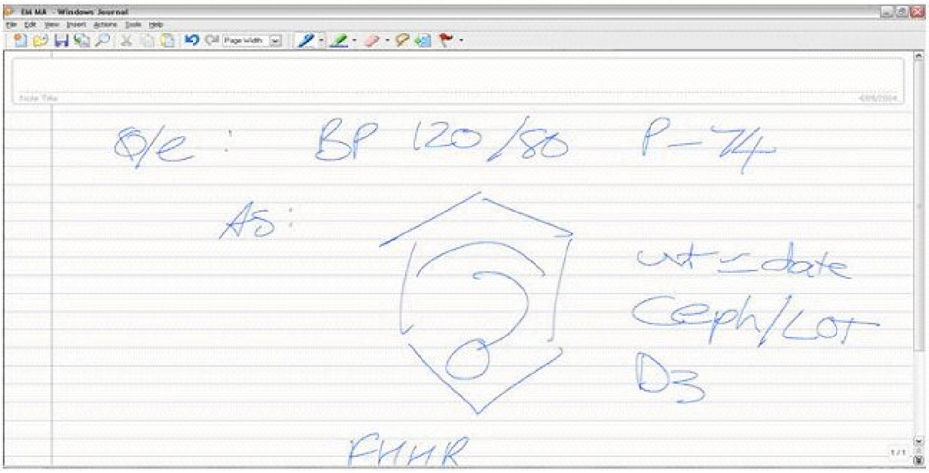
5.7 Tablet PC implementation

Tablet computing devices have become increasingly sophisticated in the last few years, and have become more commonplace with the release of more powerful and widely distributed pen interface software such as palm OS and Microsoft TabletPC (Dray, Siegel, Feldman, & Potenza, 2002). Devices vary, but they all include some facility for word or letter recognition via a pressure or magnetic pen detection sensitive surface. Often the bulk of the face of the device is given over to a display, and there may not be any keyboard, or it may be hidden (see Figure 52). The stylus used for input may have an “erase” function and provision for simulation of common mouse operations. Such devices are generally portable and battery operated, with wireless connectivity to a network. They can therefore be used as collaborative input or output devices, for example in the educational setting (Berque, Johnson, & Jovanovic, 2001), with the advantage that the interface may appear more natural to those more accustomed to writing and reading from paper rather than via the traditional keyboard, screen and mouse used in computers. This may lend itself to use in environments where traditional interface methods are unwieldy or ineffective. These devices are available in ‘ruggedised’ form for use in environments where they may be dropped or made wet etc.

The Windows XP Tablet edition also includes voice input as an alternative route, and supports voice production as well.

There has been some interest in the use of pen-based systems in medical practice (Young, Leung, Ho, & McGhee, 2001). They have a number of advantages including the ability to store drawings along with text, as well as storing text as images, the so-called 'scribble' approach. Along with the ready availability of Wi-Fi networking, such devices may represent a practical alternative for medical records management, especially as some current electronic patient record systems are based around scanned images of patient notes rather than a database driven approach. An example of this is the CRIS system recently implemented in the Auckland District Health Board, where previous notes are scanned and made available via a set of indexed images. An example of what hand-written notes look like is shown in Figure 53. There is considerable interest in their use in clinical areas in Counties Manakau Health Board, although software-licensing costs remain a major issue (Phil Brimacome personal communication).

Figure 52: Tablet PC



O/e : BP 120/80 P-74

AS :

wt = date

Ceph/LOT

D3

FHHR

Figure 53: Simulated Medical notes recorded on a Tablet Device

Having such a technology available already, then adding a information retrieval system to the devices already present on the ward follows the pattern set already, where in the mid 1990’s at Auckland health board, networked PC’s were installed in clinical areas for access to clinical databases and were subsequently used for access to MEDLINE etc. The current fixed PC’s have usability issues, because of their size and vulnerability to damage, they tend to be installed in areas devoted to clerical tasks, rather than direct patient contact. They also tend to be used in a very bursty way, with large numbers of users at key times, when ward rounds are starting or finishing for example, and data being transcribed or memorised for use at the patient bedside. Clinical information systems have had a very mixed success because of various issues (Littlejohns, Wyatt, & Garvican, 2003), but usability remains one of the most important barriers to their adoption.

The tablet approach would allow machines to be allocated to individuals or groups rather than geographically. In order to test the feasibility of this approach for information retrieval tasks the application was installed on a tablet PC. This is a simple task because of the commonalities in the operating system. In order to allow Internet access a commercial “hot spot” Wi-Fi network was used, supplied by Reach software, and running in the AUT campus. Particular issues to consider include:

- Minimisation of network access if possible – achieved by using local storage of the ontology.
- Minimal use of text entry because of the high error rate of pen-based entry especially for technical terms
- Ergonomic considerations including screen size, lighting and pen position.

- Returning to the work described in (Agarwal & Karahanna, 2000), is the system fun?

The software was tested using the same user who performed the initial testing of the desktop PC application. Findings include the low success rate of handwriting recognition for technical terms but relatively high ease of use and portability. These devices are undoubtedly fun, but the large screen size and weight means that they have to be carried in the hand or a briefcase-sized bag rather than in the voluminous pockets available in white coats or theatre greens. A general discussion on the use of these devices concluded that there was potential, but that the relatively high cost, and perceived fragility of the devices would make them a dubious proposition at present. There are also issues with battery life – with a bright screen setting and continued hard-drive use it is rare to have more than about 2 hours between charges.

This situation contrasts with the large-scale use of Personal Digital Assistants (PDA's) in hospitals by clinicians. For example 270 medical applications for Palm OS devices are available via www.palmsource.com. Aside from the standard functionality of appointment and reminder management, contact lists and 'to-do' lists, a number of specific medical applications have been developed, ranging from e-books based around medical texts and local guidelines, to specific calculators for bishop score or gestation or for antibiotic sensitivity (Burdette, Herchline, & Richardson, 2004). PDA's including wireless equipped devices are becoming increasingly popular – even if they can represent an infection control hazard if not cleaned regularly (Hassoun, Vellozzi, & Smith, 2004). "PubMed on tap" has recently been made available for these devices and has been found to be generally usable at the point of care (Alexander, Hauser, Steely, Ford, & Demner-Fushman, 2004). Given the commonality between development environments for Windows CE devices and full-scale devices, it may be possible to perform another port from tablet to PDA. SQL server, for example is available in a Windows CE implementation. Restrictions on the use of PDA's may include limited screen size and lack of memory, but with memory sticks becoming available for $100 for 1 GB and the popularity of the IPOD (Apple) devices, memory may not be such a problem in the near future. It is also possible that 3G mobile telephony devices may become the device of choice in such circumstances, but their user interfaces require careful design (Goldstein, Nyberg, & Anneroth, 2003). Electronic Books (eBooks) have also attracted a lot of attention recently and issues of interface design are crucial in their adoption and use (R. Wilson, 2002).

Chapter Six

This chapter deals with a number of approaches to using available data to confirm the validity of the approach (section 6.1). The simulation (section 6.2) is designed to overcome the issue of large numbers of users being required to make decisions in order to have a preferences database. Some issues of parsing are dealt with in section 0. General document handling information and techniques are described in section 6.4. Use of these approaches in a well-known data set (The Reuters-21578 Corpus) is described in section 6.5. Most of the experimental results in the thesis are contained in this chapter.

6.1 Validation

To demonstrate validity one of the key objectives is to show that the techniques described are functional and can be practically performed. This was done using a number of document corpora. In these corpora the aim was to demonstrate that the different analysis methods were workable, and that they could produce a result. However, these results were not able to be compared to a gold-standard i.e. human classification, so these results represent the output from unsupervised learning in an unbounded space without reference results. This approach is still useful although not definitive.

6.1.1 Analysis of RCOG guidelines

In order to identify a suitable source of Obstetrics-related documents, the website of the Royal College of Obstetricians and Gynaecologists (RCOG) was examined. The RCOG website contains a number of guidelines, patient information sheets and links to other sources of guidelines – such as the National Institutes for Clinical Evidence (NICE) (Sculpher, Drummond, & O'Brien, 2001). Using the spider described above (section 4.3.1), the entire RCOG website (www.rcog.org.uk) apart from the bookshop was searched; the process was stopped when no suitable (i.e. non-bookshop) RCOG domain pages were awaiting analysis. A total 889 files were discovered, of which 576 were from within the RCOG domain. Of these 47 were .PDF files located within ‘resources’ folders, which is the standard means of identifying documents for downloads on the

RCOG site. The final 16 guidelines, selected on the basis that they are related to obstetrics and designed for use by clinicians, were then identified and are listed below:

- Antenatal_corticosteroids_No7.pdf
- Pregnancy_breast_cancer_No12.pdf
- Polys_Ovary_Syndrome_No33.pdf
- Peritoneal_Closure_No15.pdf
- Thromboprophylaxis_no037.pdf
- Tocolytic_Drugs_No1(B).pdf
- GroupB_strep_no36.pdf
- Perineal_Tears_No29.pdf
- Pelvic_Inflamatory_Disease_No32.pdf
- Small_Gest_Age_Fetus_No31.pdf
- Chickenpox_No13.pdf
- Genital_Herpes_No30.pdf
- HRT_Venous_Thromboembolism_no19.pdf
- Recurrent_Miscarriage_No17.pdf
- Ovarian_Cysts_No34.pdf
- Urodynamic_stress_incontinence_No35.pdf

These documents were then examined for occurrences of the MeSH strings. The top 30 terms, along with the number of times they occurred is shown in Table 28. The total number of words in these documents was around 61,500 – calculated by concatenating the text conversions and using the Microsoft word "word count" function.

6.1.2 Corpora generation RCOG and RCS

One of the most surprising early results from the usability study was the limitation in the UMLS language with regards to non-US English. For example the word preterm is not in the standard ontology or even in the cross-referenced ones. The term 'preterm' is not even in a number of popular online dictionaries (for example –the Cambridge advanced learners dictionary at http://dictionary.cambridge.org/ and Stedman's medical dictionary at http://www.stedmans.com/). For this reason, all words were extracted from the RCOG and RANZCOG guidelines described above to form a reference corpora for identification of terms that are likely to be used by non-US English speakers. Fortunately the UMLS also provides in the SPECIALIST lexicon a number of tools for matching alternative spellings and this was used to allow conversion from UK to US spelling as used in MeSH.

Table 28: RCOG Common terms

String	Count
steroid	1520
corticosteroid	1088
Guideline	448
control	336
clinic	320
treatment	288
corticosteroid therapy	272
analysis	272
gestation	160
reduction	160
Health	144
cation	144
glucocorticoid	128
adverse effects	112
report	112
systematic	96
experience	96
development	80
London	80
Pediatrics	80
perinatal	64
mental	64
regimen	64
statistical	64
Research	64
Database	64
Guidelines	64
birthweight	64
author	48

The RCOG and RANZCOG guideline collections were analysed by extracting all words. Guidelines are particularly well suited to this type of analysis because:

- They are often quite long (see tables for word counts), so that there is a fairly wide range of sentences and word constructions
- They cover wide ranges of terms that may be used in the domain, including terms related to diagnosis, symptoms, treatment and outcomes.

- They are written using language commonly used by users of that domain – sometimes the guidelines lead the usage and sometimes they follow.

Each occurrence of a word was recorded, along with its sequential position in the file, and the file it occurred in.

Table 29: The RCS Corpus

Document URL	Word count
publications_scheme.pdf	6043
3rdmolar.pdf	8200
aacguide.pdf	2314
clinicalaudit.pdf	8912
complexityassessment.pdf	1754
dentalerosion.pdf	5917
dianatru.pdf	26864
discolor.pdf	3887
ectopic_canine.pdf	2090
extractp.pdf	1550
hospract.pdf	1017
icppld.pdf	26864
ncg97.pdf	34227
orthodontic.pdf	2213
pd1.pdf	355
pd2.pdf	27506
pd3.pdf	6887
staffgrd.pdf	2252
surg_end_guideline.pdf	5616

The parser was built in Visual Basic and used the ASCII values of the PDF-converted texts to identify punctuation and spaces in the document. Stopwords and words less than 3 characters long were extracted once the wordlist was in the SQL server database. The stopword list from (Retrieval, 2003) was used as this has 429 words compared to the smaller list used previously (STN, 2003), which made the set of remaining words smaller, and processing of the files without the stopwords considerably quicker. Two files were constructed that included all words from each set of sources. As a comparison, the websites of the Royal College of Physicians of London (RCP) and the Royal College of Surgeons of England (RCS) were examined for the presence of guidelines. From each site a number of guidelines were extracted, 19 for the RCS – see Table 29. Note that these guidelines are associated with dentistry.

Table 30: The RCP Corpus

Document URL	Word count
doc_IndepSector.asp	1180
ewtd_caseforten.asp	1250
ewtd_developOOOmt.asp	1834
ewtd_houseoflords.asp	2364
statements_animal_research.htm	247
statements_elderly_care.htm	1454
statements_interm_care.htm	1434
colorectal.pdf	3210
osteosummary.pdf	4725
articles/teachinginoutpatients.asp	710
ClinicalMedicine/index.htm	348
collegelist_home.htm	134
comm_nccg.htm	2673
handbook/gpt/index.htm	186
munk_home.htm	171
pub_census2002.pdf	4594
pub_memorabilia.htm	380
sho_corecurricforms.htm	155
strokeaudit01-02.pdf	9613
wp_actnhsai_summary.htm	1774
wp_allergyunmet.pdf	1920
wp_antiobesitydrugs.htm	568
wp_ch_summary.htm	683
wp_cu_home.htm	3548
wp_ibagmacc_home.htm	1092
wp_ic_home.htm	589
wp_interface_ae.htm	1121
wp_isolatedacute_summary.htm	862
wp_medrehab_summary.htm	1500
wp_np_summary.htm	633
wp_nucmed.pdf	7027
wp_osteo_update.htm	6429
wp_pc_home.htm	2178
wp_pcm_home.htm	2713
wp_pcomp.htm	740
wp_pet.pdf	24390
wp_skillmix_summary.htm	752
wp_thyroidcancer_summary.htm	1061
wp_tiam_home.htm	4386
wp_womeninmed_summary.htm	2067

Forty files were retrieved from the RCP site, with a mixture of topics of interest to physicians; these files are shown in Table 30.

6.2 Simulation

The concept of the fuzzy ontology could use many different learning schemes to assign a membership value. In this experiment the support for each term at each location within each document S(C) was calculated by:

$$S(C) = K \sum_{i=1}^{i=n} \frac{1}{|D|_i} \Big/ L$$

Equation 13

Where K is a scaling constant, n is equal to the number of relatives of the particular location being tested, D is the distance in character space from the start of each relative term and L is the total length in characters of the document. Initially K was assigned to be equal to one, but other approaches such as using the log of the rank for the frequency of the term could be used This approach was taken in order to assign more weight to terms that are close to the target term in the document, and avoid long documents automatically weighting terms more heavily. Support is related to the membership value of a relationship by calculating the total support S_{tot} of a particular term in the entire set along with the support S_{AB} of a particular relationship A with B.

$$\mu_{AB} = \frac{S_{AB}}{S_{tot}(A)}$$

Equation 14

If and only if $S_{tot}(A)= S_{tot}(B)$, then as $S_{AB}= S_{BA}$, then $\mu_{AB}= \mu_{BA.}$

With suitable scaling the fuzzy ontology membership value can then be modified using equation 3 above

$$\mu_{New} = \mu_{Old} * |\mu_i - \mu_{Old}| / Q_{Hist})$$

Equation 3

Where

$$\mu_i \approx S(C)$$

Equation 15

Where the new membership (μ_{New}) is determined by the old membership (μ_{Old}) the membership calculated for this query (μi), and the number of queries that have confirmed the intended meaning of this term (Q_{Hist}). Q is used as a denominator in order to reduced the effect of later learning in order to stabilise the values. For this experiment, only the support values were calculated which are equivalent to (μi), depending on the choice of K. Normalisation or the membership value is necessary to avoid multiply located terms being given undue weight.

Relatives of the location include parents and children of the location in question but not siblings, as these would be likely to refer to distinct entities. The nomenclature is shown diagrammatically in Figure 54.

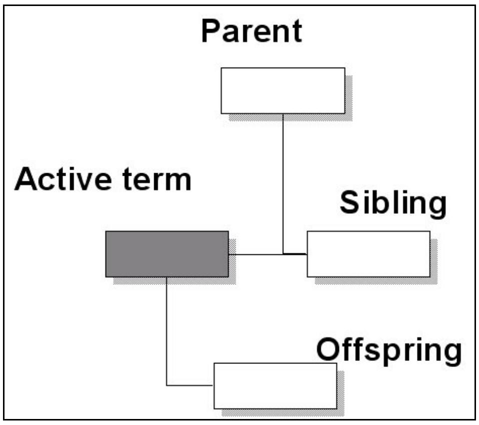


Figure 54: Siblings

Siblings – i.e. members of the ontology at the same level as the active term will relate to different concepts in a well-designed ontology. For example if there was an ontology describing animals, then cats and dogs may be siblings, with "domestic pets" as a common parent, and Siamese and Persian etc. as offspring for the cat and Alsatian and Terrier etc. as offspring for the dog. However the presence of "dog" and "cat" in the same document it is suggested, does not give such support to the idea that the document is about cat as the presence of "domestic animals" and "Persian".

In the MeSH hierarchy relatives are fairly easily identified. The code string is broken up into three character elements of the form *A12.123.123 etc.* A *Parent* consists of a subset of the original code with the final three characters deleted. A *Sibling* has the same leading character sets as the original, but the final three characters are different. A *Child* has the same groups of characters, but with the addition of a new 3-character suffix. As an imaginary example A12.123.123 has the following relations:

Parent –A12.123 (There is only ever one parent)
Siblings –A12.123.456, A12.123.789…etc.
Children - A12.123.123.456, A12.123.123.789…etc.

The foregoing applies to simple “is-a” hierarchies, more complex ontologies cannot be simply categorised in this way. For a discussion of ontology representation please see section 8.4.

6.2.1 Algorithm

At this point the nomenclature needs to be clarified.
The following algorithm was used for calculation:

```
For each ambiguous term  in the collection
 For each location of this term
        For each document in the collection
                Discover relatives
                Calculate distance
     Next document
Calculate average support for this location
Next location
Next term
```

6.2.2 Results of the automatic support calculation

A total of 188 term/code combinations were identified in the corpus where there was more than one supported location for a term. Assuming a value of K =1 for all of the documents, 48 codes had different support for the same term. The magnitude of support varied widely with the minimum being around 1x10-10 and the maximum being 4.6x10-3. Differences between the support for each term were very large, for example the three locations for public health shown in Table 31. This result suggests that large variations in support, and hence membership will occur, which will allow likely preferred locations to be identified.

Table 31: Locations of the term "Public Health" in MeSH

Concept ID	Support	Root term
N01.400.550	4.6x10-3	Population Characteristics
G02.403.776.630.560	4.7x10-9	Health Occupations
G03.850	4.6x10-6	Environment and Public Health

Terms were selected that had multiple occurrences within the general ontology for the analysis. A total of 180 term/code results were identified. The weights were then normalised so that the sum of weights added up to 1. This data is displayed in Table 32. The default ontology was based around the MeSH ontology, and then the updating process is performed based on the corpus being used, or the combination of them.

Term	Code1	Weight 1	Code2	Weight 2	Code3	Weight 3	Code4	Weight 4	Code5	Weight 5
Acupressure	E02.695.522.100	0.50	E02.831.580.499.100	0.50		0.00		0.00		0.00
Arthroscopy	E01.370.388.250.070	0.50	E04.800.250.070	0.50		0.00		0.00		0.00
Biology	G01.273	0.50	H01.158.273	0.50		0.00		0.00		0.00
Blood	A12.207.152	0.50	A15.145	0.50		0.00		0.00		0.00
California	Z01.107.567.875.580.200	0.50	Z01.107.567.875.760.200	0.50		0.00		0.00		0.00
candida	B05.354.381.147	0.50	B05.354.930.176	0.50		0.00		0.00		0.00
Candida albicans	B05.354.381.147.326	0.50	B05.354.930.176.326	0.50		0.00		0.00		0.00
Cardiotocography	E01.370.378.230.150	0.50	E01.370.520.230.150	0.50		0.00		0.00		0.00
Chicago	Z01.107.567.875.350.350.200	0.50	Z01.107.567.875.510.350.200	0.50		0.00		0.00		0.00
Clinical Trials	E05.318.760.535	0.25	E05.337.250	0.25	G03.850.520.450.535	0.25	N05.715.360.775.235	0.25		0.00
Communication	F01.145.209	0.99	L01.143	0.01		0.00		0.00		0.00
Controlled Clinical Trials	E05.318.760.535.365	0.25	E05.337.250.365	0.25	G03.850.520.450.535.365	0.25	N05.715.360.775.235.387	0.25		0.00
Critical Care	E02.760.190	0.50	N02.421.585.190	0.50		0.00		0.00		0.00
Cytogenetics	G01.273.343.180	0.50	H01.158.273.343.180	0.50		0.00		0.00		0.00
Dental Care	E06.170	0.98	N02.421.240.190	0.02		0.00		0.00		0.00
Dental Care for Children	E06.170.152	0.50	N02.421.240.190.215	0.50		0.00		0.00		0.00
Dentistry	E06	0.87	G02.163	0.13		0.00		0.00		0.00
Diet	E05.272	0.00	G06.696.384	1.00		0.00		0.00		0.00
Endoscopy	E01.370.388.250	0.50	E04.800.250	0.50		0.00		0.00		0.00
Environmental Medicine	G02.403.776.630.250	0.09	G03.850.420	0.91		0.00		0.00		0.00
Estrogen Receptor Modulators	D06.347.360	0.50	D27.505.440.450.360	0.50		0.00		0.00		0.00
Fetal Monitoring	E01.370.378.230	0.50	E01.370.520.230	0.50		0.00		0.00		0.00
Fluoridation	E06.761.382	0.49	G03.890.235	0.51		0.00		0.00		0.00
Genetics	G01.273.343	0.50	H01.158.273.343	0.50		0.00		0.00		0.00
Guidelines	N04.761.700.350	0.50	N05.700.350	0.50		0.00		0.00		0.00
HIV	B04.820.650.589.650.350	0.50	B04.909.777.731.589.650.350	0.50		0.00		0.00		0.00
HIV-1	B04.820.650.589.650.350.400	0.50	B04.909.777.731.589.650.350.400	0.50		0.00		0.00		0.00
HIV-2	B04.820.650.589.650.350.410	0.50	B04.909.777.731.589.650.350.410	0.50		0.00		0.00		0.00
Hygiene	E02.547	0.99	G03.850.670	0.01		0.00		0.00		0.00

Hysteroscopy	E01.370.388.250.360	0.50	E04.800.250.360	0.50		0.00		0.00		0.00
Illinois	Z01.107.567.875.350.350	0.50	Z01.107.567.875.510.350	0.50		0.00		0.00		0.00
Incidence	E05.318.308.985.525.375	0.20	G03.850.505.400.975.525.375	0.20	G03.850.520.308.985.525.375	0.20	L01.280.975.525.375	0.20	N01.224.935.597.500	0.20
Intensive Care	E02.760.190.400	0.50	N02.421.585.190.400	0.50		0.00		0.00		0.00
Israel	Z01.252.245.500.375	0.50	Z01.586.500.375	0.50		0.00		0.00		0.00
Jaw	A02.835.232.781.324.502	0.46	A14.521	0.54		0.00		0.00		0.00
Jordan	Z01.252.245.500.400	0.50	Z01.586.500.400	0.50		0.00		0.00		0.00
Language	F01.145.209.399	0.50	L01.143.506	0.50		0.00		0.00		0.00
Laparoscopy	E01.370.388.250.520	0.50	E04.800.250.520	0.50		0.00		0.00		0.00
Lip	A01.456.505.631.515	0.50	A14.549.336	0.50		0.00		0.00		0.00
Massage	E02.695.522	0.50	E02.831.580.499	0.50		0.00		0.00		0.00
Maternal Mortality	E05.318.308.985.550.500	0.20	G03.850.505.400.975.550.500	0.20	G03.850.520.308.985.550.500	0.20	L01.280.975.550.500	0.20	N01.224.935.698.653	0.20
Maxilla	A02.835.232.781.324.502.645	0.50	A14.521.645	0.50		0.00		0.00		0.00
Microbiology	G01.273.540	0.50	H01.158.273.540	0.50		0.00		0.00		0.00
Middle East	Z01.252.245.500	0.50	Z01.586.500	0.50		0.00		0.00		0.00
Morbidity	E05.318.308.985.525	0.20	G03.850.505.400.975.525	0.20	G03.850.520.308.985.525	0.20	L01.280.975.525	0.20	N01.224.935.597	0.20
Mortality	E05.318.308.985.550	0.20	G03.850.505.400.975.550	0.20	G03.850.520.308.985.550	0.20	L01.280.975.550	0.20	N01.224.935.698	0.20
Mothers	F01.829.263.500.320.200	0.33	I01.880.225.500.340.270	0.33	M01.620.630	0.33		0.00		0.00
Mouth	A01.456.505.631	0.33	A14.549	0.67		0.00		0.00		0.00
Odds Ratio	E05.318.740.600.600	0.25	G03.850.520.830.600.600	0.25	H01.548.832.672.471	0.25	N05.715.360.750.625.590	0.25		0.00
Oman	Z01.252.245.500.600	0.50	Z01.586.500.600	0.50		0.00		0.00		0.00
Oral Hygiene	E02.547.600	0.98	E06.761.726	0.02		0.00		0.00		0.00
Oral Medicine	E06.640	0.50	G02.163.670	0.50		0.00		0.00		0.00
Palate	A14.521.658	0.01	A14.549.617	0.99		0.00		0.00		0.00
Parents	F01.829.263.500.320	0.00	I01.880.225.500.340	0.00	M01.620	1.00		0.00		0.00
Plasma	A12.207.152.693	0.50	A15.145.693	0.50		0.00		0.00		0.00
Practice Guidelines	N04.761.700.350.650	0.50	N05.700.350.650	0.50		0.00		0.00		0.00
Prevalence	E05.318.308.985.525.750	0.20	G03.850.505.400.975.525.750	0.20	G03.850.520.308.985.525.750	0.20	L01.280.975.525.750	0.20	N01.224.935.597.750	0.20
Preventive Dentistry	E06.761	0.50	G02.163.721	0.50		0.00		0.00		0.00

Probability	E05.318.740.600	0.25	G03.850.520.830.600	0.25	H01.548.832.672	0.25	N05.715.360.750.625	0.25		0.00
Public Health	G02.403.776.630.560	0.00	G03.850	0.00	N01.400.550	1.00		0.00		0.00
Public Health Dentistry	G02.163.876.770	1.00	G03.890	0.00		0.00		0.00		0.00
Rehabilitation	G02.403.776.620.600	1.00	N02.421.784	0.00		0.00		0.00		0.00
Risk	E05.318.740.600.800	0.25	G03.850.520.830.600.800	0.25	H01.548.832.672.734	0.25	N05.715.360.750.625.700	0.25		0.00
Risk Factors	E05.318.740.600.800.725	0.25	G03.850.520.830.600.800.725	0.25	H01.548.832.672.734.800	0.25	N05.715.360.750.625.700.700	0.25		0.00
San Francisco	Z01.107.567.875.580.200.700	0.50	Z01.107.567.875.760.200.700	0.50		0.00		0.00		0.00
S. Estrogen Receptor Modulators	D06.347.360.827	0.50	D27.505.440.450.360.827	0.50		0.00		0.00		0.00
Stainless Steel	D01.485.910.900	0.33	D01.552.033.847.681	0.33	D25.058.807.681	0.33		0.00		0.00
Statistics	E05.318.740	0.25	G03.850.520.830	0.25	H01.548.832	0.25	N05.715.360.750	0.25		0.00
Steel	D01.485.910	0.33	D01.552.033.847	0.33	D25.058.807	0.33		0.00		0.00
Students	I02.851	0.87	M01.848	0.13		0.00		0.00		0.00
Survivors	M01.643.836	1.00	M01.860	0.00		0.00		0.00		0.00
Thromboembolism	C14.907.355.350.850	0.01	C14.907.355.830.850	0.99		0.00		0.00		0.00
Virology	G01.273.540.859	0.50	H01.158.273.540.859	0.50		0.00		0.00		0.00

Table 32: Membership weights from guideline corpus

6.2.3 Differences between the corpora

After the overall analysis was completed the corpora were compared with each other. Each of the codes for ambiguously located terms was included and its occurrence in each of the corpora recorded where more than one possible code has been identified from the context. The results are shown in Table 33.

Table 33: Ambiguous term sources

Corpora Name	Number of codes that are relevant
RANZCOG	15
RCOG	124
RCS	75
RCP	18

The RCOG is the largest corpus, so each of the other corpora are compared to it. For each of the following tables, term/code combinations are listed where the term or term/code combination does not occur in the RCOG result. The following tables are represent relations that did not occur in the RCOG corpus.

Table 34: RCOG - RANZCOG differences

Term	Code
Estrogen Receptor Modulators	D06.347.360
Estrogen Receptor Modulators	D27.505.440.450.360
Guidelines	N04.761.700.350
Guidelines	N05.700.350
Practice Guidelines	N04.761.700.350.650
Practice Guidelines	N05.700.350.650
Selective Estrogen Receptor Modulators	D06.347.360.827
Selective Estrogen Receptor Modulators	D27.505.440.450.360.827

Table 35: RCOG-RCP differences

Term	Code
Arthroscopy	E01.370.388.250.070
Arthroscopy	E04.800.250.070
Critical Care	E02.760.190
Critical Care	N02.421.585.190
Intensive Care	E02.760.190.400
Intensive Care	N02.421.585.190.400

Table 36: RCOG - RCS differences

Term	Code	Term	Code
Chicago	Z01.107.567.875.350.350.200	Practice Guidelines	N04.761.700.350.650
Chicago	Z01.107.567.875.510.350.200	Practice Guidelines	N05.700.350.650
Communication	L01.143	Preventive Dentistry	E06.761
Dental Care	E06.170	Preventive Dentistry	G02.163.721
Dental Care	N02.421.240.190	Public Health Dentistry	G02.163.876.770
Dental Care for Children	E06.170.152	Public Health Dentistry	G03.890
Dental Care for Children	N02.421.240.190.215	Stainless Steel	D01.485.910.900
Dentistry	E06	Stainless Steel	D01.552.033.847.681
Dentistry	G02.163	Stainless Steel	D25.058.807.681
Fluoridation	E06.761.382	Steel	D01.485.910
Fluoridation	G03.890.235	Steel	D01.552.033.847
Guidelines	N04.761.700.350	Steel	D25.058.807
Guidelines	N05.700.350	Students	I02.851
Hygiene	E02.547	Students	M01.848
Illinois	Z01.107.567.875.350.350	Palate	A14.549.617
Illinois	Z01.107.567.875.510.350	Palate	A14.521.658
Jaw	A02.835.232.781.324.502	Mouth	A01.456.505.631
Jaw	A14.521	Mouth	A14.549
Language	F01.145.209.399	Oral Hygiene	E02.547.600
Language	L01.143.506	Oral Hygiene	E06.761.726
Lip	A01.456.505.631.515	Oral Medicine	E06.640
Lip	A14.549.336	Oral Medicine	G02.163.670
Maxilla	A02.835.232.781.324.502.645		
Maxilla	A14.521.645		

The differences in term/code support in these corpora suggest that different sources of information will produce different fuzzy ontology structures. If all of the corpora had produced identical results then this would not be the case. The hypothesis that different domains will produce different fuzzy ontologies is supported and this justifies further investigation.

6.3 Parsing and document characterisation

Aside from discovering the fuzzy ontology associated with a particular document set, another task is to discover similarities between type of documents that may not be on the same topic but are of the same genre or format. In this section stylistic and syntactic measures are uses, which imply a derivation of the structure and content of the documents. Fortunately the use of HTML in documents on the web leads to a number of structures being available for analysis, but this requires a parsing tool to detect the similarities and differences.

Parsing and analysing documents are important activities in information retrieval. This section examines in detail the approaches used in this thesis work.

The parser in the intelligent search application, fsearch is very limited in scope, and performs the “content-based” value filtering. Essentially it performs 3 main tasks:

1) Extraction of data from structured records such as the PubMed XML format (see appendix).
2) Evaluation of structural information about the document, for example hyperlinks, paragraph structures etc. This material is found by examining the HTML tags.
3) Evaluation of the syntactic structure of the document, for example readability calculated from Flesch Readability index (Talburt, 1985), word count, sentence count etc.

Another approach to differentiation between electronic documents relates to the internal structure of the document. Using the analogy with paper documents, it is easy to tell at a glance the difference between a birthday card from a friend and a tax demand from the Inland Revenue. This difference is obvious from the appearance of the document, and the presence or absence of various elements. Analysis of structural information is attractive because it holds the possibility of understanding the genre of the document as well as it’s content, and thus its suitability for different audiences.

6.3.1 Introduction

Structural information refers to the way the document is put together, in particular the mark up language information. This influences appearance, for example whether there are large numbers of paragraphs or images or links in the document. These design decisions may be made on the basis of satisfying the expectations of the target audience of the document and/or expressing the style of the producer of the document.

Information about the documents contained in the document structure (of HTML pages) include:

- Number of links
- Number of images
- Number of italics
- Number of forms

- Number of bold statements
- Number of paragraph statements
- Number of listed items

These page attributes are derived from the Document object model. In addition the following parameters could be recorded:

  - Depth of URL, in effect the number of "/" characters in the URL used for document recovery
  - Whether the URL uses query strings
  - Document type, for example whether HTML or PDF and whether generated by the server for example .asp
  - Domain Type – .edu, .ac, .com, .org etc. and also national location, such as .nz, .au, .uk

  These aspects are automatically extracted from documents downloaded during the use of the system.

The document set described in the previous section was analysed in this way and a number of clustering approaches used on this set.

6.3.2 Syntactic information

This information is concerned with the way that the text of the document is put together. It may represent a means of identifying genre, or at least intended audience, as it is mostly derived from reading age and other information, linked to the way the text is split by punctuation and the choice of words. Syntactic information included the Flesch readability index along with a set of indices derived using Microsoft word. These are:

- Flesch reading ease score
- Flesch-Kincaid Grade Level score (based on the US grade-school model)
- Words per sentence
- Characters per sentence
- Number of words and sentences

This data was obtained by using some elements from Microsoft Word that are made visible via the Visual Basic for applications (VBA) methods. This approach is described in more detail in section 5.2.3.

The results were analysed using SOM and K-Means as described in sections 6.4.1 and 6.4.2. Paragraph data was calculated, but the conversion process was not adequate for the reliable identification of paragraphs given the wide variety of techniques available for presentation of paragraphs in HTML. These include the HTML paragraph tags (<p>), but also in some cases line breaks (<br>) and even table structures. Paragraphs may also be created by the use of images.

It seems reasonable to suppose that the syntactic structure of documents would be a guide to the genre that they come from. However, the process of text extraction is not perfect, and is likely to overestimate the number of paragraphs, and split up sentences because of the nature of the tagging process in HTML documents. PDF documents retrieved using the PDF conversion process, performed using the software tools described in section 5.2.6, actually produce better-formed documents, which are more amenable to the syntactic analysis. This is partly due to the more sophisticated document format tagging in the PDF approach along with the sophisticated algorithm used by the third-party tools, and partly to the fact that PDF documents are more focussed on being printable, so that the mixture of text and other material is easier to separate.

6.3.3 Document characteristics

As a test of the system, the DOM parser was used to investigate the AUT website. Each page on the website was crawled and the above attributes recovered. This experiment was designed to give a baseline on the document variability and the likelihood of successfully analysing documents by their structure. The variability of the websites in this regard is shown in Table 37. The means are shown merely to indicate that the variation is large between sites. This means that such indicators may be useful for demonstrating differences, but does not prove that such an approach will work.

Table 37: Variation in structural variables

Domain	Link Average	Bold Average	Paragraph Average	Frame Average	List Average	Form Average	Image Average	Italic Average	Length Average	Number of Pages
Apple	46.90	15.40	1.62	0.00	3.15	1.01	48.52	0.09	36392.32	588.00
AUT	51.99	8.28	16.30	0.01	4.76	0.26	42.09	0.92	50886.73	2170.00
Guardian	41.31	12.62	9.07	2.98	0.00	1.56	59.08	0.00	39530.76	234.00
HON	42.79	3.26	11.16	0.16	1.68	0.63	37.63	0.53	21687.42	19.00
Microsoft	37.22	5.00	5.75	0.34	6.33	0.86	32.70	0.30	708961.27	441.00
OBGYN	62.94	12.83	22.45	0.10	7.05	1.56	18.36	1.50	22554.48	167.00
Overall	**49.10**	**9.50**	**12.42**	**0.24**	**4.47**	**0.60**	**41.95**	**0.68**	**126461.45**	**603.17**

This data was also plotted as a percentage of the mean value with the depth of the pages added. This data is shown in Figure 55. This data appears to indicate that there are wide variations in these values between domains.

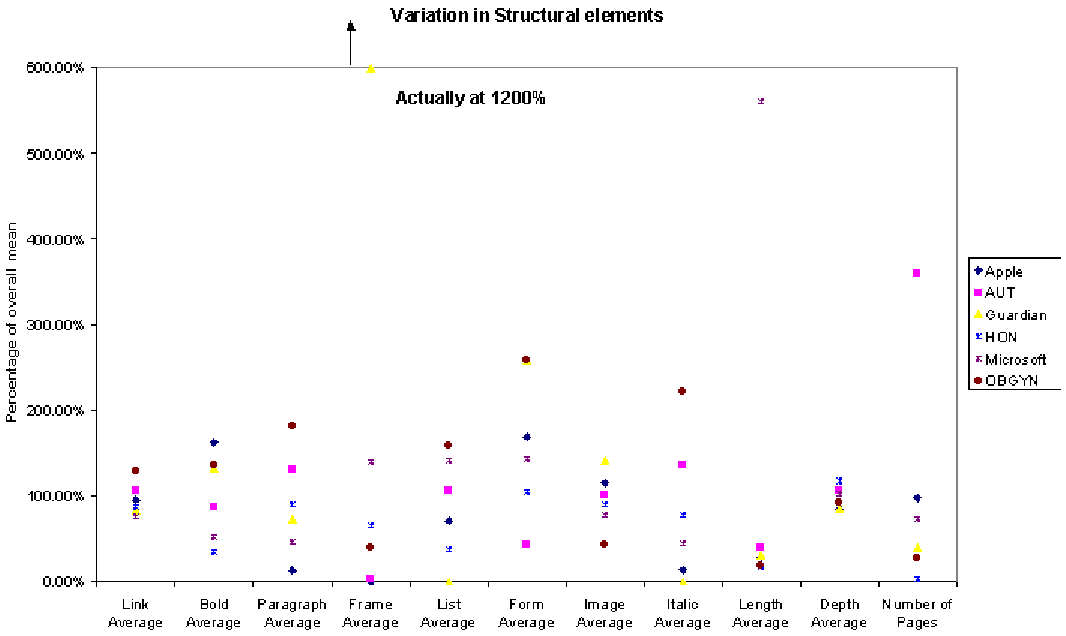


Figure 55: Distribution of structural elements

This part of the work showed that the parser could extract these structural elements from large numbers of HTML pages. The parser was then used on other data sets.

6.3.4 Google Corpus development

In order to get a baseline understanding of the types of documents likely to be retrieved in any search, a corpus of documents was downloaded from the Internet. The corpus was created by selecting 735 mesh terms that were related to the field of Obstetrics and Gynaecology and using them as the search terms for a Google API based search. The documents retrieved by this search were then examined and combined into a number of corpora.

Terms were selected using the MeSH list described in section 0 in the appendix. The Google search was performed using a batch file created to make use of the Google API.

Along with the URL of each document, the title and text fragment associated with the document, along with any date information associated with it were retrieved. In some cases the Google search failed and the process had to be repeated. A total of 735 search strings were used.

A total 17,242 Web Pages were identified. This search was repeated in June and July 2003 and April 2004 in order to gain some understanding of the rate of change of the ranking within Google and the loss and gain of indexed pages. Details of the difference between the highest rank domain each run are shown in Table 38. Domains, rather than actual pages were selected to note the variance because websites are reorganised and also URL's may change for particular Web Pages if they are retrieved from databases via querystrings, and an irrelevant index number changes. This calculation does not identify changes within pages that stay in the same domain.

Table 38: Variability of web domains of top URL

	Number of changes	Total number of URL's	Percentage change
June-July 2003	335	963	34.79%
July 2003- April 2004	578	965	59.90%

Some reasons for these changes include, removal or renaming the pages, blocking of the page from web spiders, changes in the contents of the page, changes in the number of pages linking to or otherwise contributing to the ranking of the page and changes in the algorithm used by Google itself. Google declares that it ranks each page every four weeks. Assuming that around 50% of the pages retrieved for the same search are different in a year, then this implies that searching tools based on queries sent to a search engine cannot make a snapshot copy of the retrieved data set as a substitute for reusing the search engine. Whether the newer results are 'better' than the older set may be open to question. Some of these pages were removed from the analysis because they were pdf or ppt files, or they came from secure sources and could not be downloaded. This left a total of 16,385 pages available for analysis. Of these a total of 7,027 were successfully downloaded. The smaller set was chosen randomly in order to make processing and storage easier.

One measure of the difficulty of identifying useful pages is the number of pages returned by a keyword search. In this experiment an average (arithmetic mean) of 385,047 documents were returned for each call to the search engine. (Figure 56).

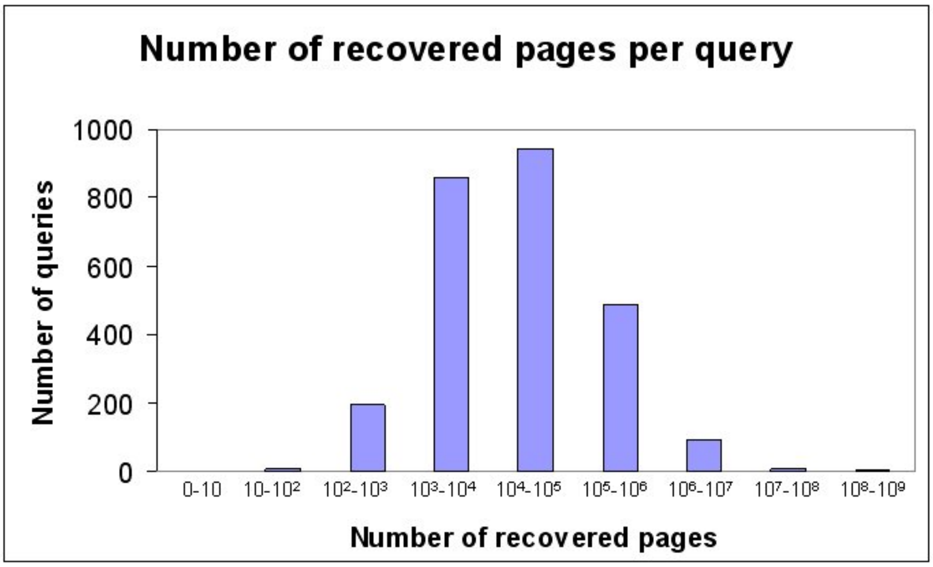


Figure 56: Query Recovery

A calculation of the Zipf law distribution of terms in this corpus is shown in Figure 57. This shows that the corpus has a reasonable compliance with the standard Zipf power-law model, without any great bias to low frequency or high frequency terms. Thus it is not made up purely of obscure technical terms or stopwords.

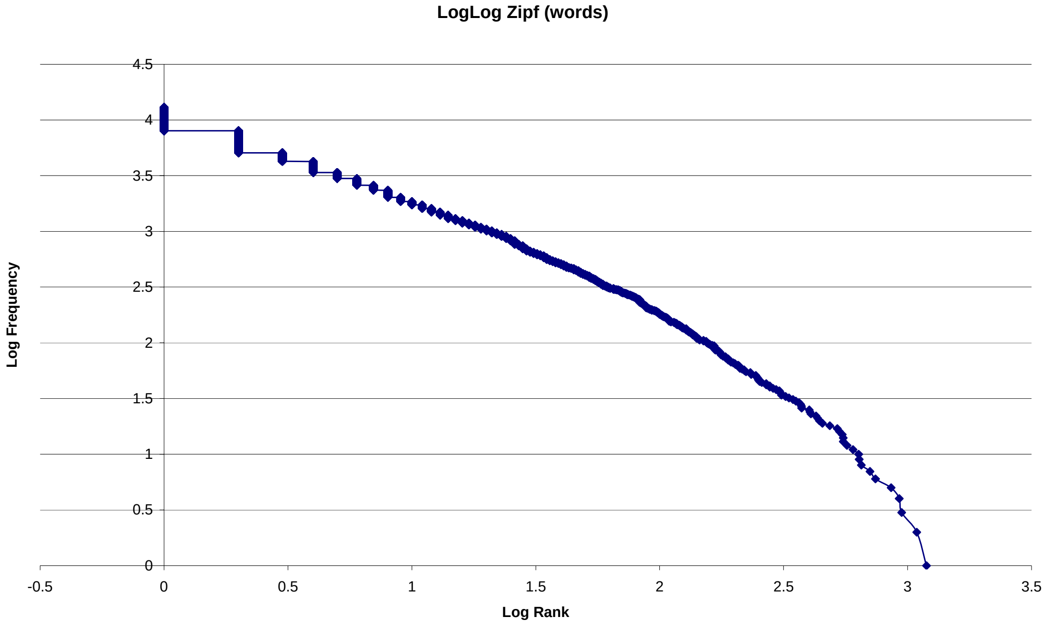


Figure 57: Zipf curve for Google

6.3.5 Document analysis measures

The documents downloaded were subjected to a number of analyses. For all those documents that were downloaded, the first test was to calculate whether the document was a valid document, or whether some sort of "not found" message resulted in attempting to retrieve them. The documents analysed were examined using the document object model for such factors as, number of links, number of images, number of italics, number of forms, number of bold statements, number of paragraph statements and number of list items. ‘Visible Text’ files were created from the HTML using the method described in section 4.4.3. Within the text of the document, measures of number of words, number of paragraphs, numbers of sentences and reading scale were calculated using the Flesch reading ease scale, programmed in Visual Basic, using the algorithm described by (Talburt, 1985) and also the parameters obtained by using the Microsoft word readability measures described in section 5.2.3.

6.4 Document handling

A number of transformations were performed on these documents. Firstly, the documents themselves were zipped and the difference in zip and unzipped file size was measured. The results of this analysis and others were interpreted using a number of different approaches. The clustering techniques used are briefly described below.

6.4.1 K-means Clustering

This was performed on the retrieved documents using the “WEKA” program (Witten & Frank, 1999). WEKA is a set of programs that allow a wide range of data mining algorithms to be run from a single interface. K-means clustering is an extremely simple technique whereby a number (K) of centroids are identified – one for each cluster. The N data points are assigned at random to each of the K clusters (S). Then the centroid is computed for each cluster. This process is repeated until there is no change in the assignment of data point to clusters.

The equation for the centroids calculation is a sum-of-squares calculation given by:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \| x_n - \mu_j \|^2,$$

Equation 16

Where x is a vector being clustered and u is the centroid of the cluster. The approach can be thought of as moving centroids around the data space until the data space is optimally clustered.

6.4.2 Self organising map

The self-organising map software was the one supplied in the MATLAB release 12 Neural Network Toolbox. Self-organising maps allow dimensionality to be reduced in a problem by having competing neurons that distribute themselves according to the input vectors that are presented to them. The neurons learn according to the Kohonen learning rule:

$${}_{i\,i}IW^{1,1}(q) = {}_{i}IW^{1,1}(q-1) + \alpha(p(q) - {}_{i}IW^{1,1}(q-1))$$

Equation 17

Where IW represents the input weight of the i^{th} (winning) neuron and p is the input vector it is responding to. The ♋ represents the bias of the system, which is altered during the learning process to ensure that at the end each of the presented vectors is recognised by only one neuron.

6.4.3 Document analysis

Firstly a “bucket of words” approach was used. Each document had each word extracted from it by a parser written in VB.net. This parser replaced spaces and punctuation marks with tags such as “*” to indicate the start and end of words, and then read the encapsulated words into a SQL Server table. This approach was also used for the analysis module shown in Figure 79. A total of 115,316 words of 3 characters or more were discovered – no check was made of the validity of each word, so for example numbers such as ‘2003’ were counted in the total. Zipf’s law states (Zipf, 1949) that there is a relation between the frequency of a words occurrence, and its ranking in terms of frequency, essentially the frequency is related to the inverse of the rank. The highest frequency word was “links” with 1598 occurrences. The top twenty words and their frequencies are shown in Table 39.

A Zipf’s law analysis was performed of the frequency/rank relation. The $\log_{10}$ frequency versus the $\log_{10}$ rank was plotted and the result shown in Figure 57. This result appears to follow Zipf’s law for the main body of the data where the frequency is proportional to the inverse of the rank.

Table 39: Google frequent words

Count of occurrences	Word
1598	LINKS
1308	DOCUMENT
1080	SEARCH
996	HEALTH
980	DISEASES
860	DISEASE
799	THIS
765	MEDICINE
764	MESH
739	PREGNANCY
738	WRITE
699	HOME
637	MEDICAL
628	PAGE
592	ABOUT
591	PUBMED
583	NEOPLASMS
567	SITE
559	SYNDROME

The domain identification data was also used for an attempted classification into domains by means of the use of structural information. It can be seen that many of these words are not directly related to the domain, but rather are structural. These words (eg links, home) may be useful for genre analysis – i.e. is this a 'home page' or a list of links or a content page.

Data from pages of length greater than 1000 Bytes was used for this analysis. There were 3,944 examples. Each example in the set was used, with the Link, bold, Para, frame, list, form, image, italic, file_len, and depth as attributes with domain code as the classifier. The data was linearly normalised using the WEKA function – that is the range from the highest to the lowest value was calculated, then each value was divided by this number so that the maximal value became 1 and the minimal 0.The data was then randomly split into 66%/33% ratio. Figure 58 and Figure 59 show there was little variation between the two sets of training and testing data. Domain 4 was small but non-zero in both the training and testing set.

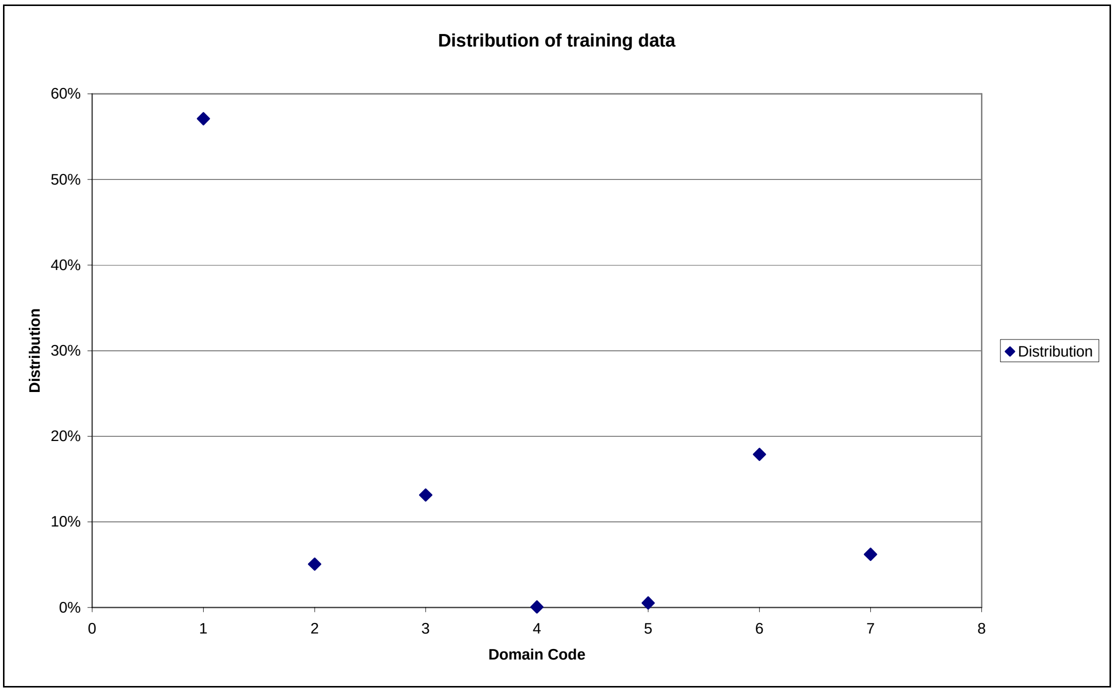


Figure 58 :Distribution between domains in training data n=2981

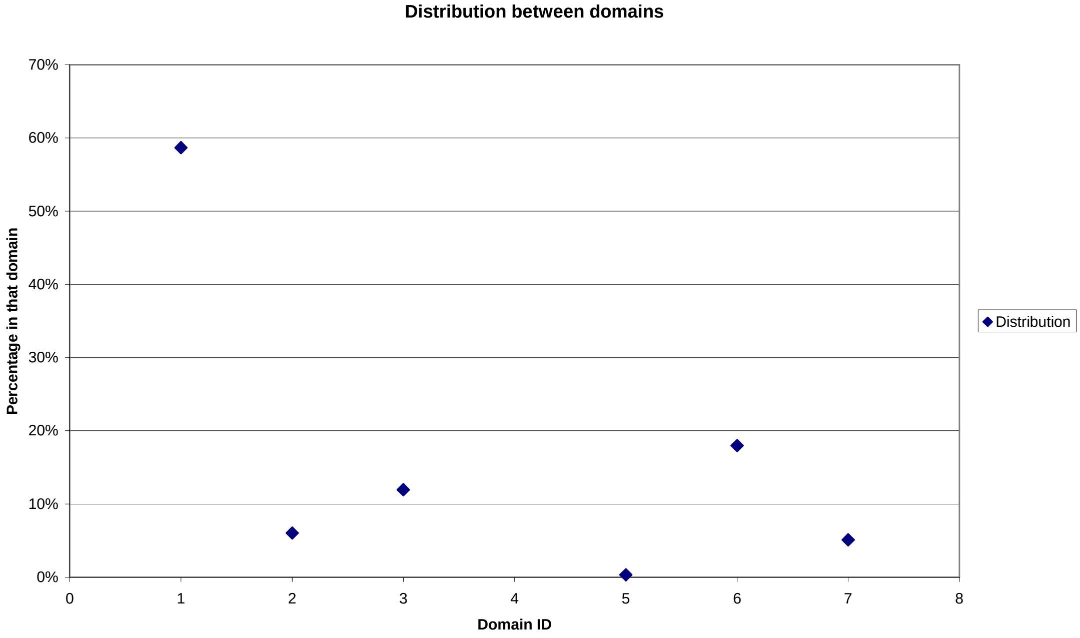


Figure 59: Distribution between domains in testing data n=963

There were a total of 7 domains in the dataset, so each clustering approach used 7 clusters where possible Table 40 shows the distribution of elements in the whole set of the data. The SOM technique was implemented in MATLAB using the Neural Network toolbox along with the database connectivity toolbox to access the data.

Table 40: Structural data characteristics

Attribute	Mean Value	Normalised mean
Link	45.58012	0.098658
Bold	8.727181	0.031393
Paragraph	11.56922	0.012269
Frames	0.287779	0.015146
Lists	4.152637	0.015323
Form	0.573529	0.081933
Image	39.24645	0.101151
Italic	0.622972	0.010741
File Length	116257.9	0.001044
Depth	2.980984	0.372623

The SOM used was 3x3 elements, and the network was trained over 1000 epochs. In an ideal situation the number of neurons would be equal to the number of classes, however there are 7 classes and 9 neurons because of the need for symmetry. The testing set was then fed into the network and the winning neuron for each example recorded. The Purity was calculated for the whole network. The purity is calculated by looking at the sum of the proportion of the largest class in each cluster, multiplied by the relative size of that cluster, so if Pc_i is the proportion of the largest cluster in cluster i, and N_i is the number of examples included in that cluster, and Tot represents the total number of examples then purity is given by:

$$Purity = \sum_i Pc_i \times N_i \Big/ Tot$$

Equation 18

A clustering technique that only puts one class in each cluster will have a purity of 100%. In this case the overall purity of the tested network was 61 %. The purity of each neuron (cluster) is given in Table 44.

Table 41: Purity of SOM neurons for structural measures

Neuron	Largest Class	Purity
9	1	57.14%
8	3	30.12%
7	3	27.78%
6	1	39.62%
5	1	43.51%
4	1	53.24%
3	1	77.17%
2	1	80.00%
1	1	62.24%

These purity results are disappointing for two reasons. First of all the majority of neurons trained to recognise a single class – class 1. Secondly the purity of the overall result was low. There were three other methods of analysis used. For each method a 66/33 training testing split was used (apart from the K-means which used all the data).

Table 42: Analysis of structure to discover domains

Analysis technique	Source	Result	Number of populated classes
Radial Basis Function	WEKA	80% Correct Classification	5
Multilayer Perceptron	WEKA	60% Correct Classification	2
Simple K-means	WEKA	48% Correct Classification	5

The results are shown in Table 42. RBF performs particularly well in this scheme – as a rule of thumb from personal experience any classification below 70% is not particularly useful. Again, these results are not particularly heartening as the number of populated classes is too low.

6.4.4 Document analysis by stylistic measures

For this analysis the number of words, characters and sentences were used along with two “readability measures”. The details of the recovered data are shown in Table 43.

Table 43: Stylistic characteristics

Domain	Number Of Documents	Mean Words	Mean Chars	Mean Sentences	Mean Reading Ease	Mean Flesch Grade
Apple	717	606.42	3260.55	24.50	40.33	9.98
AUT	2202	443.36	2535.28	20.53	33.59	9.72
Disney	3	22.00	124.00	2.00	53.00	8.00
Guardian	246	859.84	5000.33	24.40	31.96	10.59
Microsoft	510	372.37	2212.97	13.64	32.55	9.79
OBGYN	111	336.10	1939.24	10.01	34.21	10.07

The overall purity of the SOM solution was 46.47 %, and the individual SOM purity is shown in Table 44. Table 45 shows the results of the analysis. The Neural network approaches outperform the k-means in terms of accuracy of classification but this is mitigated by the smaller number of classes that are populated. Because there is not an even distribution between the number of examples in each class, a learning scheme that utilises smaller numbers of classes may well outperform one with the number of output classes more comparable to the number of input classes. In the limit, if there is only one output class then the correct classification is around 58% for this dataset. All these results are below 70% and are probably not very useful.

Table 44: Purity of SOM neurons for stylistic measures

Neuron	Purity
8	46.88%
7	45.26%
6	48.57%
5	50.00%
4	45.16%
3	62.50%
2	44.12%
1	46.88%

Table 45: Analysis of style to discover domains

Analysis technique	Source	Result	Number of populated classes
Radial Basis Function	WEKA	61% Correct Classification	2
Multilayer Perceptron	WEKA	61% Correct Classification	3
Simple K-means	WEKA	29 %Correct Classification	5

6.4.5 Derivation of a fuzzy ontology from the PubMed ontology.

Although the aim of the fuzzy ontology is to accurately represent the ontology of particular groups, by understanding their searching behaviour and information needs a fuzzy ontology can also be constructed from already published sources. In particular, a bibliographic database such as MEDLINE allows for the "fuzzification" of an existing structure. Indeed, given the extremely large number of entries in the PubMed database (over 5 million) and the existing ontology (based on the "MesH" structure). PubMed represents both a large corpus of data and a benchmark that can be used for comparison. This work attempted to derive a plausible value of μ for the degree of membership of a particular term, within a particular location of a fuzzy ontology, using a modified version of the method described in (Kruschwitz, 2003), where weights for terms in particular sections, such as abstract or title are higher than for the body of the text.. This method is particularly suitable for documents with a strong, regular structure, in this case PubMed documents. Another method was used – a modification of the subsumption method described in (Sanderson & Croft, 1999), for reordering or realigning the ontology. The subsumption method can be used to create and destroy relations within an ontology, whereas in this case the Kruschwitz techniques were modified to alter the membership values.

A couple of points on terminology are in order at this point. Kruschwitz is careful to label the outcome of his work a hierarchy, rather than an ontology. This reflects his belief that an ontology includes a richer collection of relations than a hierarchy. However, I argue in chapter 2 that the definition of an ontology does not specify a particular set of relations, and indeed that translations between relations are essential for the combination of ontologies.

In practice, subsumption was used when the membership value, as calculated by the fuzzification method fell below 0.1, with at least 2 different position for the term. 0.1 is used as an arbitrary cut off largely to reduce processing. It may be that relations with a membership function less than 0.1 are important sometimes, especially if multiple relations are used,.

6.4.6 Implementation and experimental work on fuzzification

Initially, the UMLS "MeSH tree" table was used as a starting point and branches related to obstetrics and gynaecology (including birth and paediatrics) was used.

The terms selected were the same as those used in the Google search method, as described in section 3.4.1. Each term was followed down the tree – leading to a total number of 472 search terms.

Each of these terms was then searched for using the ENTREZ “esearch” facility. This allows the id numbers of MEDLINE documents to be recovered using a simple URL script (National Library of Medicine, 2003). When using these scripts, the NLM requests that large-scale searches are only made between 10pm and 6 am Eastern Standard Time and that no more than one hit per 3 seconds is made. The first part of this request is easily complied with because of time differences – the queries were run late in the afternoon NZ time. The 3-second rule was accomplished by using the .net frameworks “timer” class; the program loop halting between downloads via the http transfer class, until 3 seconds had elapsed. The lists were recovered as XML documents, using a script written in Visual Basic.net to insert the correct parameters into the string sent to ENTREZ from the data stored in the SQL server 2000 database from the previous part of the process. The URL format to use the ENTREZ tool is given by:

http://eutils.ncbi.nlm.nih.gov/ENTREZ/eutils/esearch.fcgi?db=<database>&term=<term>&retmode=<mode>
where
<database>=PubMed
<term> = term being searched for – in this case one of the descriptions from the MeSH tree
<mode>= Xml in this case.

Each of the returned XML documents includes the number of documents discovered by the query, the query and any expansion terms or alternative terms that the PubMed parser added. The complete XML files returned for each search were stored and then parsed using the XMLdoc class of VB.net the following values were extracted:

- Number of records retrieved by the search
- The id numbers of those records

This data was stored in a database table along with the original URL used and the unique ID number assigned to the original search term.

After this process was completed the ENTREZ “efetch” utility was used to recover a maximum of 20 of the full bibliographic entries for each term, again in XML format from the PubMed database. This was an arbitrary decision – it could be that having a fixed percentage of the potential entry would be better.

These XML documents are of the form shown in section 8.8 in the appendix. A total of 5,437 XML documents were recovered and parsed. The recovered XML documents were parsed to extract the following elements:

- Title
- Abstract
- Country
- Language

Each of these elements occurs once per document and was stored in a table in the database.

The affiliation, journal name and NLM journal code were also retrieved, each had only one value per XML document – PubMed only records one affiliation per document. This data was stored in another table.

The publication type was stored in another table. Each document could have multiple values of this element – see Table 46 for details of this, along with the number of each publication type.

This corresponds to an average of 1.3 article types per document with the maximum being 5 and the minimum 1. This multiple membership of genres can be seen to have obvious similarities with the structural fuzzy ontology described in section 8.2

MeSH heading descriptors were then analysed for each retrieved document. The descriptors – including the major topic status were retained, along with the qualifiers. This data was stored in the MEDLINE keyword table. Note that the qualifiers were not generally used in the analysis after this point.

Table 46: Types of publication in the MEDLINE corpus

Type	Count
Journal Article	5130
Review	502
Review, Tutorial	356
Letter	176
Clinical Trial	149
Comment	147
Randomized Controlled Trial	93
Review, Academic	59
Multicenter Study	59
Editorial	50
Review of Reported Cases	48
Evaluation Studies	41
Newspaper Article	36
News	30
Historical Article	28
Review Literature	25
Biography	17
Guideline	13
Meta-Analysis	11
Controlled Clinical Trial	10
Twin Study	10
Practice Guideline	9
Validation Studies	9
Review, Multicase	8
Congresses	7
Legal Cases	6
Clinical Trial, Phase II	5
Overall	5
Clinical Trial, Phase III	5
Classical Article	4
Clinical Conference	3
Consensus Development Conference	3
Patient Education Handout	2
Clinical Trial, Phase I	2
Lectures	2
Legislation	1
Interview	1
Directory	1

After this, the data from the title, abstract and keyword section was analysed. Firstly a list of all words present in these elements and their frequency was constructed, by parsing each of the relevant fields to retrieve individual words. A stopword list was used to remove stopwords. A further reduction was made using the MRCXT table in order to identify those terms that exist as concepts, so that only those strings were identified. See Figure 60 for details.

Zipf charts were constructed for keywords, abstracts and titles, (Figure 61, Figure 62, Figure 63) and each demonstrates reasonable adherence to Zipfs law. This is particularly interesting in the case of keywords, as it does not show a particularly high (or low)

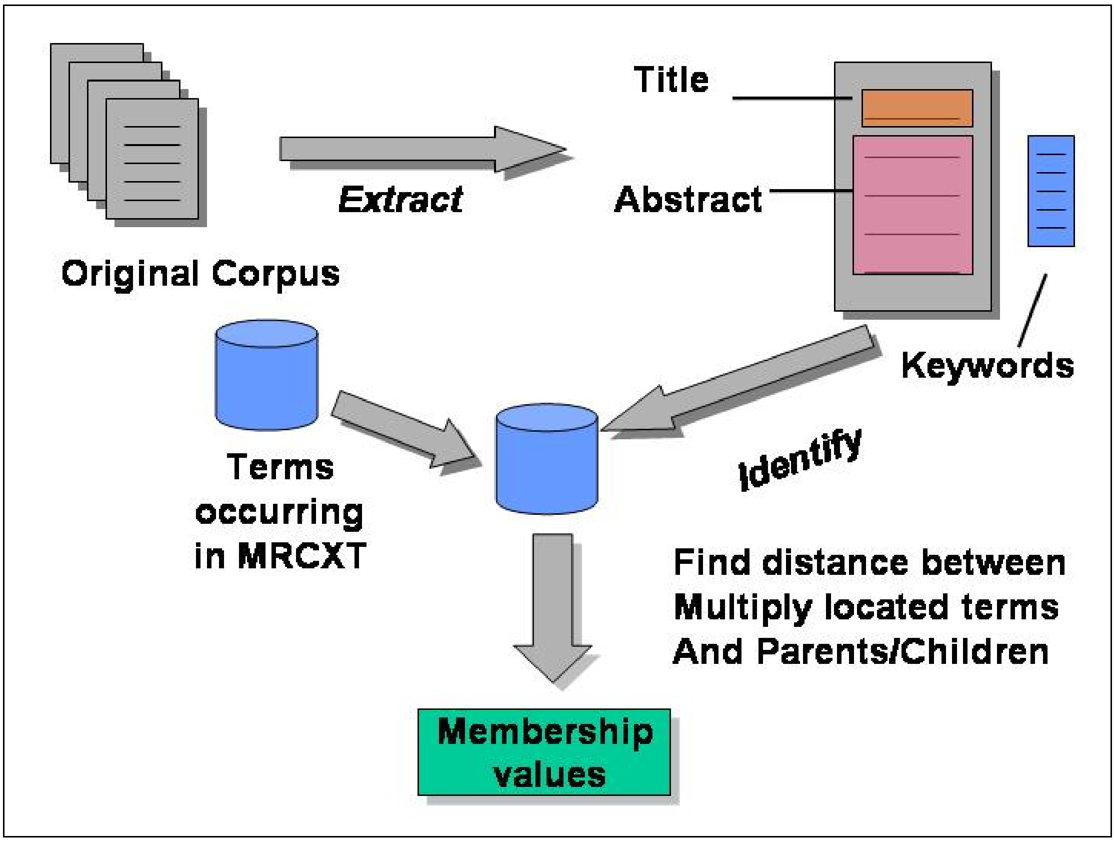


Figure 60: The Data Preparation Procedure

density of technical terms, compared to normal text. It should be noted that there are some compound keywords, but in order to facilitate comparisons, these were regarded for this analysis as separate words.

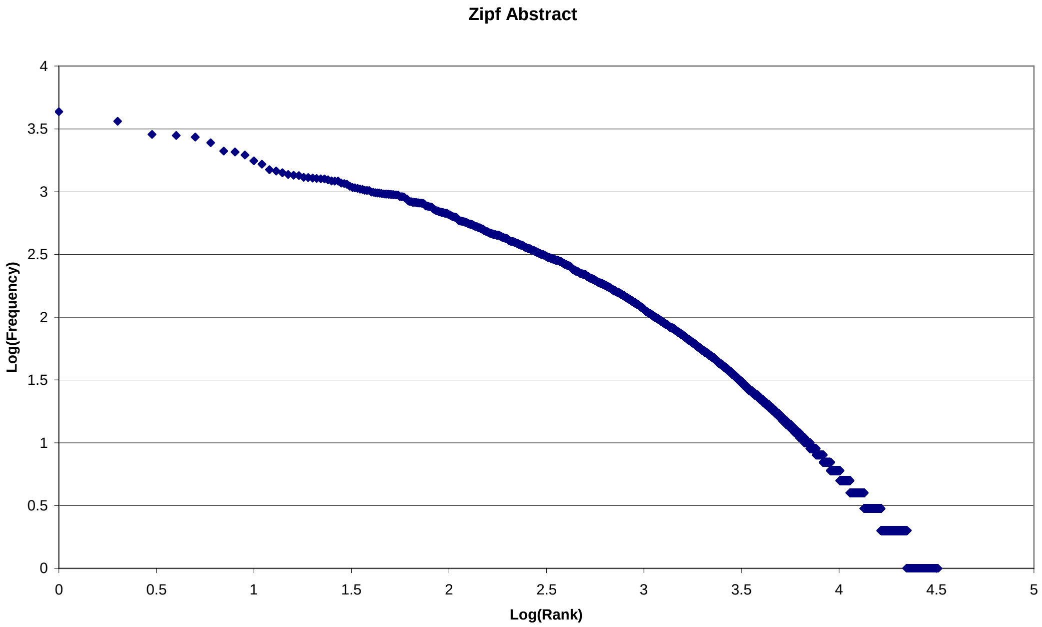


Figure 61: Zipf plot for abstract

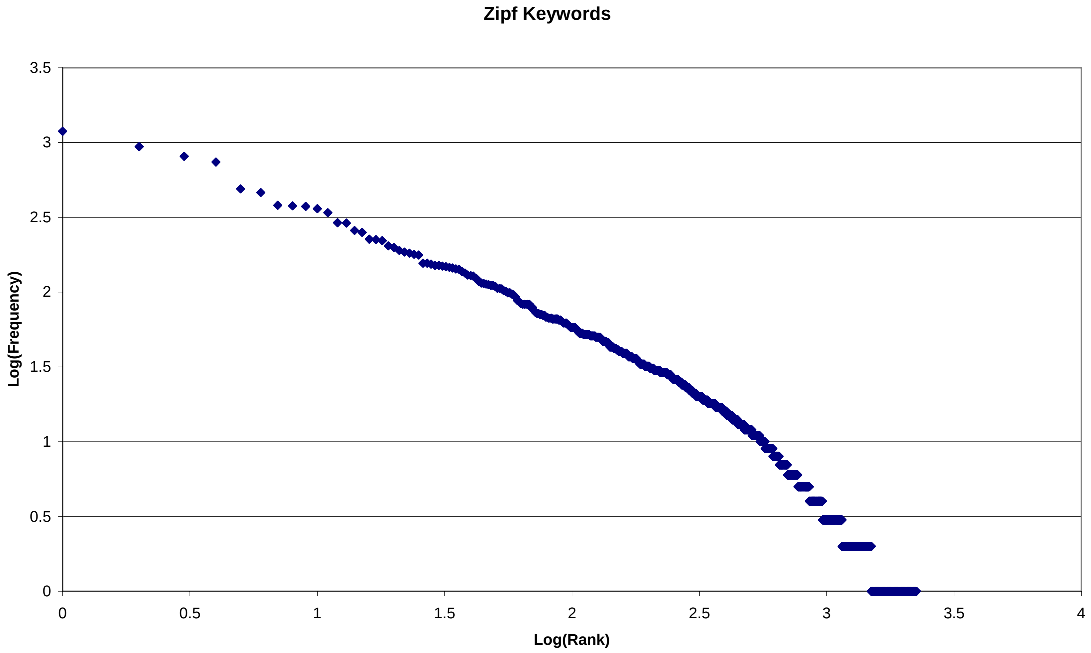


Figure 62: Zipf plot for keyword

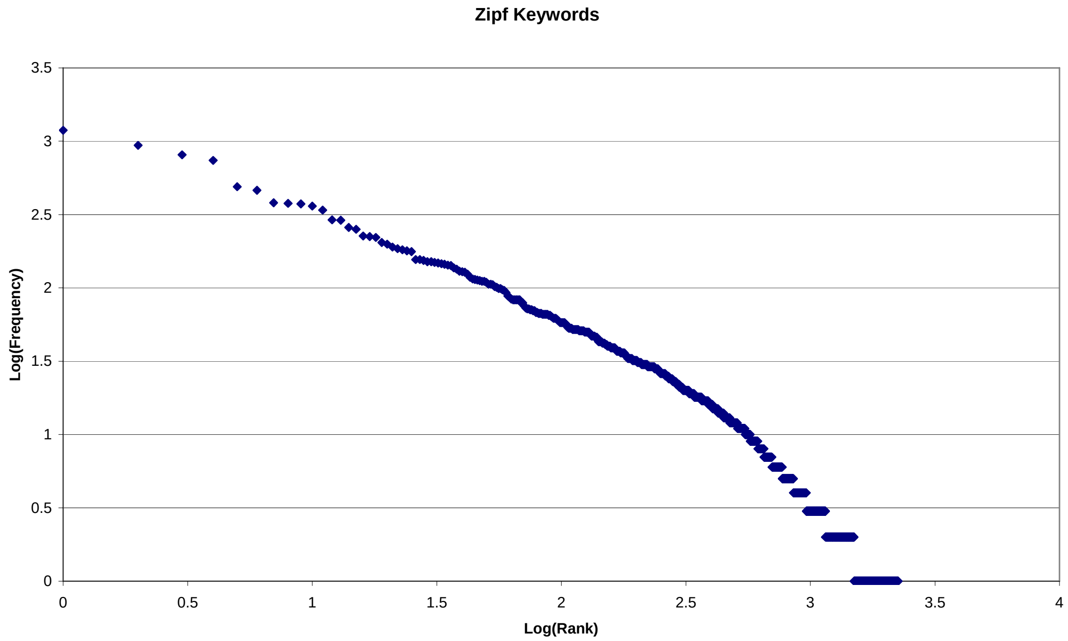


Figure 63: Zipf plot for title

After performing this analysis, an occurrence-scoring table was drawn up with the following characteristics:

If a word appeared in the abstract, the occurrence score was 1, if a word appeared in the title the occurrence score was 2 and if a word appeared in the keyword list the occurrence score was 4. These scores were cumulative, but for each only the first occurrence in each section was included. These scores are arbitrary, but the choice of integer was made in order to allow the AND operator to separate out the values at a later date. Therefore the maximum score for each word/document combination was 7 (i.e. 4+2+1) and the minimum score was 0. A table in the database stored this data in the form id number (of the document), the word in question and the "occurrence score" as calculated above. The initial calculation compared the terms discovered with the search terms – as these search terms are members of the ontology.

This formulation is very close to Bayes formula in that the calculation is effectively the probability of each word occurring with that score independently divided by the probability of the words occurring together. It is then multiplied by the occurrence score for each match, and then summed over each match combination, i.e. each value of the occurrence score.

Normalization of this value is not trivial. It should be remembered that the documents studied represent a small subset of the total number of documents in the universe of documents. In order for these values to be effectively calculated each word is normalized independently, that is over every document it occurs in, rather than the whole set, so that the total score for each word - with all other words in combination is calculated. This number is normalized as 1 and then all other values are calculated appropriately. Note that this means that the membership value is not commutative – a rarely occurring term may have a particularly high membership in relation to a common one, but the reverse may not be true.

Concepts were then identified using the UMLS MRCON data; this list includes various strings that have been identified as belonging to particular concepts. By using this approach synonyms and words with differing endings were identified, as well as homonyms. In addition the UMLS MRCXT table was used in order to link strings associated with concepts with particular items in the MeSH hierarchy. MRCXT is a particularly large table with over 20 million records, because it links concept terms with all of the ontologies present in the thesaurus. In fact the native format file is so large that it causes failure of most import schemes. A small programme was written to parse the text version of this table, the latest version of which was the 2004 update downloaded in December 2003.   Only those records in MRCXT that corresponded to MeSH locations were used in this analysis, and the parser selected for them and loaded them into an SQL server table. At this point the fuzzy ontology begins to be constructed. This ontology is based around the MeSH hierarchy.

6.4.7 Results of this analysis

The number of MEDLINE documents received was reduced from that recovered in the structural analysis to 5,437, because of the removal of documents with no ontology terms. Only terms with a total weight of >5 in order to reduce the number of terms displayed, were included in the analysis that produced the term analysis shown in Table 47.

This meant that only 395 terms were used as original search terms, because others did not produce sufficient weighted values. The distribution of search terms/ weighted value is shown in Figure 64. This shows as expected a strong negative correlation between $\log_{10}$(count) and the rank of the weights – that is to say that strongly weighted terms are much rarer than low-weighted terms.

Table 47: The search process

Parameter	Value
Number of Searches	395
Number of terms discovered	20154
Mean sum of weights of discovered terms	2 (range 1-28)
Number of Documents involved	5437

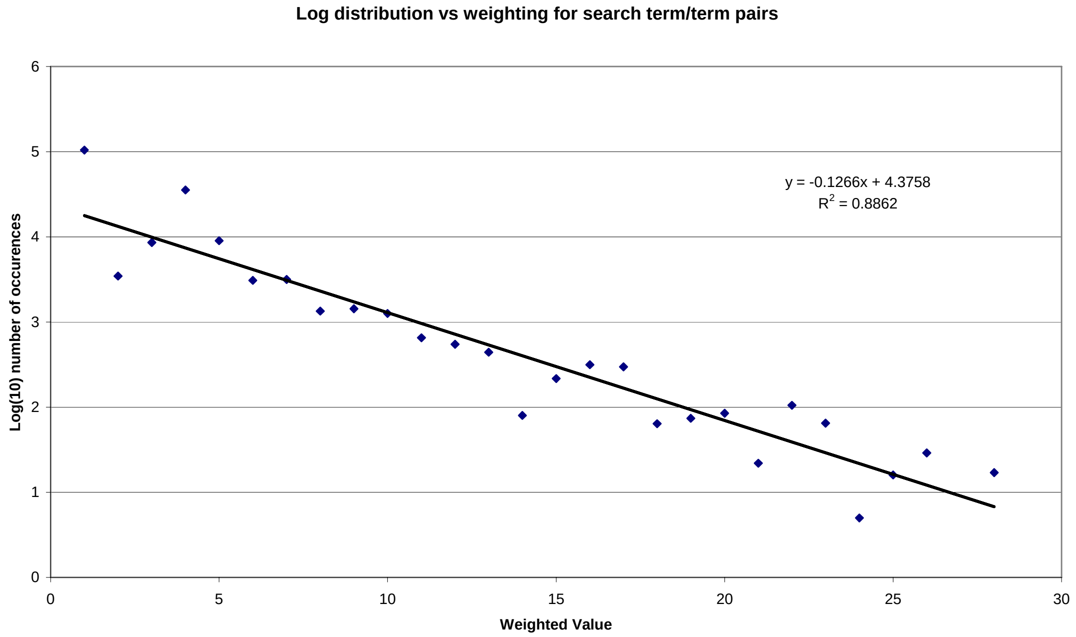


Figure 64: Distribution of weights

However in order to prove that this is occurring, and hence show that the highly weighted words are significant and not just common, a normalisation process was followed.

$$N_i = {}^{W_i}\!/_{F_i}$$

Equation 19

Where Ni is the normalised value, for the recovered concept/search string combination, Fi is the frequency of the recovered term in the corpus and Wi is the original weighted value for the concept/search string combination.

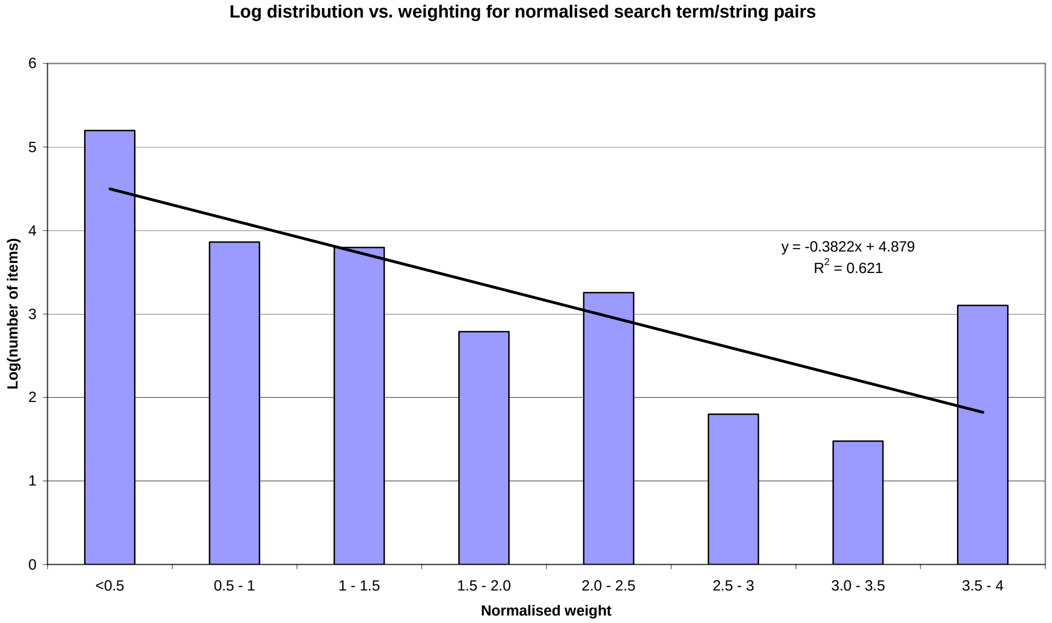


Figure 65: Distribution of normalised weights

The distribution of weights for this is shown in Figure 65. The concepts and search terms were then compared to the MeSH hierarchy. Some concepts do not exist in the obstetric part of the hierarchy, so the numbers of concept/search string pairs is reduced too. The search terms and concepts discovered are represented as xx.xx.xx type MeSH identifiers. It is unlikely that the discovered concept is going to be higher in the hierarchy than the search term for a number of reasons:

- The search terms tend to be from a high point in the hierarchy
- There are more lower level terms than high level terms
- Searching will usually discover terms lower than the search terms,
- Search strings used include high level and child terms in the hierarchy.

In order to discover relationships between the search terms and discovered concepts, the following algorithm was used:

```
For each Discovered Concept
    For each Search String
        Find number of occurrences where discovered concept is part of the search string
        Calculate the weight for each location.
    Next
Next concept
```

The final stage was to identify found terms that occur multiple times in the MeSH ontology – there were 2,631 of the possible 10,071 terms in the whole ontology that were located in the corpus that was examined. After that, search terms that had multiple locations were removed, to avoid confusion. See Table 48 for the distribution of the types of relation.

Table 48: Relationship counts

Relationship	Count
Child	132
None	2,631
Parent	2

Note that these totals are overlapping – a single term may be in all or one of these categories.

The weighted scores for each occurrence of each code were then summed and the range was 7.35 x10-03 to 4 this data was then stored in a database table. 24 Terms had occurrences with normalised weights> 0 for more than one location. The results are shown in Table 49.

6.4.8 Testing of the data

A large number of different methods have been proposed for determining the quality of the relevance of a querying system. The most commonly used method in the information retrieval community may be the F-measure, described by (van Rijsbergen, 1979). The F-measure is calculated using the precision (P) and the recall (R):

$$\text{F - measure} = \frac{2xPxR}{(P+R)}$$

Equation 20

However such an approach is not possible if there is no “Gold Standard” for the relevance of the retrieved set. In order to overcome this difficulty, the effect of combinations of terms in a query was studied. The details are given below, but the main aim is to use the change in the size of the retrieved set depending on whether the terms are Andy’s or Ord’s together.

Table 49: Discovered membership values for this corpus

Term	Code1	Value1	Code2	Value2	Code3	Value3	Code4	Value4
Anovulation	C13.371.056.630.050	10.00%	G08.520.440.508.080	90.00%		-		-
Blastula	A16.254.270.274	50.00%	A16.254.300.600.274	50.00%		-		-
Candidiasis, Vulvovaginal	C13.371.894.190	33.33%	C13.371.944.190	66.67%		-		-
Cervical Ripening	E04.520.252.968.100	77.78%	G08.520.769.326.100	22.22%		-		-
Cervix Incompetence	C13.371.852.150.280	60.00%	C13.703.039.089.339	40.00%		-		-
Cervix Neoplasms	C13.371.270.875.170	16.67%	C13.371.820.800.418.875.170	16.67%	C13.371.852.150.310	33.33%	C13.371.852.762.100	33.33%
Chorioamnionitis	C13.703.277.030	11.11%	C13.703.420.339.260	77.78%	C13.703.590.268	11.11%		-
Chorionic Villi	A16.254.403.473.200	76.47%	A16.759.189	23.53%		-		-
Cleavage Stage, Ovum	A16.254.270	50.00%	A16.254.300.600	50.00%		-		-
Depression, Postpartum	C13.703.844.253	33.33%	F03.600.300.350	66.67%		-		-
Dyspareunia	C13.371.665.313	20.00%	C13.371.894.217	20.00%	F03.800.250	60.00%		-
Fallopian Tube Neoplasms	C13.371.056.390.390	33.33%	C13.371.270.500	33.33%	C13.371.820.800.418.685	33.33%		-
HELLP Syndrome	C13.703.799.314.309	44.83%	C13.703.799.314.619.500	55.17%		-		-
Morula	A16.254.270.550	50.00%	A16.254.300.600.550	50.00%		-		-
Ovarian Neoplasms	C13.371.056.630.705	33.33%	C13.371.270.750	33.33%	C13.371.820.800.418.685	33.33%		-
Placenta Accreta	C13.703.420.643	25.00%	C13.703.590.609	75.00%		-		-
Postpartum Hemorrhage	C13.703.420.725	38.10%	E04.520.050.060.600	61.90%		-		-
Pregnancy Reduction, Multifetal	E04.520.050.060.600	50.00%	E04.520.050.600	50.00%		-		-
Trophoblasts	A16.254.085.162	80.00%	A16.759.802	20.00%		-		-
Twins, Dizygotic	G08.520.800.708.800	50.00%	M01.438.873.920	50.00%		-		-
Twins, Monozygotic	G08.520.800.708.838	50.00%	M01.438.873.940	50.00%		-		-
Uterine Neoplasms	C13.371.820.800.418.875	50.00%	C13.371.852.762	50.00%		-		-
Uterine Rupture	C13.371.852.904	40.00%	C13.703.420.904	60.00%		-		-

The calculation uses all 23 base terms identified in Table 49, along with the 119 terms that are related to each potential location. That is the related terms are either parents, siblings or children of these base terms in the base ontology.

Table 50: Record recovery for terms combined with relatives

Base Term	Average OR	Average AND	Ratio	Code
Blastula	3252.5	97	33.53	A16.254.270.274
Blastula	3252.5	97	33.53	A16.254.300.600.274
Cervix Neoplasms	57840	18165.5	3.18	C13.371.270.875.170
Cervix Neoplasms	57840	18165.5	3.18	C13.371.820.800.418.875.170
Cervix Neoplasms	57840	18165.5	3.18	C13.371.852.762.100
Cleavage Stage, Ovum	18574	297.41	62.45	A16.254.270
Cleavage Stage, Ovum	20608	988.67	20.84	A16.254.300.600
HELLP Syndrome	11125	939.5	11.84	C13.703.799.314.309
HELLP Syndrome	9722	922	10.54	C13.703.799.314.619.500
Morula	3062	441	6.94	A16.254.270.550
Morula	3062	441	6.94	A16.254.300.600.550
Ovarian Neoplasms	40595.07	3057.64	13.28	C13.371.056.630.705
Ovarian Neoplasms	38704.86	551.29	70.21	C13.371.270.750
Ovarian Neoplasms	38704.86	551.29	70.21	C13.371.820.800.418.685
Placenta Accreta	6309.43	169.86	37.15	C13.703.420.643
Placenta Accreta	2712	139	19.51	C13.703.590.609
Pregnancy Reduction, Multifetal	4529	346	13.09	E04.520.050.060.600
Pregnancy Reduction, Multifetal	4529	346	13.09	E04.520.050.600
Twins, Dizygotic	16456	3068.5	5.36	G08.520.800.708.800
Twins, Dizygotic	16456	3068.5	5.36	M01.438.873.920
Twins, Monozygotic	16456	4459.5	3.69	G08.520.800.708.838
Twins, Monozygotic	16456	4459.5	3.69	M01.438.873.940
Uterine Neoplasms	78731.71	16135.29	4.88	C13.371.820.800.418.875
Uterine Neoplasms	67827.5	21336.5	3.18	C13.371.852.762

The results are shown in Table 50. A small OR/AND ratio indicates that the terms are tightly coupled with the relatives, that is to say that it is more likely that they will occur together than separately. If this ratio becomes 1 then the terms only occur together. In Table 50 the terms are given along with the code for each possible location of that term. The average AND and average OR are calculated by calculating the arithmetic mean of the number of records retrieved by ANDing or ORing each of the base terms with each of the close relatives. Very high ratios imply that there is very little connection between the term and potential relatives. In the context of a fuzzy ontology this implies a very low membership value of the term at that location.

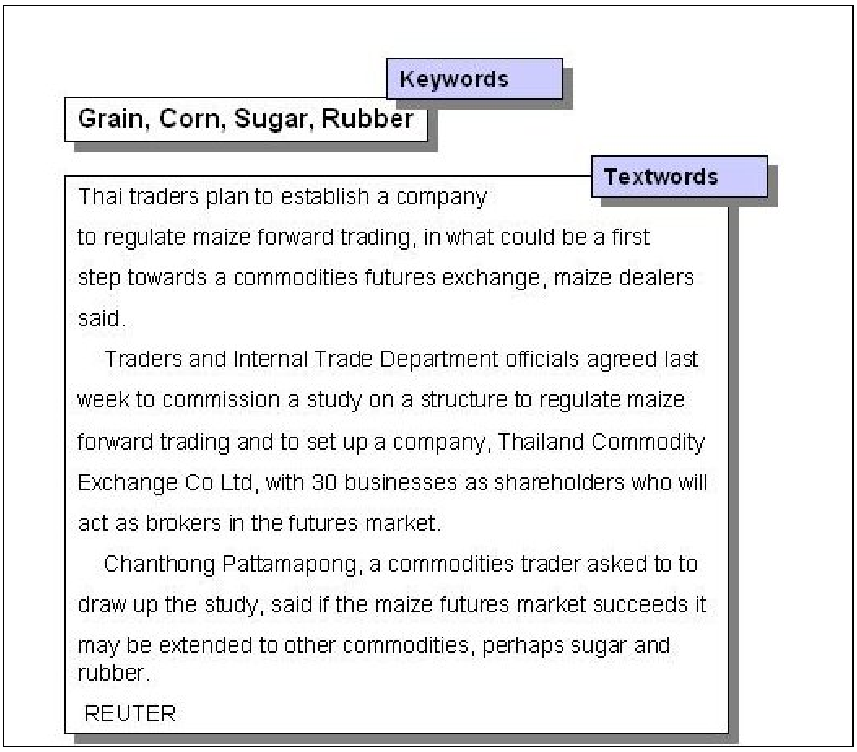
This technique can be compared with the various methods described previously (e.g. in section 6.4.5), and may appear to be somewhat circular. However, this approach relies primarily on the characteristics of the query engine – i.e. which documents are retrieved for a particular query, rather than the documents themselves. Obviously there is a link to the document contents as well as the documents being recovered in section 6.4.5 are recovered via queries, but combination queries may represent a more searching examination of the "librarian's" view than a simple single term query. Recalling the views of document model (Figure 20), the librarian in this section and the author have been investigated via the examination of the document and the query process, the next chapter deals with the reader.

6.5 The Reuters-21578 Corpus

The Reuters-21578 corpus comprises 21578 short reports (available from (Reuters, 1987)). They generally comprise commercial information and have been extensively used for experiments in text classification (Yang, 1999). Many of the reports have associated assigned index terms, and the download is organised to include these. The Apté categorisation of Reuters-21578 was used (C. Apté, F. Damerau, & S. Weiss, 1994)(ranging from 61 to 13715 characters long, mean 855) to identify reports that have valid index terms assigned, and these are split into the Apté testing set (3299 examples) and the Apté training set (9603 examples). A list of topic terms is also supplied, comprising 135 topic terms which are nouns, for example copper, copra-cake etc. Up to 14 topic terms may be applied to each report. Use of these sets allows comparison with other clustering and classification algorithms. A typical document is shown in Figure 66.

6.5.1 Fuzzy ontology and the Reuters-21578 Corpus

The training for the fuzzy ontology was performed using the Apté training set, with the ontology based around the 135 topic terms available. Those reports that did not contain any of the topic terms were excluded so the training set used comprised 8079 reports. The fuzzy ontology values were calculated using a similar technique to that described in section 6.2.1. In this case, however there was no existing structure, so a support calculation using a co-occurrence matrix including distance within the document was used. Only 83 of the possible topic terms occurred in the training set.



Grain, Corn, Sugar, Rubber

Thai traders plan to establish a company to regulate maize forward trading, in what could be a first step towards a commodities futures exchange, maize dealers said.

Traders and Internal Trade Department officials agreed last week to commission a study on a structure to regulate maize forward trading and to set up a company, Thailand Commodity Exchange Co Ltd, with 30 businesses as shareholders who will act as brokers in the futures market.

Chanthong Pattamapong, a commodities trader asked to to draw up the study, said if the maize futures market succeeds it may be extended to other commodities, perhaps sugar and rubber.

REUTER

Figure 66: One of the Reuters-21578 documents

The algorithm followed for the co-occurrence matrix calculation was:

```
For each topic string  in the collection
        For each document in the collection
        Discover  location
    Next document
Enter results into co-location matrix


Next term
For each term in co-location matrix ($T_a$)
    For each document in which term exists (C) from(1..s)
    For each topic string in overall list ($T_b$)
    Where $T_a$<> $T_b$ and $D_a$=$D_b$
        For each instance of $T_a$ (i) where i from (1..n)
            For each instance of $T_b$ (j) where j from (1..m)


            Calculate  particular distance TDij = $|L((T_a(i))- (T_b(j)))|$
            Next j
        Next i
    Calculate number of pairs(i*j)
```

A total of 249550 term-pairs were identified, and the mean term distance was calculated.

The final term mean distance (TD(Ta,Tb) is given by:

$$TD(T_a,T_b) = \sum_1^s \sum_1^m \sum_1^n \frac{(| \mathrm{L}((\mathrm{T_a}(c,i), (\mathrm{T_b}(c,j)) |)}{(\mathrm{L_c})}$$

Equation 21

Where:

L_C=Length of the current document in bytes

TD(a,b) = The overall term –term distance for each pair

L(a,b)= distance between instances of terms a and b where both a and b are in the same document.

n= number of instances of term a in document c

m= number of instances of term b in document c

s= number of documents containing both a and b.

The final co-occurrence table comprises 4538 term pairs. In addition to this calculation, the total support for each term was calculated, that is the sum of the pair support by each term.

Identification of all words in the Apté training corpus. All words longer than 3 characters and not on the stop word list described in (Retrieval, 2003) were identified.

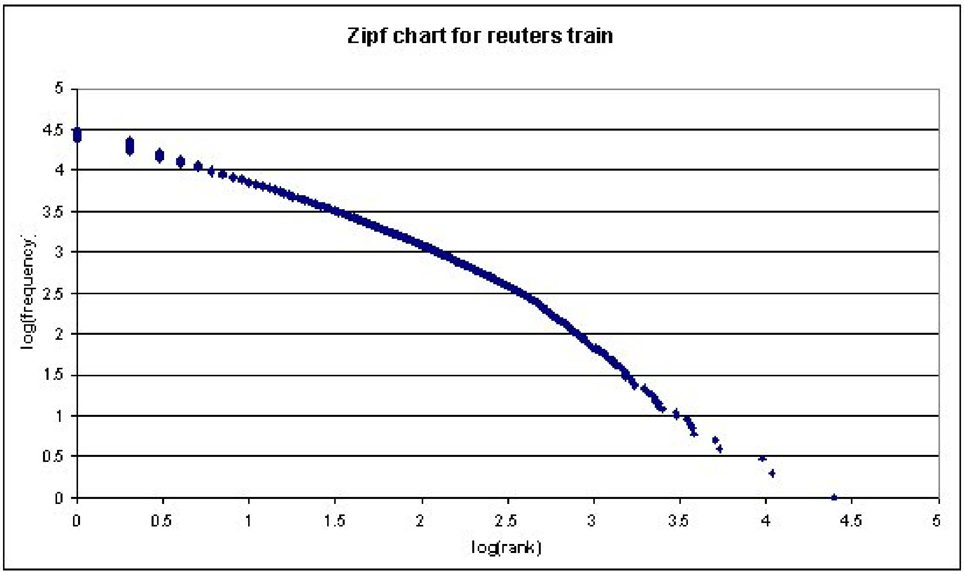


Figure 67: Zipf distribution of all words in Reuters training set

Expansion of the fuzzy ontology for non-index terms. In order to demonstrate the effectiveness of the fuzzy ontology approach, a similar term distance calculation was performed on non-index terms compared to index terms in the training corpus. Non-index terms were selected on the following basis:

- More than three characters
- Not present in the stopword list
- Not part of index terms
- More than 200 and less than 1000 occurrences in the corpus (roughly in the centre of the Zipf plot above.)

The term distance T_nT_k was calculated with T_n being a non-index term and T_k an index term. The calculation was performed as above, and the total support was calculated for each T_n. The term distance, T_nT_k was divided by Stot(T_n) in order to normalise the pair support to sum to a total of 1. This was done to avoid biasing future searches, and represents the relative degree of membership for each text word T_n with each index term T_k if each T_k is seen as a base member of the fuzzy ontology. Relations between T_k items were not considered in this case, because the calculation would not include them later.

A website with the relationship between each text word term included and each keyword is available at: http://csrs2.aut.ac.nz/fuzzont/default.html (see appendix, section 8.12).

Searches were then performed on the training set. The top five text words (in terms of membership value μ_{AB} where A is the index term and B is the other term were selected) for all A's where:

- There were at least 5 B's with μ_{AB} >=0.1
- There were some documents with corresponding keywords in the testing set.

The first criterion was selected in order to see the effect of increasing "ORs", where there are a reasonably small number of important related terms.

Table 51: Base terms and derived fuzzy relations

Base Term	Top Term	μ	Second term	μ	Third Term	μ	Fourth Term	μ	Fifth Term	μ
acq	Terms	0.18	Systems	0.16	merger	0.14	Management	0.12	proposed	0.12
dlr	companys	1.00	dlrs	0.96	mths	0.84	shrs	0.80	Qtly	0.75
heat	Soviet	0.16	Union	0.14	tonnes	0.13	crop	0.11	tonne	0.10
interest	interest	0.95	payments	0.39	payment	0.35	principal	0.19	prime	0.18
lead	manager	0.50	industrial	0.14	countrys	0.14	coupon	0.12	Swiss	0.10
rice	priced	0.95	index	0.48	commodity	0.39	fixed	0.38	coupon	0.35
sugar	intervention	0.18	French	0.15	program	0.13	quota	0.13	industry	0.13
tin	INsurance	1.00	meeting	0.92	meet	0.88	continue	0.87	continued	0.85
trade	surplus	0.55	deficit	0.52	plus	0.38	balance	0.37	Reagan	0.35
yen	bond	0.26	bonds	0.26	1992	0.24	dollars	0.21	Tokyo	0.17
Mean		0.57		0.41		0.36		0.33		0.31

The keywords were then used to query the testing set in two ways, one by using keywords from the training set to query documents in the testing set using the keyword field of the testing set and also using keywords against the body text. This latter approach was used to simulate searching of unindexed documents.

Table 52: Recall, Precision and f-measure for keywords from training set onto testing set

	Precision	Recall	F-measure
Keyword on keyword	0.85	1	0.92
Keyword in text	0.58	0.46	0.51

F- Measures (see section 6.4.8) were used to check the quality of these searches. The results in Table 52 demonstrate the considerable loss in both recall and precision when the query is run against text terms rather than index terms only.

6.5.2 Results

The most closely related text word terms were used to search the text of the testing set, in a series of OR operations, where the first run was performed just using the highest μ text word term (from Table 51). These terms are named "dependent terms", represented as DT_1,DT_2 etc. The next run used this term Ord's with DT_2, the following run Ord's DT_1 ,DT_2 ,DT_3 and so on.

F-Measures of queries using keyword terms from the training set were calculated from queries searching the testing set text words using the combinations described above and the f-measures calculated. The results are shown in Table 53. Interestingly the f-measures generally increase as more of the terms are Ord's together, although the overall results are generally low. This implies that the fuzzily related terms may be useful for query formulation, but the index terms themselves are more useful. To discover whether this was the case a second experiment was performed, where the index term was also included in the query. The aim of this investigation is to see whether terms not previously selected as index terms in the data set are themselves useful – this combines the librarian view with the author view, as the author chose the words in the document but the indexer chose the index terms.

The results for the second experiment are displayed in Table 54. In this experiment the dependent terms were Ord's in the same sequence, but with the keyword term ("keyterm"), included in the query with "OR". Although higher, the f-measure values are still lower than obtained using the keyterms against the text of the testing set, This is due to the lower precision obtained, as the recall must be equal or greater than that obtained.

Table 53: F-measure values for Ord's related terms

Base term	Top term	Top two terms	Top three terms	Top four terms	Top five terms
acq	0.24	0.33	0.52	0.58	0.61
dlr	0.00	0.01	0.01	0.00	0.00
heat	0.00	0.00	0.00	0.00	0.00
interest	0.18	0.14	0.12	0.11	0.24
lead	0.00	0.00	0.00	0.00	0.00
rice	0.00	0.00	0.04	0.04	0.03
sugar	0.12	0.09	0.05	0.05	0.05
tin	0.00	0.03	0.03	0.03	0.03
trade	0.61	0.76	0.68	0.66	0.68
yen	0.05	0.03	0.03	0.05	0.05
Mean	**0.12**	**0.14**	**0.15**	**0.15**	**0.17**

Table 54: F-Measure values for keyterm in text Ord's with related terms

Base term	Top term OR keyterm	Top two terms OR keyterm	Top three terms OR keyterm	Top four terms OR keyterm	Top five terms OR keyterm
acq	0.24	0.33	0.52	0.58	0.61
dlr	0.00	0.01	0.01	0.00	0.00
heat	0.00	0.00	0.00	0.00	0.00
interest	0.45	0.40	0.35	0.34	0.39
lead	0.15	0.10	0.09	0.08	0.07
rice	0.76	0.58	0.50	0.40	0.34
sugar	0.98	0.76	0.52	0.45	0.33
tin	0.00	0.03	0.03	0.03	0.03
trade	0.98	0.97	0.92	0.90	0.89
yen	0.05	0.03	0.03	0.05	0.05
Mean	**0.36**	**0.32**	**0.30**	**0.28**	**0.27**

6.5.3 Fuzzy querying results

However, these experiments do not include the fuzzy membership value in the query. A simple scheme was devised to include the relative importance of a dependent term that would be based on its relatedness to the original keyterm and the degree to which it is present in a particular document. Thus a term which is closely related to the target, but rare, may give less importance to a retrieved document than a term which is less-closely related but common in desired documents. This approach may give flexibility in cases where the most closely related terms to obscure terms are themselves obscure. In order to use this value, the test documents were re-examined and the number of occurrences of each dependent term (DT_1 …DT_5) from each keyterm were recorded.

The number of occurrences of each DT in each document were multiplied by the membership value for each DT, to obtain the "Importance" of this term in each document.

$$I_i(D) = \sum n * \mu$$

Equation 22

Where $I_i(D)$ is the importance of DT_i in document D, n is the number of occurrences in D of DT_i and μ is the membership value of DT_i in respect to the desired term.

Table 55: Combinations with Importance >0.5

Base term	Keyword Only	Top term OR keyterm	Top 2 terms OR keyterm	Top 3 terms OR keyterm	Top 4 terms OR keyterm	Top 5 terms OR keyterm
acq	0.00	0.00	0.00	0.08	0.11	0.13
dlr	0.00	0.00	0.01	0.01	0.00	0.00
heat	0.00	0.00	0.00	0.00	0.00	0.00
interest	1.00	0.35	0.26	0.22	0.22	0.39
lead	1.26	0.00	0.00	0.00	0.00	0.00
rice	1.47	0.00	0.00	0.00	0.00	0.00
sugar	1.51	0.00	0.00	0.00	0.00	0.00
tin	0.00	0.00	0.03	0.03	0.03	0.03
trade	1.51	0.76	0.86	0.82	0.81	0.83
yen	0.00	0.04	0.02	0.02	0.02	0.01
Mean	**0.68**	**0.11**	**0.12**	**0.12**	**0.12**	**0.14**

The results of this analysis are shown below in Table 55. The first column gives the F-measure values for the index term only, versus the text of the test set. It can be seen that the F-measures are reduced even further by this approach, if the threshold is lowered to 0.2 the results are as shown in Table 56.

Table 56: Combinations with Importance >0.2

Base term	Top term OR keyterm	Top 2 terms OR keyterm	Top 3 terms OR keyterm	Top 4 terms OR keyterm	Top 5 terms OR keyterm
acq	0.24	0.33	0.52	0.58	0.61
dlr	0.00	0.00	0.01	0.00	0.00
heat	0.00	0.00	0.00	0.00	0.00
interest	0.35	0.23	0.17	0.16	0.30
lead	0.00	0.00	0.00	0.00	0.00
rice	0.00	0.00	0.04	0.04	0.03
sugar	0.09	0.05	0.03	0.04	0.05
tin	0.00	0.00	0.03	0.03	0.03
trade	0.76	0.68	0.74	0.62	0.63
yen	0.05	0.00	0.03	0.05	0.06
Mean	**0.12**	**0.10**	**0.14**	**0.13**	**0.16**

This approach improves the F-measures somewhat, but overall, this approach is not successful mostly because of low recall values. Recalling the views of document model (Figure 20), the librarian in this section and the author have been investigated via the examination of the document and the query process, the next chapter deals with the reader. It may be for search optimisation purposes that the most useful aspect of this sort of analysis is to expand queries when the initial yield is too small. However, intelligent searching often requires the user to make decisions about the query refinement process and this approach may be most useful for identifying potential expansion terms that the user can themselves choose.

Chapter Seven

The case study component of the thesis is described in this chapter. The case study involved the experimental deployment and usability testing of an implementation of fSearch in an academic department of Obstetrics and Gynaecology. The usability was measured via questionnaires and observation. The use of the fSearch system was logged and analysed. Section 7.1 introduces the concepts and section 7.2 deals with the background to the usability evaluation measures adopted and some of the experimental details. Section 7.3 is concerned with the results of the testing. Experimental results are contained in this section for both usability and the learned ontology.

7.1 Introduction

The purpose of this chapter is to describe the use of the implemented system in practice. The system tested comprised a query parser based on UMLS, sending queries to PUBMED and GOOGLE and returning documents via a browser that allowed manual assignment of usefulness scores and term selection for relatedness scores. Although the system has been described as an intelligent system for medical information retrieval, the practical work has focussed exclusively on the domain of Obstetrics and Gynaecology. This has been done in order to reduce the size of the database's needed, but mostly because of the availability of suitable subjects, and the availability of suitable facilities in the University of Auckland Department of Obstetrics and Gynaecology at National Women's Hospital. Ethical approval was obtained from the AUT ethical committee - code *03/81 Investigation of usability of medical searching system.* The information sheet given to subjects, and the consent form the subjects were asked to sign is shown in appendix ethics. The site chosen for the investigation had Internet access and computers available.

One of the interesting aspects of the domain of Obstetrics and Gynaecology is the wide variety of people and specialties involved in it. Major groups involve medical Obstetricians and Gynaecologists, Midwives, Researchers from these backgrounds, and from basic sciences are also associated with it. In addition, General Practitioners and other primary healthcare providers are often involved in the care of pregnant women. Last but not least, women themselves who currently are, planning to be, have been or wishing to avoid being pregnant are a huge audience for specifically obstetric information. Partners, friends and relatives are also interested in this area. Members of the public and patients were not involved in the usability study because of ethical issues at this stage.

Within each professional group, there are different roles and ranks. Medical staff range from House officers/Senior House Officers (HO/SHO), Registrars to Specialists. Ranks up to specialist are known as ‘junior’ doctors. Specialist may be employed as academics (which involves clinical duties as well as teaching and research) or on clinical contracts. Midwives have less formal rank titles, but may also be involved in research and teaching as well as clinical duties. Both Obstetricians and Midwives care for women in pregnancy and labour but as a general rule Obstetricians are more likely to involved in the care of women with complications of pregnancy. In terms of information requirements, obstetricians are more likely to require the sort of information that comes from clinical research journals such as those indexed in MEDLINE. Midwives may focus more on the information that is indexed in CINHAL for example. However, there is considerable overlap, and clinical staff of all types may well use sources such as books, online guides and conference reports.

Clinical staff need to communicate with patients, and vice versa, and there is a wide variety of material available for patients on the Internet. Thus clinical staff may wish to be able to identify documents that they would recommend to patients, especially for those who have rare or particularly difficult conditions. General practitioners in particular, are often called upon to explain a wide variety of conditions to patients without themselves having a great deal of experience in the field. All people involved in healthcare have a huge challenge in dealing with the torrent of information being produced in the field.

Considering information requirements, researchers are likely to have more time to consider their search strategies and reviewing the documents retrieved. Clinical staff are likely to have initially less time available but may be satisfied with simpler results to answer a particular question. Given that a standard consultation period is around 15 minutes or less, and clinical examination and discussion all have to happen in this period, rapid results are important for busy clinicians.

7.2 Usability measures

Useability assessment was performed on the fsearch implementation in an number of stages – firstly to identify “usability bugs” via a simulated use with a single user, and then more general evaluation with a collection of users who were observed using the system. Finally the users were asked to rate the

Initial Usability Experiment

Performed with one of the collaborators, this was in the nature of an expert walkthrough.

The subject was a 35-year-old Obstetric Specialist with a self-assessed "Advanced Beginner" knowledge of searching. She chose the "HELLP Syndrome" item to investigate.

A number of software bugs were identified:

- The "analysis window" did not clear on moving between pages
- Saving the usefulness score caused an error

The following usability points were identified:

- Browser windows were too small to see the entire width of a standard page
- The individual word analysis, separated naturally paired words – such as "Blood Pressure"
- The analysis window did not allow enough items to be displayed.
- Punctuation marks should have been removed from the analysis window words
- Assigned words i.e. those moved into the category boxes, should be retained between pages.
- An ignore button should be provided for pages that are similar to others.
- The "age" category during login was confusing.
- The keyword tree entry to the system was far too long, and a supported browse for terms should be substituted.

Overall it was interesting that the subject initially believed that the system would only be useful for SHO's or patients, and that registrars and above would need to use PubMed directly. However, one of the pages retrieved was deemed suitable for registrars use (Padden, 1999). The time taken to identify the related terms in each page was around 10 minutes, but an assessment of quality was made more quickly.

The following actions were taken in response to this experiment:

- The software bugs were fixed
- The time for each users testing was fixed at around 30 mins actual use, with around 10 mins teaching and question selection beforehand and a 5-20 minute debrief. This effectively only allows one set of pages to be used.

7.2.1 Questionnaire selection

The questionnaire used in the study was adapted from one of those generated from the site provided by Perlman (Perlman). The questionnaire is shown in Appendix 8.11. This questionnaire was originally reported in (Davis, 1989), and has subsequently been used in a number of studies. This questionnaire is based around the Technology Acceptance model (TAM). This questionnaire focuses on the use of a system for work –related tasks, and the scale runs from -2 to +2, to allow a 0 for neutrality. All the questions are phrased so that a positive result implies satisfaction with the system. One of the most interesting aspects of this questionnaire is that it specifically links ease of use and usefulness. The original work suggested that increased perceived ease of use has a causal influence on perceived usefulness. This model has also been applied to the world wide web and (Lederer, Maupin, Sena, & Zhuang, 2000), appears to suggest that this analysis is not complete. It could be suggested that in turn, that an information system that is perceived as useful must be retrieving useful information.

7.2.2 Initial group

Four subjects were recruited for the initial testing. User profiles of these subjects are shown in Table 57.MFM=Maternal Fetal Medicine (High Risk Pregnancy), REI= Reproductive

Table 57: First group details

Number	**Job description**	**Professional group**	**Computer experience**	**Gender**	**Age Range**
1	Senior Academic	Doctor, interest in MFM	Moderate	Male	50+
2	Senior Academic	Doctor interest in MFM	Moderate	Female	50+
3	Junior Academic	Doctor, interest in REI	High	Female	30+
4	Research Midwife	Midwife background, clinical researcher	Moderate	Female	50+

The setting for this investigation was the Academic Obstetrics department at National Women’s Hospital. The desktop computer used was running Windows 2000 with a screen resolution of 800x600. Subjects read the information sheet (see section 8.10.1 in the appendix) and signed the consent form (section 8.10.2 in the appendix). The developer acted as a facilitator and recorder of the interaction. The users were encouraged to comment on any difficulties or comments they might have as they went through the procedure. There was a debrief after the procedure finished, and the users completed the questionnaire. The users were recruited by means of an announcement in the staff meeting.

7.2.3 Tasks

The users were asked to perform the following tasks:

a) Log into fSearch and select the appropriate demographic, and area of interest.

b) Perform a search using the "obstetric" keyword on the Google interface. This was done mostly to familiarise the users with the system, in particular the appropriate use of the mouse and the use of the + anchor in lists to expand them, as in windows explorer.

c) They then performed another search using terms of their own choosing, again using the Google interface.

d) With the open browser windows, they were asked to rate a number of the pages shown in terms of usefulness via the slider. They were also asked to perform a term extraction and selection analysis on pages that they rated highly. In most cases this amounted to around five pages.

e) They were then asked to perform a similar information finding task using the full MeSH tree.

The results for each user in terms of satisfaction are shown in Table 58.

Table 58: Initial group satisfaction -Please refer to the appendix for the full questions

Question	User 1	User 2	User 3	User 4
Perceived Usefulness				
1 (Quick)	0	2	1	1
2 (Performance)	0	1	0	0
3 (Productivity)	2	1	1	0
4 (Effectiveness)	1	1	0	0
5 (Easier)	1	1	0	1
6 (Useful)	1	1	2	1
Perceived Ease of Use				
7 (Easy to Learn)	2	1	2	2
8 (Easy to Control)	-1	2	1	2
9 (Clear Interact)	2	2	1	2
10 (Flexible)	-1	2	1	2
11 (Skill)	2	2	2	2
12 (Easy to Use)	0	2	1	2

Comments about the system are recorded below but general satisfaction seemed quite high. As the developer was also the usability test facilitator there could have been some bias introduced by the user's trying to please the developer, but efforts were made to minimise this by emphasising that the system was expected to have problems, that the user was not on test, but the system was and that anonymity was preserved. At the start of the session the user was given a brief guide to the system, following a set of instructions in order to bring up a search on the term "obstetrics". Of particular interest was the ease of use of the analysis system, despite the fact there were bugs in this version that allowed duplicate words to occur in the pick list. There was certainly a preference towards identifying positive (very, somewhat relevant), rather than negative (irrelevant, unwanted) words. The users preferred to analyse those documents they found useful, and tended to ignore those they found useless.

Comments by user 1

"Need a hierarchy of references with pub med having priority, unclear how priority decided."

"Word versus phrase searching"

"Can this be linked to Cochrane?"

"Fast", "Clear", "Scope is wide", "Learning system?"

Generally this user saw the system as potentially useful, but identified particular technical issues that would need to be resolved before the system would be generally usable.

Comments by user 2

"Broad range of resources accessible especially information for women".

This user was particularly impressed by the common interface to multiple databases or information sources allowed for by this system.

Comments by user 3

"Short time needed to learn the software, wide variety of sites searched – not confined to medical sites".

Again, this user felt that the access to information not contained in PubMed or Cochrane was helpful.

Comments by user 4

"Need some background knowledge on searching", "Easy", "Clear", "Time saver". The user appeared very happy with the system and felt that it would improve their searching ability.

General observations of users included the fact that they found dealing with large numbers of windows a little confusing. By attempting to improve visibility, the use of multiple windows tended to remove the obvious focus. Mouse movement became more uncertain when there were overlapping windows and the users were often uncertain as to the difference between closing and minimising windows. In many cases the users maximised the active window.

One of the recurring themes was the uncertainty of whether such a system was primarily for medical professionals or for patients. When browsing the documents recovered via Google the users were sometimes surprised to find what they regarded as legitimate medical pages amongst the obviously patient-centred ones. This is an unexpected benefit of using multiple search engines – multiple search strategies are used simultaneously. Various meta-engines already use this approach, but they currently do not appear to use non-commercial data sources such as PubMed.

Particular usability “bugs” that were discovered in this session included the following:

- Slow response when phrases were searched for.
- A confusing number of windows being displayed.
- Uncertainty as to the arrangement of the MeSH Tree display.
- The analysis window remains open after its parent browser is shut.
- Some terms not present in the pre-loaded analysis

Particular efforts made to resolve these issues include:

- Use of phrase detection from the MRCXT table, so that not every combination is located initially, e.g. “Premature Labour” searches for the phrase as well as “premature” and “labour”.
- Window numbers reduces, and different icons chosen for the minimised version
- The analysis form made into a MDI child of the browser
- Online term searches made available.
- More general improvements were made using improved indexing of large tables, use of ADO.NET rather than ADO, and inclusion of the PubMed search and display as an option.

7.2.4 Second Group

There were four users in the second group – see table 48, unfortunately the system crashed during the fourth users test because of a corrupt database and no searches were performed by that user. It should be noted that the system was still running in a "learning" mode – that is it recorded the preferences etc. but was not actually presenting documents based on a learned fuzzy ontology. The software was a slightly later version with some improvements in speed, but no changes to the interface. The second set of tests took place around a month after the first ones.

Table 59: Second group details

Number	Job description	Professional group	Computer experience	Gender	Age Range
5	New Consultant	Doctor, General Obstetrics and Gynaecology	Moderate	Female	30+
6	Senior Academic	Doctor, interest in Contraception	High	Female	40+
7	Junior Academic	Doctor, interest in Infertility	Moderate	Female	30+
8	New Consultant	Doctor, General Obstetrics and Gynaecology	Moderate	Female	30+

Comments by user 5

"I may get lost if nobody helping me."

"Seems simple to use"

"Would be good to avoid all unwanted sites"

This user was particularly impressed with the idea of a learning system, and being able to exclude useless sites. The "lost in hyperspace" phenomenon is still present here, and some sort of support for current and past location may be in order, possibly along the lines of "scent trails" as discussed in section 2.1.4.

Comments by user 6

"Unable to find very specific topics related to contraception"

"Easy to navigate through"

"Came up with some interesting pages on menopause"

This user had a very specific set of requirements that she was using to assess the quality of documents that she was searching for. In particular she was looking for references to the change in advice to women as part of the consequences of the Women's Health Initiative (WHI) Study (Manson et al., 2003). This was a particularly good "litmus test" as this recent, large study concluded that Hormone Replacement therapy (HRT) did not provide cardiovascular benefits despite theoretical reasons why it could. Since this report came out approximately a year before the usability test, and many organisations have changed their recommendations in response to it, references to the previous advice, based on the incorrect perception are not only misleading and potentially harmful, but may indicate a generally lackadaisical attitude on the part of the document producers. This would then call into question the rest of the advice or information produced from that source.

However this user was also impressed with the ease of use of the system although she wanted finer resolution than the MeSH headings gave for particular searches.

Table 60: Second group results

Question	User 5	User 6	User 7
Perceived Usefulness			
1 (Quick)	1	2	2
2 (Performance)	1	2	2
3 (Productivity)	1	2	2
4 (Effectiveness)	1	2	2
5 (Easier)	1	2	2
6 (Useful)	1	2	2
Perceived Ease of Use			
7 (Easy to Learn)	2	2	2
8 (Easy to Control)	2	1	2
9 (Clear Interact)	1	1	2
10 (Flexible)	1	1	2
11 (Skill)	1	2	2
12 (Easy to Use)	1	2	2

Comments by user 7

“[Good] Key word choices”

“Ease of Use”

This user was generally happy with the system. One aspect of particular benefit was the presentation of the derived MeSH keywords, which allowed the user to reconsider her search before it began. There has been a tendency to avoid presenting keywords to users – for example both PubMed and Google avoid doing so, but it allows an implicit “more” or “less” like this to be constructed when the index terms are displayed as in the ACM digital library or IEEE Xplore which actually has them as hyperlinks (see Figure 69).

The combined results are shown in Figure 68. The mean values are always above zero, which appears to indicate general satisfaction with the system. The values for ease of use are generally higher than those for perceived usefulness, which at least indicates that the system can be used – and would hopefully be found to be useful over time, as the observational rather than ethnographic usability test represents a simulation, rather than actual use.

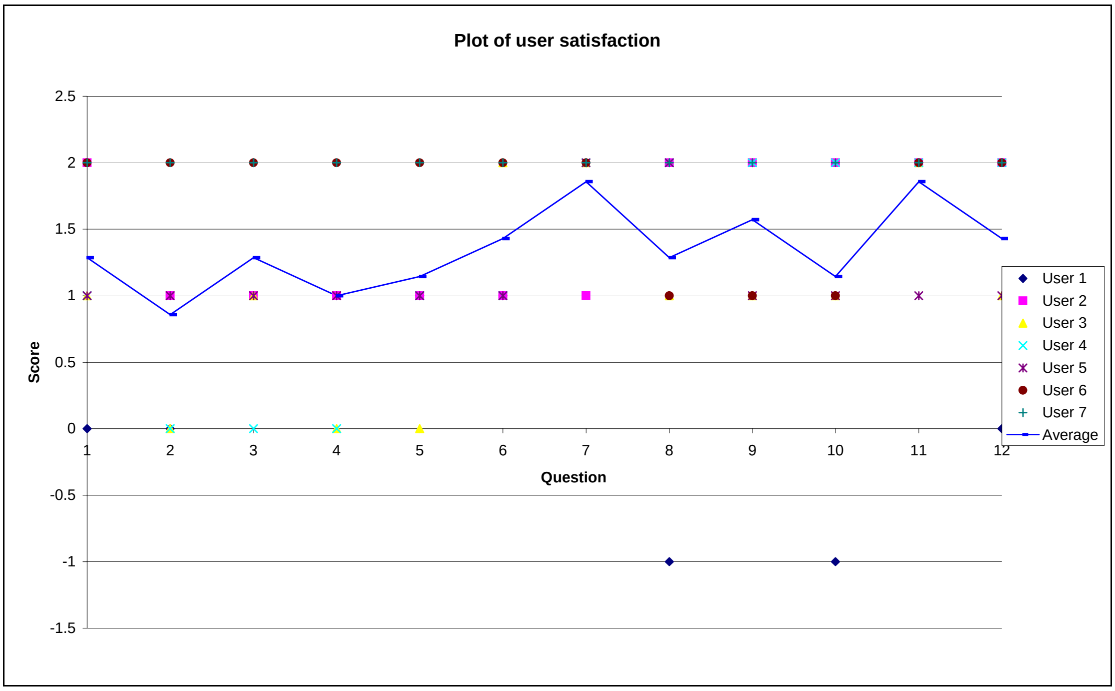
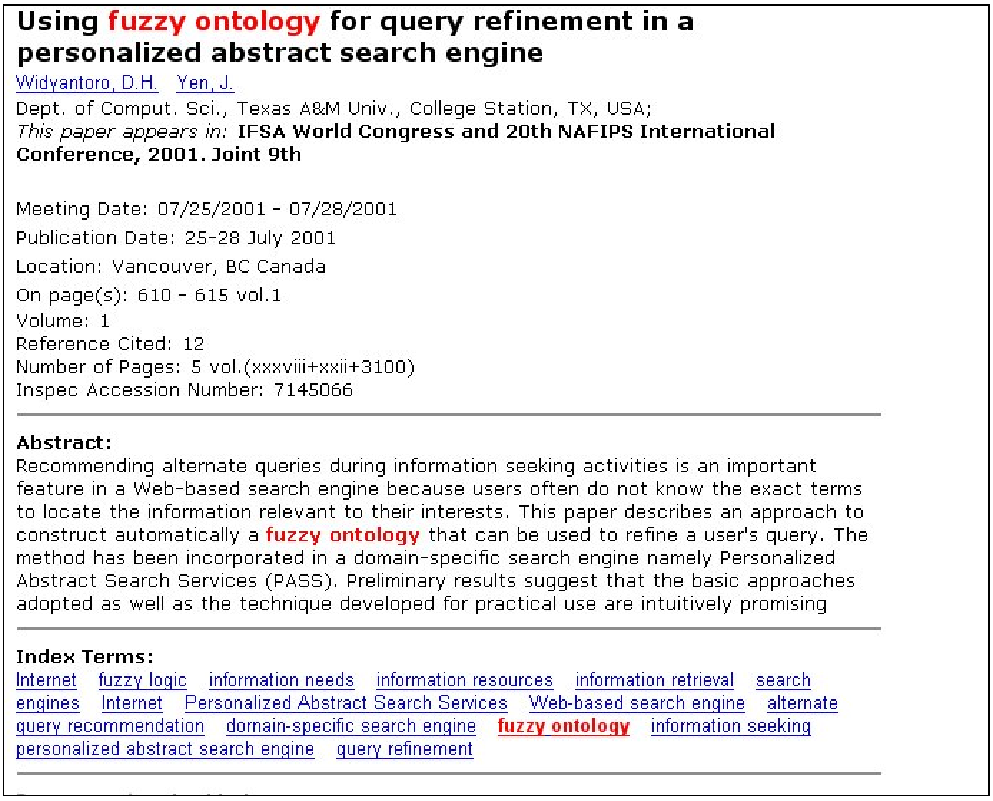


Figure 68: Plot of overall satisfaction



Using fuzzy ontology for query refinement in a personalized abstract search engine

Widyantoro, D.H.   Yen, J.

Dept. of Comput. Sci., Texas A&M Univ., College Station, TX, USA;

This paper appears in: **IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th**

Meeting Date: 07/25/2001 - 07/28/2001

Publication Date: 25-28 July 2001

Location: Vancouver, BC Canada

On page(s): 610 - 615 vol.1

Volume: 1

Reference Cited: 12

Number of Pages: 5 vol.(xxxviii+xxii+3100)

Inspec Accession Number: 7145066

Abstract:

Recommending alternate queries during information seeking activities is an important feature in a Web-based search engine because users often do not know the exact terms to locate the information relevant to their interests. This paper describes an approach to construct automatically a **fuzzy ontology** that can be used to refine a user's query. The method has been incorporated in a domain-specific search engine namely Personalized Abstract Search Services (PASS). Preliminary results suggest that the basic approaches adopted as well as the technique developed for practical use are intuitively promising

Index Terms:

Internet   fuzzy logic   information needs   information resources   information retrieval   search engines   Internet   Personalized Abstract Search Services   Web-based search engine   alternate query recommendation   domain-specific search engine   **fuzzy ontology**   information seeking   personalized abstract search engine   query refinement

Figure 69: IEEE Xplore –note index terms

7.3 Results of Use

Along with the perceived ease of use and usefulness, the usability testing was an opportunity to collect data from users in order to assess the viability of the learning schemes.

7.3.1 Rating of Pages

The users rated a total of 98 Pages during the case study. The distribution of the ratings is shown in Figure 70. The users rated some pages at 0% - that is “useless”.

7.3.2 Term association

The users selected a total of 138 terms for analysis– see Table 61.

Table 61: Terms that were rated

Relatedness	Number of terms
Strong	63
Somewhat	32
Slightly	24
Unrelated	9
Opposite	10

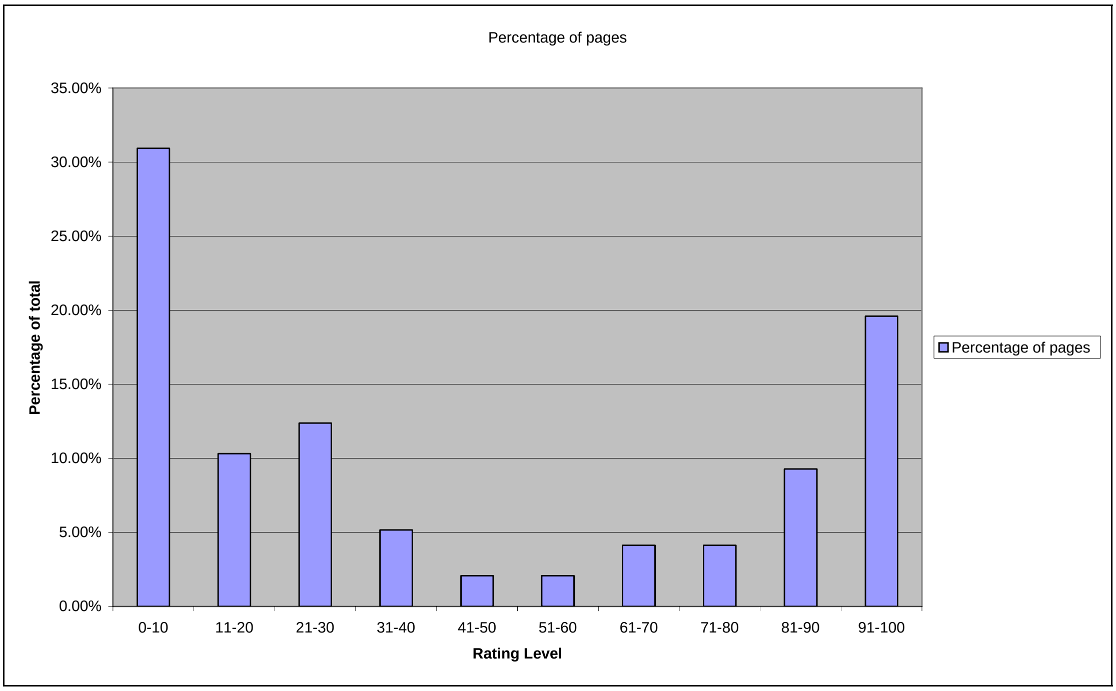


Figure 70: Distribution of ratings

Note that the users were much more likely to rate “positive” relations rather than negative ones. This may have been due to bias, or particularly effective search – as with all learning systems based on user feedback to cases the pattern of feedback is at least partly determined by the examples shown. The figure suggests that the point from (Jin et al., 2004) that user’s rating scheme can be very different, is worth noting, although a direct comparison of the inter-user variability could not be made as the data was anonymised.

Relating the terms rated to the original search string, Table 62 shows the strongly related terms. No term/search string combination occurred more than once.

This data supports the idea that rating is possible, and in fact term and relation harvesting may be quite efficient – this represents only around 3 and ½ hours of user time.

In summary, the users generally felt the system was both useful and usable. They were able to understand the method of use easily and they were able to both rate the usefulness of retrieved pages and the degree of association between terms in the retrieved document and their desired concept of interest. This suggests that the third pillar of the view of the document, the reader, will be able to assist in the development of a fuzzy ontology

Table 62 Terms "strongly related" to search strings

Word	Search String
AFTER	Vaginal Birth after Cesarean
AMNIOCENTESIS	Chorioamnionitis
AMNION	Chorioamnionitis
AMNIOTIC	Chorioamnionitis
BIRTH	Vaginal Birth after Cesarean
CESAREAN	Vaginal Birth after Cesarean
CESAREAN/VBAC	Vaginal Birth after Cesarean
CESAREANS	Vaginal Birth after Cesarean
CHORIOAMNIONITIS	Chorioamnionitis
CHORION	Chorioamnionitis
DELIVERY	Vaginal Birth after Cesarean
INFECTION	Chorioamnionitis
INFECTIONS	Chorioamnionitis
INFLAMMATION	Chorioamnionitis
INTRAAMNIOTIC	Chorioamnionitis
MEMBRANES	Chorioamnionitis
OBSTETRICAL	Vacuum Extraction, Obstetrical
OBSTETRICIANS	Vacuum Extraction, Obstetrical
PREVIOUS	Vaginal Birth after Cesarean
PROS	Episiotomy
PROTEIN	Vacuum Extraction, Obstetrical
QUADPLETS	Quadruplets
QUADRUPLETS	Quadruplets
QUADRUPLETS'	Quadruplets
QUADS'	Quadruplets
SECTION	Vaginal Birth after Cesarean
VAGINAL	Vaginal Birth after Cesarean
VBAC	Vaginal Birth after Cesarean
VBACS	Vaginal Birth after Cesarean

Chapter Eight

This thesis concerns the optimisation of an information retrieval system for heterogeneous data sources to provide useful information in a medical domain. This chapter is the conclusion and discussion. Section 8.1. describes some analysis of the results. Section 8.2 deals with the possible use of structural fuzzy ontology. Some research questions are discussed in section 8.3. More general representation of ontologies and the relationship with OWL are discussed in section 8.4. Possible future approaches for further work are introduced in section 8.5. The overall conclusions are drawn in section 8.6.

8.1 Analysis of results

The results from both the automatic clustering and human simulation are included here. As with all information retrieval work, and indeed machine learning, performance of algorithms and systems is at least partially dependent on the data set, and the method of judging performance.

In the review (Kodratoff, 2001) comparing machine learning (ML) and KDD the author makes the point that there is an epistemological difference between KDD and ML. If the process modifies behaviour of a human or a mechanical system then it is KDD. ML contributes to these changes, but the criteria for success are different in that ML is generally assessed in terms of its statistical performance whereas KDD is assessed in terms of whether it does something useful. This analysis is not universally accepted but, there is an important point here especially in regard to the usefulness of such systems. Even very poorly performing ML tools can be useful in the KDD process – for example if an IR system produces one good result in a list of ten, if that result is what you need then it has been a success, especially if the other examples are obviously not useful. The reverse also occurs – if a ML tool has very good performance in terms of classification accuracy, but the knowledge produced is not useful, then it is no use in terms of KDD. An example of this might be a speech recognition system with word recognition at around 99%, for legal letters this system is still unacceptable without editing, as there is effectively a zero tolerance for errors. This leads to the understanding that the use of ML needs to fit within a process that allows its quality and performance to be continuously assessed. It is also important to be able to override such systems when they are not working usefully and in a perfect system such events are used by the system to learn to perform better. ML systems are by their nature difficult to make self-reflective and it is generally a mistake to expect them to be able to audit themselves in an unbounded domain.

Information retrieval systems have often been analysed in term of precision, and recall and f-measures against various standard data sets. However there is some doubt as to whether such analysis is the most appropriate, and new forms of assessment are being proposed (Allan et al., 2003). This reference is extremely useful as a “wish-list” for improvements in information retrieval systems and measurement of those systems and modifications that are needed to bring the worlds of Information retrieval and the WWW together. The challenges of using diverse information sources from the web have been raised in (Allan et al., 2003), and this thesis has attempted to do so in terms of the harvesting of corpora (see sections and 4.3.2, 4.4.1 and 6.1.2) from the web, which occur in many divers formats, for example HTML, PDF and text extracted from these pages. User modelling is also identified as an important area in (Allan et al., 2003), and in particular the importance of providing integrated search systems that allow easy incorporation in the users’ workflow – this is the aim of the fSearch - and the development (Chapter Five) and testing (Chapter Seven) of it forms a major part of this thesis. The importance of not using “approaches that are ‘good enough’ for everyone and therefore ‘never great’ for everyone”(Allan et al., 2003)(page 38) is recognised in this thesis in particular in the development of the fuzzy ontology approach (Chapter Three). Identification of aspects of documents that may be important for different retrieval models was described in the framework for useful information in Chapter 1. Particular ways of discovering authorship or source, automatically, by means of information theory and Kolmogorov distance were discussed and tested in Chapter Four. Some problems of distributed search were tackled by the use of web services, (section 5.2.7) and the use of an XML –based representation of the fuzzy ontology (section 5.4). The need to include learning in a system for filtering and result presentation was explored in the learning algorithms for the fuzzy ontology, (section 3.3, and 6.2) as well as hierarchical and other clustering techniques applied using the information theory approach (section 6.4).

8.1.1 Kolmogorov distance

The majority of results in this area are contained in Chapter Four, and Information entropy was demonstrated to be effective at identifying authorship in simple electronic documents in section 4.2. It may be that such an approach would be fruitful in other problems such as forensic software or email analysis or even spam detection, along with the potential for its use in bioinformatics, and language identification described in the original paper (Benedetto et al., 2002). The extension of this approach to identification of the domain of origin of web pages as described in section 4.3 is more directly relevant to this thesis. Clustering via this method, as described in section 4.5 is possible, but the practical identification of useful clusters is difficult without more extensive use of the manual system to identify popular, useful styles. This approach, although attractive because of its language and domain independence is more likely to be of use in a subsidiary or specialised role in a system for discovering useful medical information. Such niches may exist in the examination of such domains as newsgroups or other discussion boards. The Kolmogorov distance is likely to be a useful additional tool to any other clustering approach however because of its dissimilarity and incorporation of otherwise unused data about a document. It should also be particularly useful because it is not dependent on language information.

8.1.2 Automatic fuzzy ontology

Results of the automatic support calculation from section 6.2.2 demonstrate that differences in the weightings of membership for each potential term location can be generated from a corpus of documents. Different results could be expected to be produced for different corpora, with differing authors and target audiences. One particularly pertinent aspect is the relatively small number of multiply located terms that can be resolved. For example in section 6.1.2, the number of ambiguous terms identified is roughly equal to the number of documents studied. However, this may not be an insurmountable issue if a collaborative system is put in place. In this scenario, pages downloaded via the system for individuals could also be analysed for the central database. One particular approach would be the construction of a specialised web - based search engine that could record the preferences of the users in the same way as fsearch, but via a standard browser.

The choice of updating algorithm for the fuzzy ontology is only one of a range of possible approaches. In fact the approaches demonstrated here include simple linear systems, such as equation 3, or the weighted by search number approach suggested in equation 5. It is designed to be simple, and to converge to a stable value in order to make processing practical. However extra resources could be devoted for particularly difficult areas, or a Genetic Algorithm approach could be used to traverse the solution space. It may be that an optimal settling or "forgetting" time may be discovered if the process is continued for longer. The results from the RCOG Corpus demonstrate that wide variations in membership are possible. The Reuters-21578 corpus results (section 6.5) demonstrate some support for the method of learning membership values by distance between terms in a document, although the f-measures were not as good as the case where humans selected the index terms. It is suggested that human- selected index terms are unlikely to be widely available on the WWW, and that, in addition the differing requirements of different users may make them less useful in practice.

8.1.3 Structural Analysis

The structural and stylistic analysis work was not as productive as originally hoped. The neural network learning system was effective, with classification rates in the 80% range in the domain classification task, but there did not appear to be a great deal of meaning to be gained from it, because of the low purity scores and the small number of classifications discovered. It is suggested that this would be part of a hybrid approach including other data such as the kol-distance calculation, with related documents specifically identified by the users. This needs a period of extended use of the system. It may also be that such an approach requires more complex feedback than the simple useful/useless visual analogue scale, although by excluding particular variables it may be possible . The scheme would be:

Develop a corpus of "highly rated" documents

Recover other documents related to these documents – e.g. by the same author or publisher or genre.

Present the new and old documents to users and note the degree of agreement in their usefulness rating.

It is interesting that other approaches to this sort of work for genre identification (Rauber & Mller-Kgler, 2001) have also shown quite disappointing results.

8.1.4 Case study results

The case study results as described in section 8.14 indicate that the fSearch system is regarded as both useful and usable. There remain a number of issues to resolve, but the concept of a group searching system appears comprehensible to users and this mechanism is a practical way of eliciting associations and usefulness. The relatively high rating of both perceived usability and usefulness will according to the technology acceptance model of (Davis, 1989), be likely to lead to adoption. The comments by the users, and the observations support this conclusion.

8.1.5 Extensibility

There are a number of facets to potential extensibility of this work. Firstly, the system can be extended to other medical domains, e.g. Paediatrics etc. There are two possible approaches to this – generalisation, where the whole MeSH domain is made available, or extension, where the same approach is taken as in the current system with a limited domain.

8.1.6 Fuzzy ontology development

Referring to the fuzzy ontology concept with a group and individual ontology, the concept of a general ontology could be used, which covers the whole of the MesH or any other ontology. The drawback of this approach is that specific information about useful structures or ontology membership values may be washed out by the combination. In addition, a standard fuzzy ontology may be biased by the contribution of particular domains. The fuzzy ontology membership values and other parameters could be collected from domain experts or learnt from specialised corpora – for example the BMJ topic corpus. However the combination is problematic -should each group parameter set be weighted equally, or should their contributions be weighted according to some scheme. Possible schemes could reflect:

- The number of queries associated with a particular domain from a particular portal – calculated from log files – this approach is often taken for search engine optimisation.
- The relative importance in terms of number of patients, or number of practitioners in the domain.
- The number of publications or citations in a particular area – an approach taken by Citeseer and other citation based systems.

The other issue concerns what constitutes a domain, as there is often combination and fission of specialities in the medical domain, and even controversy as to who is best equipped to look after particular patients, or teach particular students (McCahan, 1998).

It is hoped that further research in this area will continue, in particular in the following areas; the replacement of the executable form of the system with a browser based client server system that will allow much larger user groups to interact with it, and provide a substantial base for learning about group preferences. Mobile and wireless information retrieval may be more appropriately integrated into clinical workflow, especially by means of “information appliances” (K. F. Eustice et al., 1999). More research needs to be undertaken in the use and standardization of aspects of information reliability, usefulness and relevance to improve research and classification in this area, especially from the perspective of the clinical worker.

8.1.7 Extension

The extension approach involves accepting that different domains need to retain their own fuzzy ontologies and other parameters. In this approach, the system login remains the same apart from the addition of a domain selection along with speciality etc. A “General” category is also required, along with “Unknown”. The general category could have it’s own domain information, reflecting common queries and results, however the unknown domain would act as a higher level system which would be designed to lead users to the correct lower-level domain. This could be done via a system comparing the query terms used to terms contained in the various fuzzy ontologies. Unlike the other domains the unknown domain would generally not learn its own ontologies etc. but search others. It would appear that the extension approach would be simpler in the long run as such an approach would allow the construction of fuzzy ontologies by those best able to confirm their suitability for the task.

8.2 Structural fuzzy ontology

Fuzzy ontologies can have a wider use outside term definition and selection. In the present case the fuzzy ontology approach could be used to relate different sorts of documents in terms of their structure. The combination of fuzzy ontologies as described in section 3.5, could be used for the identification of documents likely to be of suitable structure. An example is given below:

Consider a user who is a member of a group, for example (English speaking, physiotherapist, advanced user). If we represent their document structure preferences as a set of fuzzy relations, then we can use the combination rules described above, if documents are assigned to a set of similar groups. In this example the document structure preferences of this example user may be as shown in Table 63.

Table 63: Simulated preferences

Parameter	Value	Membership
Official source	1	High (1)
Graphic Content	0.8	Low (0.2)
Number of Links	0.7	Medium (0.5)
Non-English	1	Unwanted (-1)

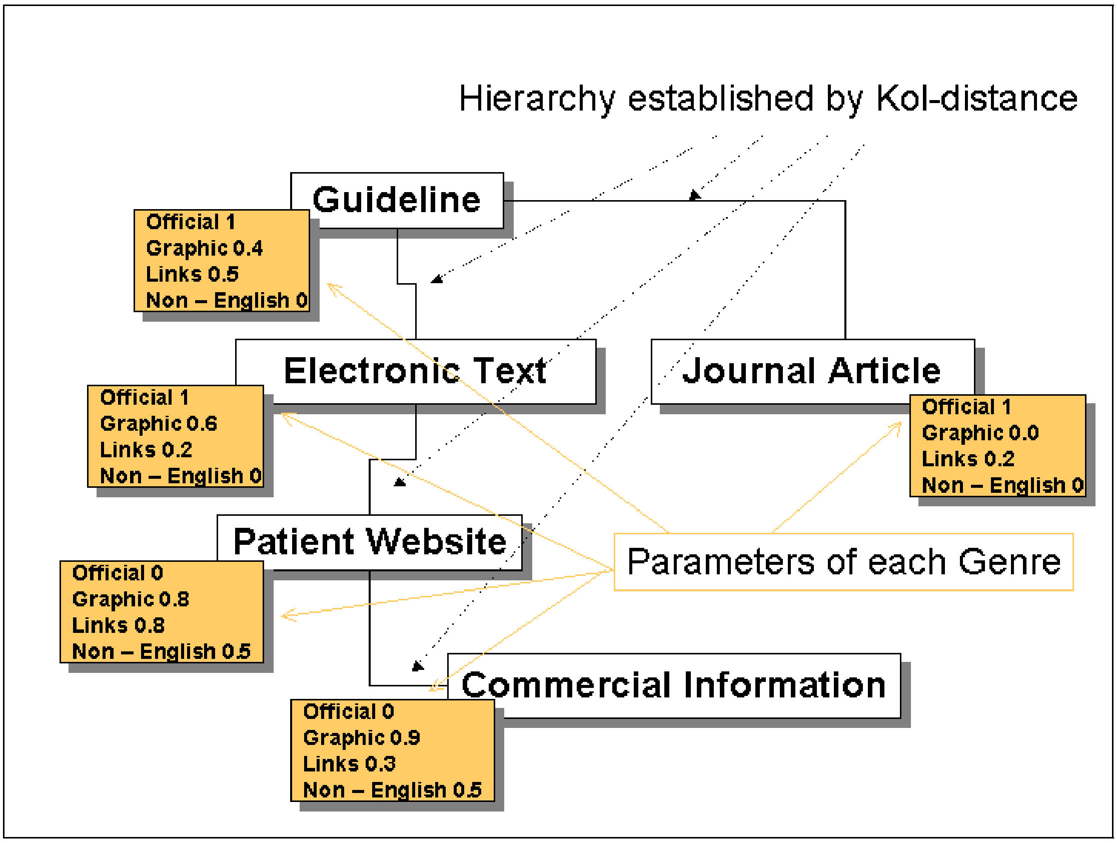


Figure 71: Structural fuzzy ontology

The parameters are derived from the combination of the parameters for "useful" information described in Chapter 2 and a combination of the DOM elements described in section 6.3.3. The documents retrieved by the system can also be classified by means of the Kolmogorov method in terms of their distance including the style elements i.e. the mark up tags as well as the text. In this case the Kol-distance is calculated between documents with a high "usefulness rating" as recorded by members of this group or individuals, and the documents being considered. This is then represented as an overall "quality factor" between 0 and 1. This means that a new document can then be located in a structural fuzzy ontology where it is assigned to a particular genre with a particular membership.

Figure 71 shows the construction of such an ontology, where an hierarchical clustering process using Kol-Distance produces the base relations between the genres. A sample of documents belonging to each genre is then identified and the mean value for each parameter calculated. A new document is then assigned to one or more different genres with different membership values for each. This approach does not of course need to use Kol-distance clustering, k-means or other parameter-based techniques could be used, or multiple techniques could be used in combination. This work may well be incorporated into more sophisticated learning schemes, where such information as Kolmagorov distance may be but one part of the overall analysis.

8.3 Some Research Questions

Groups of users have particular requirements for information in terms of content and presentation and structure.

This is supported by the literature, that is that groups have particular information needs (Glover et al., 1999; Quintana, 1998; Stephens & Huhns, 2001; R. W. White, Jose, & Ruthven, 2003). This conclusion was borne out by some of the comments in the case study, where users commented that patients and doctors may be using the same terms, but have different information needs.

Using current, diverse, knowledge sources e.g. PubMed and Google can be difficult, but essential. *And*

An integrated environment for information retrieval will help with learnability, reducing time spent on fruitless searches, increasing the range of sources used and allowing seamless updating of those sources,

That multiple sources are difficult to search is born out by some of the comments in the usability testing. The case study also highlights the learnability of a common system. The increase in web service usage, and the continuing development of the entrez eutilites for searching PubMed (National Library of Medicine, 2003) bear this out. The digital library community is also becoming more interested in the possibilities afforded by modular search systems (Fu & Mostafa, 2004).

Ontologies are useful for information retrieval but are more likely to be specialised among groups than universal or individual.

The use of ontologies and the differences between particular user's ontologies is confirmed by the variation in membership values discovered in the practical analysis of the various corpora. The use of ontologies for information retrieval has been extensively studied, for example in (Alani et al., 2003; Kyung-Sam Choi, Chi-Hoon Lee, & Phill-Kyu Rhee, 2000; Sirin et al., 2004; Widyantoro, 2001; Xiaolan Zhu, 1999). The use of the fuzzy ontology for information retrieval was tested in section 6.5.

Structural and similarity measures are useful for identifying useful documents.

Structural and in particular information theory based approaches can be used to differentiate between groups of documents, but the data preparation may need refinement. Chapter Four demonstrates obvious success in author identification, and hopeful results in genre and source detection. However, the success of this approach is not overwhelming.

Examining corpora of documents allows clustering and ontology creation/modification to take place

The use of corpora for ontology creation and modification has been demonstrated in the literature, for example in (Kruschwitz, 2003). In this thesis the work described in section 4.3.1 and section 4.3.2 has practically demonstrated the ease of creation of web-based corpora, and the work in Chapter Six has shown the practicality of modifying ontologies based on this data.

Usefulness and relatedness feedback can be obtained from people, and the effort of this process is reduced if this procedure happens in the concept of a particular document.

The usefulness and relatedness feedback was shown to be usable in the case study in Chapter Seven, and previous work has shown that the visual analogue scale is an appropriate way to do this (Lenert LA, 2001). The usability scores, and the observed ease of use of the fSearch system also demonstrate the effectiveness of this approach.

8.4 More General Fuzzy ontologies and OWL

The fuzzy ontology schema presented in this work is a single 'proprietary' example. However there are increasing numbers of ontology standards emerging, for example the enterprise ontology (Maedche, 2003). More generally, with the emergence of standards such as OWL (Smith et al., 2004), the fuzzy ontology membership value could be easily added to the specification. One approach would be to add it as part of the parameter for a particular class, but a more productive approach may be to include a membership value as a optional part of the standard schema for relations. In this approach, rather than the implicit is-a relation with a particular membership proposed above, each possible relation would have an optional parameter corresponding to the fuzzy membership. As with the current OWL representations, extra relations are easily added, and the definitions of classes and objects within the ontology are stored separately. In terms of the logic of the fuzzy ontology, if a relation is not present in the OWL code, then the membership value is set to zero. If a relation is present with no explicit value, then it can be assumed to be a crisp relation with value 1.

There are a number of advantages of using the OWL approach. Firstly a published standard exists, (W3C, 2004a) and one consequence of this is that OWL editors are becoming available. Approaches such as OWL are optimised for support of ontology use and traversal via machine, and browsers that use this approach will no doubt become available.

A more general ontology, one with multiple relations between the objects within it, may also use the fuzzy ontology approach. 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 reuse and knowledge management to be extended.

There are differences between the OWL and the fuzzy ontology approaches. As shown in section 2.6, OWL requires information about the type of relations that are being used, and represents each one as an entry in an XML-type document. Only pairs of objects that have relationships with each other need to be listed, if there is no relationship an entry does not exist. A fuzzy ontology will add to the relations between objects the membership value of that relation. However this leads to a number of issues:

- How should the membership value be normalised?
- What is the effect of the multiplicity of potential relations ?
- Is there any additional meaning to be found in such an approach?

8.4.1 Normalisation of membership values

The issue with normalisation of membership values arises because of the multiple potential relations inherent in a rich-relations ontology. In a simple “is-a” hierarchy, the normalisation process is fairly trivial, as each location of a term can be regarded as an instance of a potential desired location for that term. Assuming that the values are within the range 0-1 then the sum of each membership value at each location for each term should add up to 1. This has been described earlier, and the calculation performed, for example in section 6.4.5.

With multiple possible relations such an approach will need revision. Consider a term that has multiple relation types with other terms. If the same approach was used then the concept of a “slot” can be taken to mean one side of every relation that the term is involved in, so for example if there are two relations between two terms then the number of slots go from n to n+1. To give an example, consider the terms “apple” and “tree”. The apple can be seen as being “part-of” a tree (i.e. a physical component of the tree) and also “product-of” the tree (i.e. the economically valuable part of it). The normalisation would occur over every item over every relation. The advantage of this type of normalisation is that it can be done using the process described in section 6.4.4 or any other scheme that detects relatives in documents.

One potential drawback of this approach is an explosion in the number of relationships that need to be recorded in order to describe the ontology. In a naïve approach every term is potentially related to every other term by every relation, with some limitations where the number or arrangement of terms is limited – for example if there is a relation such as “closest-to”. However, generally, the majority of term/relation/membership value tuples will have a membership value of zero, because most relationships are not used by most terms, so they can be excluded. The non-zero relations can be used as the basis of a sparse 3 dimensional relationship matrix, with each position representing a membership value of two objects and the relationship. This assumes that each relationship allows two terms to be linked – some such as “needs all of” may require three or more terms. An example of this could be “Fire needs a combustible material and oxygen and a heat source to occur”. This approach may allow query refinement via feedback on the type of relationship that exists between the terms, along with a clearer view of the structure of knowledge in a domain.

8.5 Possible future approaches

This section describes a number of future approaches; in some cases work has been done already in this area. In other cases this is more speculative.

8.5.1 Radio frequency identification devices and tablet input devices.

Radio Frequency Identification Devices (RFID’s) have become increasingly popular and cheap within the last 5 years (Stanford, 2003). RFID tags and detectors can be purchased that adhere to particular standards, for example ISO 15693, using particular frequencies, for example 13.56 MHz. One important feature of the tags is that they are effectively powered by the interrogation pulse and thus require no batteries. This has enabled them to be made extremely small and light, down to the size of grains of rice, or even printed onto paper. They can be read at various ranges, depending on the size and orientation of the tag and the detector antenna. The low cost has made them attractive in supply –chain applications, as a replacement for barcodes as a line of sight is not necessary for reading (Raza, Bradshaw, & Hague, 1999). There have been concerns about potential loss of privacy associated with the use of the tags (Weiss, 2003), especially as they can be implanted under the skin! However, some of the most promising work appears to be associated with the use of RFID to link the virtual and physical worlds (Want, Fishkin, Gujar, & Harrison, 1999). By being able to uniquely identify objects in the physical environment, links between the virtual manipulation or navigation between objects, and physical interaction with them can be easily reinforced. There are three features of RFID that make them particularly attractive for use in personalisation and localisation:

I. Multiple RFID’s can be detected simultaneously by individual detectors .

II. RFID tags have unique ID numbers, unlike barcodes, and can also be provided with RAM.

III. RFID Tag detection does not require a line of sight link to the detector, or particular activation to begin the detection process.

Combinations of such devices in a number of ways could enhance the medical searching system:

- Users could wear "smart badges" that uniquely identify them to the system – as they approach the system it could log them on with less preamble.
- Similarly the system could identify classes of user, patient, doctor, nurse etc. automatically.
- In conjunction with mobile, wireless computers, the system could recognize its location, e.g. at the bedside, in a study area, at the nurses station, in a particular clinical area. This would enable the system to bias its search to particular keywords or concept modifiers (such as treatment versus diagnosis, paediatrics versus geriatrics etc.
- Ideally using tablet type interfaces, or even voice recognition, the system could allow searches to be carried out and then physically passed around a group of users, with each user seeing the appropriate result set for their group and preferences. The search process could then become an almost simultaneous use of collective intelligence, sharing insights and relations discovered by different individuals who are in different groups as well as the continuing process within groups.

To expand the final point, the system could be extended to include linkages between structural preferences and fuzzy ontologies, based around geographical location. This could allow the creation of a set of local recommendations in the ward, hospital, practice or other location. It is conceivable that for particularly complex patients, this would occur for one particular patient. In the context of the "smart patient" such a system could even act as a communication device between clinician and patient, with those articles of interest or perceived usefulness to each being communicated to the other, directly as happens today with the patient information sheet or the infamous pile of printouts from the Internet. More interestingly, such an approach might be used to separate structural preferences from particular search terms, so that documents that contain information that one user discovered in their favoured format could be matched with documents containing similar information in the other users preferred format.

8.5.2 Information appliances

Although personal, wireless enabled devices have been available for some time, they have had a mixed reception from users. In particular such technologies such as WAP have suffered from poor usability and long latencies for page downloads (Palen & Salzman, 2002). Their use for information retrieval in practice has been unsuccessful. In particular the long download times mean that searching or browsing activities – both of which inevitably involve the examination of documents that are not useful, as well as those that are – become unfeasible.

With the increasing availability of very large capacity personal information storage devices (such as the Ipod, and MP3 Players), along with wireless networking capacity, an information appliance (IA) approach (K. F. Eustice et al., 1999) may be the logical outcome of this work. IA's are personal, portable, specialised devices designed to provide information in context.

Recent work on the personalisation of shared ubiquitous devices (Hilbert & Trevor, 2004) has emphasised the concept of an "information cloud" around the user where the device is aware of its location, the user and even the task the user is likely to be currently performing. Such devices will almost certainly need to learn the behaviour of the user and the routine tasks performed in order to be usable. In addition, as the article by (K. F. Eustice et al., 1999) points out, the interface for such a device will be changed in context, although with modern devices allowing familiar GUI'S on even small screens, and with the increasing quality of emulators and multi-target development environments such a problem may be less of an issue

Such an appliance may be particularly useful in the hospital environment, where large numbers of junior doctors are not only constantly mobile, but also constantly in need of useful medical information. Because their group membership is already known, and there are large numbers of individuals involved, a fairly complete term and structure fuzzy ontology can be built quite quickly. With the storage capacity available in personal devices such as the IPOD (e.g. 40GB and above), the IA can be used as a cache, storing the documents most likely to be useful. The ontology updates can take place when the devices are connected to a high-speed network when the user is not working and the central network can search for updated documents and incorporate the change in rankings for documents selected by use of the ontology. In this way, the vast majority of the searching can be done via the cached documents contained in the IA, without very large movement of data via the wireless network. This should reduce the length of time spent downloading useful pages (which are likely to already be on the local device) and useless pages – which are unlikely to be downloaded at all. Because the intelligence of the system is largely contained within the fuzzy ontology, processing power for such IA's can be limited.

The wireless networking aspects are also showing considerable improvement at this time. Currently there are three major groups of wireless approaches.

- Short Range – typically within a room. These technologies include Bluetooth and infrared links. The advantages of these include free spectrum, low power requirements and low probability of interference between network.
- Medium Range – around a building or campus. Technologies such as Wi-Fi exemplify this market, with free spectrum usage, but higher requirements for base stations, and the need for security measures to prevent drive-by attacks.
- Long Range – Effectively covers the whole country. Mobile telephony technologies and satellite links are included in this group. These approaches require use of regulated and hence paid-for spectrum, often accessed via telecommunication service providers that provide coverage via large base stations in the case of mobile telephony, or powerful receivers in the case of satellite approaches. In the past, mobile telephony has suffered from low and intermittent bandwidth, but with the advent of so-called 3G services, using higher bandwidth and always-on connection have partly solved these problems. Security and infrastructure issues are also less onerous for the end-user health institutions, although cost may be a factor. In New Zealand, 3G services using WCDMA are being introduced during 2004.

Junior doctors move between speciality groups at regular intervals, and also may be employed in areas with particular emphasis on particular information sources. By use of the fuzzy ontology IA, they can be provided with information sources that are appropriate to their current role, by changing their group membership and modifying their personal fuzzy ontology accordingly. Such an approach can for example allow the relevant domain, and the relevant meaning of particular words, which may have different emphasis between specialities.

In terms of bandwidth and device size an N-tier approach may be used with tablet or projection systems being used in static or semi-static roles – for example associated with the chart or drug trolley or even particular beds. With relatively large displays, large amounts of local memory, multiple applications running simultaneously, and sophisticated access control, these devices would allow linking to multiple central databases, both clinical and knowledge based. The next layer could be PDA's either individually or collectively owned, with smaller displays, relatively slower interfaces and single visible applications. These devices may have limited access to confidential data because of difficulties in authentication and access control. The most personal level could be high bandwidth mobile phones. The user interface would be writing surface based at the higher level with a tendency towards voice interaction at the smaller end of the range. Much depends on the relative costs of bandwidth. For example, in a small unit the cost of mobile phone connections where the infrastructure is shared with other, external, subscribers may be more attractive than in an environment where large numbers of users are able to share an internal WI-FI network.

8.6 Overall conclusions

One of the most important issues that arises in data analysis is the difference between global and local solutions to problems. Approaches that use global solutions have been used for many years – for example - statistical modelling. However local approaches such as heuristics are often used by humans – in for example deciding behaviour in particular contexts. This work attempts to show methods of using local solutions, by linking information retrieval optimisation to particular groups' needs rather than the global approach that may be based on first principles or from analysis of all users. The major problem with such a local approach is deciding what is local, and how big 'local' is. By linking the analysis to known professional groups, particular genres or well-defined corpora this thesis has attempted to justify and map the appropriate areas for local analysis.

The usability and perceived usefulness of information retrieval systems are especially important in the area of medicine. If practicing clinicians are to use the vast array of resources then integrative systems such as fSearch will need to replace the myriad of incompatible partial search systems currently available. At the same time such systems will have to be able to deliver useful results, useful that is to the particular users or groups. Differentiation between the views of knowledge between different user groups and between users, and the learning and recording of that information is vital. It is hoped that the fuzzy ontology approach may be convincing as a basis for development in this area. Undoubtedly, learning and classification of documents and preferences, whether by neural networks, traditional statistics or new methods will be a vital part of future systems, and the use of such approaches in this thesis has shown some success. The use of the term "preferences" is a little weak – nothing is more useless than information that is inappropriate, and systems that fail to match themselves to their users needs are unlikely to continue to be used.

Analysis of web corpora is an expanding field, and much of the actual experimental time spent on this thesis was used in devising appropriate methods for doing this as well as validating the resultant data sets. The increasing availability of web services for information retrieval and the increasing interest in their use will make their use in future information retrieval work essential. The use of information theory for document characterisation has been shown in this thesis to be effective for the identification of authorship and structural aspects. The work in this area will no doubt increase.

The final words should perhaps go to Jean-François Abramatic, from a presentation in 1997(Abramatic, 1997):

> “It could become a World *Wild* Web.
> Let us all make it a World *Wise* Web.”

Appendices

8.7 List of terms and MeSH headings used

There were 735 Terms or phrases used for searching in Google and PubMed - see Table 64.

Table 64: List of terms used for searching along with their MeSH Codes

Description	Code
Perineum	A01.719
Pelvis	A01.673
Fetus	A16.378
Embryo	A16.254
Placenta	A16.759
Zygote	A16.950
Ovum	A16.631
Pelvic Floor	A01.673.600
Vernix Caseosa	A16.378.857
Aborted Fetus	A16.378.099
Fetal Blood	A16.378.200
Fetal Heart	A16.378.303
Meconium	A16.378.529
Blastocyst	A16.254.085
Amniotic Fluid	A16.254.072
Branchial Region	A16.254.160
Blastomeres	A16.254.090
Cloaca	A16.254.283
Cleavage Stage, Ovum	A16.254.270
Wolffian Duct	A16.254.940
Vitelline Duct	A16.254.891
Urachus	A16.254.835
Limb Bud	A16.254.462
Mesonephros	A16.254.500
Neural Crest	A16.254.600
Mullerian Ducts	A16.254.570
Umbilical Cord	A16.254.789
Organizers, Embryonic	A16.254.650
Notochord	A16.254.610
Germ Layers	A16.254.425
Gastrula	A16.254.412
Fetal Membranes	A16.254.403
Embryo, Nonmammalian	A16.254.300
Egg Yolk	A16.631.325
Zona Pellucida	A16.631.900
Vitelline Membrane	A16.631.886
Chorionic Villi	A16.759.189
Trophoblasts	A16.759.802
Decidua	A16.759.289

Description	Code
Cleavage Stage, Ovum	A16.254.300.600
Chick Embryo	A16.254.300.200
Truncus Arteriosus	A16.378.303.930
Ductus Arteriosus	A16.378.303.395
Morula	A16.254.270.550
Blastula	A16.254.270.274
Trophoblasts	A16.254.085.162
Blastoderm	A16.254.085.067
Chorionic Villi	A16.254.403.473.200
Morula	A16.254.300.600.550
Blastula	A16.254.300.600.274
Pericytes	A16.254.425.660.600
Somites	A16.254.425.660.750
Animal Diseases	C22
Digestive System Diseases	C06
Nervous System Diseases	C10
Hemic and Lymphatic Diseases	C15
Parasitic Diseases	C03
Neoplasms	C04
Musculoskeletal Diseases	C05
Respiratory Tract Diseases	C08
Stomatognathic Diseases	C07
Bacterial Infections and Mycoses	C01
Virus Diseases	C02
Otorhinolaryngologic Diseases	C09
Eye Diseases	C11
Urologic and Male Genital Diseases	C12
Female Genital Diseases and Pregnancy Complications	C13
Cardiovascular Diseases	C14
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Nutritional and Metabolic Diseases	C18
Endocrine Diseases	C19
Immunologic Diseases	C20
Disorders of Environmental Origin	C21
Pathological Conditions, Signs and Symptoms	C23
Mycoses	C01.703
Zoonoses	C01.908
Brain Abscess	C01.323
Bacterial Infections	C01.252
Infection	C01.539
Central Nervous System Infections	C01.395
Hepatitis, Viral, Animal	C02.407
Zoonoses	C02.968

Description	Code
Autonomic Nervous System Diseases	C10.177
Chronobiology Disorders	C10.281
Nervous System Malformations	C10.500
Trauma, Nervous System	C10.900
Nervous System Neoplasms	C10.551
Neurotoxicity Syndromes	C10.720
Ocular Hypertension	C11.525
Ocular Hypotension	C11.540
Vitreoretinopathy, Proliferative	C11.975
Eye Infections	C11.294
Lens Diseases	C11.510
Asthenopia	C11.093
Eye Hemorrhage	C11.290
Vision Disorders	C11.966
Eye Diseases, Hereditary	C11.270
Retinal Diseases	C11.768
Lacrimal Apparatus Diseases	C11.496
Corneal Diseases	C11.204
Orbital Diseases	C11.675
Refractive Errors	C11.744
Eyelid Diseases	C11.338
Ocular Motility Disorders	C11.590
Eye Abnormalities	C11.250
Eye Manifestations	C11.300
Uveal Diseases	C11.941
Optic Nerve Diseases	C11.640
Eye Neoplasms	C11.319
Scleral Diseases	C11.790
Pupil Disorders	C11.710
Conjunctival Diseases	C11.187
Vitreous Detachment	C11.980
Urologic Diseases	C12.777
Fournier Gangrene	C12.147
Tuberculosis, Urogenital	C12.672
Peste-des-Petits-Ruminants	C22.706
Parturient Paresis	C22.695
Parasitic Diseases, Animal	C22.674
Pleuropneumonia, Contagious	C22.717
Bird Diseases	C22.131
Swine Diseases	C22.905
Venereal Tumors, Veterinary	C22.950
Keratoconjunctivitis, Infectious	C22.500
Wasting Disease, Chronic	C22.955
Goat Diseases	C22.405

Description	Code
Genetic Diseases, Inborn	C16.320
Abnormalities	C16.131
Infant, Newborn, Diseases	C16.614
Skin Diseases	C17.800
Connective Tissue Diseases	C17.300
Nutrition Disorders	C18.654
Metabolic Diseases	C18.452
Neoplastic Endocrine-Like Syndromes	C19.576
Gonadal Disorders	C19.391
Tuberculosis, Endocrine	C19.927
Thyroid Diseases	C19.874
Breast Diseases	C19.146
Diabetes Mellitus	C19.246
Parathyroid Diseases	C19.642
Hyperinsulinism	C19.458
Polyendocrinopathies, Autoimmune	C19.750
Pituitary Diseases	C19.700
Adrenal Gland Diseases	C19.053
Endocrine Gland Neoplasms	C19.344
Thymus Hyperplasia	C19.813
Dwarfism	C19.297
Autoimmune Diseases	C20.111
Hypersensitivity	C20.543
Glomerulonephritis, Membranoproliferative	C20.425
Immunologic Deficiency Syndromes	C20.673
Blood Group Incompatibility	C20.188
Immunoproliferative Disorders	C20.683
Purpura, Thrombocytopenic	C20.841
Graft vs Host Disease	C20.452
Wounds and Injuries	C21.866
Environmental Illness	C21.223
Substance-Related Disorders	C21.739
Motion Sickness	C21.335
Poisoning	C21.613
Pregnancy Toxemias	C13.703.799
Urogenital Diseases	C13.371.820
Tuberculosis, Urogenital	C13.371.803
Vaginal Diseases	C13.371.894
Uterine Diseases	C13.371.852
Adnexal Diseases	C13.371.056
Vulvar Diseases	C13.371.944
Endometriosis	C13.371.163
Genital Neoplasms, Female	C13.371.270
Sex Disorders	C13.371.665

Description	Code
Gynatresia	C13.371.320
Herpes Genitalis	C13.371.350
Infertility	C13.371.365
Uterine Rupture	C13.703.420.904
Postpartum Hemorrhage	C13.703.420.725
Placenta Praevia	C13.703.420.714
Placenta Accreta	C13.703.420.643
Fetal Membranes, Premature Rupture	C13.703.420.339
Dystocia	C13.703.420.288
Abruptio Placentae	C13.703.420.078
Fetal Anoxia	C13.703.277.100
Fetal Alcohol Syndrome	C13.703.277.080
Erythroblastosis, Fetal	C13.703.277.060
Chorioamnionitis	C13.703.277.030
Meconium Aspiration	C13.703.277.785
Fetal Macrosomia	C13.703.277.570
Fetal Growth Retardation	C13.703.277.370
Fetal Distress	C13.703.277.200
Uterine Neoplasms	C13.371.270.875
Ovarian Neoplasms	C13.371.270.750
Fallopian Tube Neoplasms	C13.371.270.500
Vaginal Neoplasms	C13.371.270.937
Vulvar Neoplasms	C13.371.270.968
Fetal Resorption	C13.703.243.300
Parovarian Cyst	C13.371.056.739
Pelvic Inflammatory Disease	C13.371.056.750
Ovarian Diseases	C13.371.056.630
Fallopian Tube Diseases	C13.371.056.390
Placenta Praevia	C13.703.590.734
Placenta Accreta	C13.703.590.609
Chorioamnionitis	C13.703.590.268
Abruptio Placentae	C13.703.590.132
Placental Insufficiency	C13.703.590.800
Skin Diseases, Viral	C02.825
Fatigue Syndrome, Chronic	C02.330
Sexually Transmitted Diseases	C02.800
Central Nervous System Viral Diseases	C02.182
Eye Infections, Viral	C02.325
DNA Virus Infections	C02.256

Description	Code
Salmonella Infections, Animal	C22.812
Anal Gland Neoplasms	C22.073
Rinderpest	C22.780
Disease Models, Animal	C22.232
Anaplasmosis	C22.085
Erysipelothrix Infections	C22.331
Cat Diseases	C22.180
Pseudorabies	C22.742
Primate Diseases	C22.735
Horse Diseases	C22.488
Hepatitis, Animal	C22.467
Muscular Dystrophy, Animal	C22.595
Cattle Diseases	C22.196
Enterotoxemia	C22.313
Mammary Neoplasms	C22.520
Abortion, Veterinary	C22.021
Foot-and-Mouth Disease	C22.380
Pathological Conditions, Anatomical	C23.300
Pathologic Processes	C23.550
Signs and Symptoms	C23.888
Sexually Transmitted Diseases, Viral	C02.800.801
Herpes Gestationis	C13.703.320
Labor Complications	C13.703.420
Labor, Premature	C13.703.453
Puerperal Disorders	C13.703.844
Abortion, Spontaneous	C13.703.039
Diabetes, Gestational	C13.703.170
Chorea Gravidarum	C13.703.141
Pregnancy in Diabetics	C13.703.766
Pregnancy, Ectopic	C13.703.733
Pregnancy Complications, Neoplastic	C13.703.720
Pregnancy Complications, Infectious	C13.703.700
Pregnancy Complications, Hematologic	C13.703.667
Umbilical Arteries	A16.254.789.641
Umbilical Veins	A16.254.789.807
Amnion	A16.254.403.277
Allantois	A16.254.403.147
Yolk Sac	A16.254.403.981
Chorion	A16.254.403.473
Deciduoma	A16.759.289.500
Endoderm	A16.254.425.407
Ectoderm	A16.254.425.273
Mesoderm	A16.254.425.660

Description	**Code**
Genital Neoplasms, Female	C13.371.820.800.418
Endometrial Neoplasms	C13.371.270.875.750
Cervix Neoplasms	C13.371.270.875.170
Leukorrhea	C13.371.894.700.500
Polycystic Kidney Diseases	C13.371.820.700.800
Nephritis, Hereditary	C13.371.820.700.742
Multicystic Dysplastic Kidney	C13.371.820.700.558
Sex Differentiation Disorders	C13.371.820.700.842
WAGR Syndrome	C13.371.820.700.921
Epispadias	C13.371.820.700.374
Bladder Exstrophy	C13.371.820.700.132
Anovulation	C13.371.056.630.050
Ovarian Cysts	C13.371.056.630.580
Oophoritis	C13.371.056.630.450
Menopause, Premature	C13.371.056.630.250
Ovarian Neoplasms	C13.371.056.630.705
Ovarian Hyperstimulation Syndrome	C13.371.056.630.642
Ovarian Failure, Premature	C13.371.056.630.611
Metrorrhagia	C13.371.852.691.622
Menorrhagia	C13.371.852.691.449
Cervix Neoplasms	C13.371.852.150.310
Cervix Incompetence	C13.371.852.150.280
Cervix Erosion	C13.371.852.150.220
Cervix Dysplasia	C13.371.852.150.185
Cervicitis	C13.371.852.150.150
Uterine Inertia	C13.703.420.288.728
Mastitis	C13.703.844.506.677
Galactorrhea	C13.703.844.506.389
Chiari-Frommel Syndrome	C13.703.844.506.192
HIV Wasting Syndrome	C02.800.801.400.520
HIV Seropositivity	C02.800.801.400.500
HIV Enteropathy	C02.800.801.400.480
HIV-Associated Lipodystrophy Syndrome	C02.800.801.400.400
AIDS-Related Complex	C02.800.801.400.080
AIDS Dementia Complex	C02.800.801.400.070
AIDS Arteritis, Central Nervous System	C02.800.801.400.060
AIDS-Associated Nephropathy	C02.800.801.400.050
Acquired Immunodeficiency Syndrome	C02.800.801.400.040
Fallopian Tube Neoplasms	C13.371.056.390.390
Salpingitis	C13.371.056.390.890
Pre-Eclampsia	C13.703.799.314.619
HELLP Syndrome	C13.703.799.314.309
Vulvovaginitis	C13.371.944.902.737
Vulvovaginitis	C13.371.894.906.820

Description	**Code**
Hematometra	C13.371.852.495
Uterine Rupture	C13.371.852.904
Uterine Prolapse	C13.371.852.833
Hematocolpos	C13.371.894.300
Dyspareunia	C13.371.894.217
Candidiasis, Vulvovaginal	C13.371.894.190
Vaginitis	C13.371.894.906
Vaginal Neoplasms	C13.371.894.834
Vaginal Fistula	C13.371.894.763
Vaginal Discharge	C13.371.894.700
Condylomata Acuminata	C02.800.801.220
HIV Infections	C02.800.801.400
Herpes Genitalis	C02.800.801.350
Tuberculosis, Female Genital	C13.371.803.940
Urogenital Abnormalities	C13.371.820.700
Urogenital Neoplasms	C13.371.820.800
Infertility, Female	C13.371.365.700
Abortion, Incomplete	C13.703.039.093
Abortion, Habitual	C13.703.039.089
Embryo Loss	C13.703.039.711
Abortion, Veterinary	C13.703.039.422
Abortion, Threatened	C13.703.039.339
Abortion, Septic	C13.703.039.256
Abortion, Missed	C13.703.039.173
Dyspareunia	C13.371.665.313
Embolism, Amniotic Fluid	C13.703.634.404
Pregnancy, Abdominal	C13.703.733.536
Pregnancy, Tubal	C13.703.733.703
Abortion, Septic	C13.703.700.173
Puerperal Infection	C13.703.700.715
Pregnancy Complications, Parasitic	C13.703.700.680
Trophoblastic Neoplasms	C13.703.720.949
Depression, Postpartum	C13.703.844.253
Puerperal Infection	C13.703.844.757
Postpartum Hemorrhage	C13.703.844.700
Lactation Disorders	C13.703.844.506
Eclampsia	C13.703.799.314
Gestosis, EPH	C13.703.799.399
Hyperemesis Gravidarum	C13.703.799.562
Candidiasis, Vulvovaginal	C13.371.944.190
Pruritus Vulvae	C13.371.944.626
Kraurosis Vulvae	C13.371.944.338
Vulvitis	C13.371.944.902
Vulvar Neoplasms	C13.371.944.819

Description	Code
Dyspareunia	F03.800.250
Sexual Dysfunctions, Psychological	F03.800.800
Paraphilias	F03.800.800.600
Impotence	F03.800.800.400
Transsexualism	F03.800.800.800
Bipolar Disorder	F03.600.150.150
Seasonal Affective Disorder	F03.600.300.700
Dysthymic Disorder	F03.600.300.400
Depression, Postpartum	F03.600.300.350
Depression, Involutional	F03.600.300.300
Cyclothymic Disorder	F03.600.150.150.300
Sadism	F03.800.800.600.700
Voyeurism	F03.800.800.600.900
Transvestism	F03.800.800.600.800
Pedophilia	F03.800.800.600.600
Masochism	F03.800.800.600.500
Fetishism (Psychiatric)	F03.800.800.600.350
Exhibitionism	F03.800.800.600.300
Reproduction	G08.520
Maternal Age	G08.520.420
Insemination	G08.520.392
Fertilization	G08.520.277
Oviposition	G08.520.480
Pregnancy, Unwanted	G08.520.850
Pregnancy Trimesters	G08.520.840
Sex Reversal, Gonadal	G08.520.920
Sex Characteristics	G08.520.905
Sex	G08.520.900
Puerperium	G08.520.882
Puberty	G08.520.876
Pseudopregnancy	G08.520.870
Prenatal Exposure Delayed Effects	G08.520.860
Pregnancy, Multiple	G08.520.800
Pregnancy, Animal	G08.520.780
Pregnancy	G08.520.769
Placentation	G08.520.720
Paternal Age	G08.520.700
Parthenogenesis	G08.520.689
Parity	G08.520.615
Gravidity	G08.520.351
Gametogenesis	G08.520.310

Description	Code
Thecoma	C13.371.820.800.418.685.765
Meigs' Syndrome	C13.371.820.800.418.685.531
Luteoma	C13.371.820.800.418.685.464
Granulosa Cell Tumor	C13.371.820.800.418.685.398
Carcinoma, Renal Cell	C13.371.820.800.820.535.160
Nephroblastoma	C13.371.820.800.820.535.585
Nephroma, Mesoblastic	C13.371.820.800.820.535.790
Sertoli-Leydig Cell Tumor	C13.371.056.630.705.132.500
Turner Syndrome	C13.371.820.700.842.309.872
Gonadal Dysgenesis, Mixed	C13.371.820.700.842.309.391
Gonadal Dysgenesis, 46,XY	C13.371.820.700.842.309.388
Gonadal Dysgenesis, 46,XX	C13.371.820.700.842.309.193
Sarcoma, Endometrial Stromal	C13.371.270.875.750.374.500
Sertoli-Leydig Cell Tumor	C13.371.820.800.418.685.132.500
WAGR Syndrome	C13.371.820.800.820.535.585.950
Denys-Drash Syndrome	C13.371.820.800.820.535.585.220
Endometrial Stromal Tumors	C13.371.820.800.418.875.200.374
Carcinoma, Endometrioid	C13.371.820.800.418.875.200.124
Denys-Drash Syndrome	C13.371.820.700.842.316.627.220
Sarcoma, Endometrial Stromal	C13.371.820.800.418.875.200.374.500
Obstetric Surgical Procedures	E04.520
Abortion, Induced	E04.520.050
Colposcopy	E04.520.150
Cerclage, Cervical	E04.520.100
Culdoscopy	E04.520.160
Colpotomy	E04.520.155
Delivery, Obstetric	E04.520.252
Hysteroscopy	E04.520.360
Hysterotomy	E04.520.365
Fetoscopy	E04.520.280
Episiotomy	E04.520.252.750
Cesarean Section	E04.520.252.500
Version, Fetal	E04.520.252.996
Vaginal Birth after Cesarean	E04.520.252.992
Natural Childbirth	E04.520.252.984
Labor, Induced	E04.520.252.968
Home Childbirth	E04.520.252.937
Extraction, Obstetrical	E04.520.252.875
Pregnancy Reduction, Multifetal	E04.520.050.600
Abortion, Therapeutic	E04.520.050.060
Abortion, Legal	E04.520.050.055

Description	**Code**
Pregnancy Trimester, First	G08.520.840.408
Pregnancy Trimester, Third	G08.520.840.520
Quadruplets	G08.520.800.508
Superfetation	G08.520.800.608
Quintuplets	G08.520.800.550
Twins	G08.520.800.708
Triplets	G08.520.800.680
Insemination, Artificial	G08.520.392.492
Oogenesis	G08.520.310.500
Spermatogenesis	G08.520.310.760
Ovulation	G08.520.440.508
Menstruation	G08.520.440.428
Luteal Phase	G08.520.440.410
Follicular Phase	G08.520.440.310
Anestrus	G08.520.188.249
Proestrus	G08.520.188.875
Metestrus	G08.520.188.750
Estrus	G08.520.188.500
Diestrus	G08.520.188.374
Menopause	G08.520.150.500
Premenopause	G08.520.150.600
Maternal Age 35 and over	G08.520.420.630
Pregnancy in Adolescence	G08.520.420.770
Sperm Capacitation	G08.520.277.760
Ovum Transport	G08.520.277.360
Sperm Transport	G08.520.277.820
Sperm-Ovum Interactions	G08.520.277.800
Sperm Motility	G08.520.277.780
Fetal Organ Maturity	G08.520.170.290
Fetal Movement	G08.520.170.210
Twinning	G08.520.170.760
Sex Differentiation	G08.520.170.520
Organogenesis	G08.520.170.450
Gestational Age	G08.520.170.380
Fetal Weight	G08.520.170.325
Fetal Viability	G08.520.170.320
Embryo Implantation, Delayed	G08.520.175.100
Preimplantation Phase	G08.520.175.690
Postimplantation Phase	G08.520.175.670
Pregnancy, Prolonged	G08.520.769.689

Description	Code
Litter Size	G08.520.780.300
Coitus	G08.520.900.100
Penile Erection	G08.520.900.460
Orgasm	G08.520.900.400
Ejaculation	G08.520.900.138
Copulation	G08.520.900.110
Twins, Monozygotic	G08.520.800.708.838
Twins, Dizygotic	G08.520.800.708.800
Vitellogenesis	G08.520.310.500.880
Anovulation	G08.520.440.508.080
Superovulation	G08.520.440.508.768
Ovulation Inhibition	G08.520.440.508.493
Luteolysis	G08.520.440.508.380
Luteinization	G08.520.440.508.355
Estrus Synchronization	G08.520.188.500.500
Postmenopause	G08.520.150.500.625
Menopause, Premature	G08.520.150.500.500
Follicular Atresia	G08.520.440.310.358
Uterine Contraction	G08.520.769.326.700
Cervical Ripening	G08.520.769.326.100
Trial of Labor	G08.520.769.326.280
Labor Onset	G08.520.769.326.200
Breech Presentation	G08.520.769.362.150
Sperm Maturation	G08.520.310.760.700
Acrosome Reaction	G08.520.277.800.100
Milk Ejection	G08.520.882.608.460
Labor Stage, Third	G08.520.769.326.200.110
Labor Stage, Second	G08.520.769.326.200.090
Labor Stage, First	G08.520.769.326.200.080
Multiple Birth Offspring	M01.438
Twins	M01.438.873
Quadruplets	M01.438.486
Triplets	M01.438.768
Quintuplets	M01.438.587
Twins, Dizygotic	M01.438.873.920
Twins, Monozygotic	M01.438.873.940
Fetal Diseases	C16.300
Cretinism	C16.180
Heartwater Disease	C22.434
Sheep Diseases	C22.836

Description	**Code**
Parturition	G08.520.769.490
Maternal-Fetal Exchange	G08.520.769.455
Labor, Obstetric	G08.520.769.326
Oligohydramnios	C13.703.560
Fetal Diseases	C13.703.277
Fetal Death	C13.703.243
Corpus Luteum Maintenance	G08.520.769.100
Cesarean Section, Repeat	E04.520.252.500.150
Mood Disorders	F03.600
Sexual and Gender Disorders	F03.800
Gestational Trophoblastic Neoplasms	C13.703.720.949.416
Endometrial Neoplasms	C13.371.852.762.200
Cervix Neoplasms	C13.371.852.762.100
Pregnancy Rate	G08.520.769.775
Pregnancy Outcome	G08.520.769.675
Pregnancy Maintenance	G08.520.769.640
Pregnancy, High-Risk	G08.520.769.525
Prenatal Nutrition	G08.520.769.507
Labor Presentation	G08.520.769.362
Abortion, Eugenic	E04.520.050.050
Cervical Ripening	E04.520.252.968.100
Pregnancy Reduction, Multifetal	E04.520.050.060.600
Vacuum Extraction, Obstetrical	E04.520.252.875.970
Pregnancy Complications, Cardiovascular	C13.703.634
Polyhydramnios	C13.703.610
Placenta Diseases	C13.703.590
Phenylketonuria, Maternal	C13.703.575
Urogenital Diseases	C12.740
Genital Diseases, Male	C12.294
Genital Diseases, Female	C13.371
Pregnancy Complications	C13.703
Cardiovascular Abnormalities	C14.240
Heart Diseases	C14.280
Vascular Diseases	C14.907
Tuberculosis, Cardiovascular	C14.826
Hyperemia	C14.371
Pregnancy Complications, Cardiovascular	C14.583
Syphilis, Cardiovascular	C14.728
Scimitar Syndrome	C14.700
Lymphatic Diseases	C15.604
Hematologic Diseases	C15.378
Aleutian Mink Disease	C22.062
Rodent Diseases	C22.795
Paratuberculosis	C22.688
Fish Diseases	C22.362

Description	Code
Urologic Neoplasms	C13.371.820.800.820
Salpingitis	C13.371.056.750.875
Parametritis	C13.371.056.750.750
Uterine Inversion	C13.371.852.726
Uterine Neoplasms	C13.371.852.762
Uterine Hemorrhage	C13.371.852.691
Oophoritis	C13.371.056.750.500
Embryo Implantation	G08.520.175
Estrous Cycle	G08.520.188
Lactation	G08.520.882.608
Vaginosis, Bacterial	C13.371.894.906.800
Trichomonas Vaginitis	C13.371.894.906.633
Uterine Perforation	C13.371.852.904.500
Depressive Disorder	F03.600.300
Affective Disorders, Psychotic	F03.600.150
Menarche	G08.520.876.410
Pregnancy Trimester, Second	G08.520.840.490
Vesicovaginal Fistula	C13.371.894.763.816
Rectovaginal Fistula	C13.371.894.763.558
Menstrual Cycle	G08.520.440
Climacteric	G08.520.150
Fertility	G08.520.227
Embryo and Fetal Development	G08.520.170
Placenta, Retained	C13.703.590.767
Endometritis	C13.371.852.299
Endometrial Hyperplasia	C13.371.852.228
Cervix Diseases	C13.371.852.150
DNA Damage	C21.111
Occupational Diseases	C21.447
Foot Rot	C22.394
Actinobacillosis	C22.039
Zoonoses	C22.969
Lameness, Animal	C22.510
Dog Diseases	C22.268
Borna Disease	C22.152
Myxomatosis, Infectious	C22.627
Steatitis	C22.880

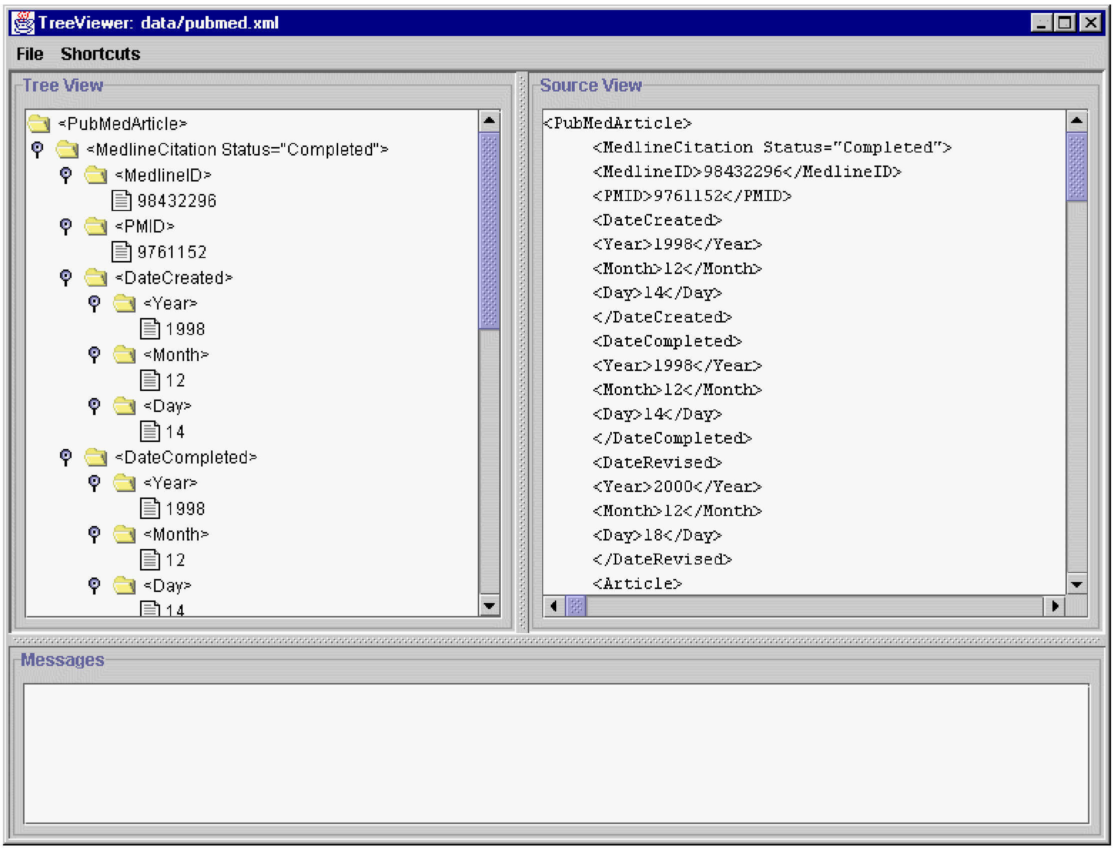
8.8 PubMed XML structure

1: Rose EA, et al. Is a 2-day course of oral dex...[PMID:11711021]

```
<PubMedArticle>
<MEDLINECitation>
<MEDLINEID>21568205</MEDLINEID>
<PMID>11711021</PMID>
<DateCreated>
<Year>2001</Year>
<Month>11</Month>
<Day>16</Day>
</DateCreated>
<Article>
<Journal>
<ISSN>0094-3509</ISSN>
<JournalIssue>
<Volume>50</Volume>
<Issue>11</Issue>
<PubDate>
<Year>2001</Year>
<Month>Nov</Month>
</PubDate>
</JournalIssue>
</Journal>
<ArticleTitle>Is a 2-day course of oral dexamethasone more effective than 5 days of oral prednisone in improving symptoms and preventing relapse in children with acute asthma?</ArticleTitle>
<Pagination>
<MEDLINEPgn>993</MEDLINEPgn>
</Pagination>
<Affiliation>Wayne State University, Department of Family Medicine, Detroit, MI. erose@med.wayne.edu</Affiliation>
<AuthorList>
<Author>
<LastName>Rose</LastName>
<FirstName>E A</FirstName>
<Initials>EA</Initials>
</Author>
<Author>
<LastName>Schwartz</LastName>
<FirstName>K</FirstName>
<Initials>K</Initials>
</Author>
</AuthorList>
<Language>eng</Language>
<PublicationTypeList>
<PublicationType>Journal Article</PublicationType>
</PublicationTypeList>
</Article>
<MEDLINEJournalInfo>
<Country>United States</Country>
<MEDLINETA>J Fam Pract</MEDLINETA>
```

```
<MEDLINECode>I4L</MEDLINECode>
<NlmUniqueID>7502590</NlmUniqueID>
</MEDLINEJournalInfo>
<CitationSubset>AIM</CitationSubset>
<CitationSubset>IM</CitationSubset>
</MEDLINECitation>
<PubMedData>
    <History>
        <PubMedPubDate PubStatus="PubMed">
            <Year>2001</Year>
            <Month>11</Month>
            <Day>17</Day>
            <Hour>10</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="MEDLINE">
            <Year>2001</Year>
            <Month>11</Month>
            <Day>17</Day>
            <Hour>10</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
    </History>
    <PublicationStatus>ppublish</PublicationStatus>
    <ArticleIdList>
        <ArticleId IdType="PubMed">11711021</ArticleId>
        <ArticleId IdType="pii">jfp_1101_0979d</ArticleId>
    </ArticleIdList>
</PubMedData>
</PubMedArticle>
```

8.9 PubMed entry in graphical format



```
<PubMedArticle>
    <MedlineCitation Status="Completed">
        <MedlineID>
            98432296
        <PMID>
            9761152
        <DateCreated>
            <Year>
                1998
            <Month>
                12
            <Day>
                14
        <DateCompleted>
            <Year>
                1998
            <Month>
                12
            <Day>
                14
```

```
<PubMedArticle>
     <MedlineCitation Status="Completed">
     <MedlineID>98432296</MedlineID>
     <PMID>9761152</PMID>
     <DateCreated>
     <Year>1998</Year>
     <Month>12</Month>
     <Day>14</Day>
     </DateCreated>
     <DateCompleted>
     <Year>1998</Year>
     <Month>12</Month>
     <Day>14</Day>
     </DateCompleted>
     <DateRevised>
     <Year>2000</Year>
     <Month>12</Month>
     <Day>18</Day>
     </DateRevised>
     <Article>
```

Figure 72: A single PubMed record in graphical format

Figure 72 shows a single-record PubMed search result displayed in a graphical representation via TreeViewer software (Apache Software Foundation, 2000). This Java-based software is particularly useful for understanding the structure of the document being retrieved.

8.10 List of topics from the BMJ

Items in bold were included in further analysis.

Table 65: The BMJ Corpus

Number of Pages	Topic Name
14	**injury**
14	**poisoning**
14	**resuscitation**
10	**public_health:other**
10	**Physiotherapy**
10	**other_emergency_medicine**
10	**microbiology**
10	**Dentistry**
9	**arrhythmias**
9	**breastfeeding_and_infant_nutrition**
9	**mood_disorders**
9	**medical_managers**
9	**health_of_indigenous_peoples**
9	**helicobacter_pylori**
9	**incontinence**
9	**other_sexual_medicine**
9	**occupational_health**
9	**poisoning**
9	**poisoning?notjournal=bmj**
9	**prison_medicine**
9	**otherhealthpolicy**
9	**qualitative_research_descriptions**
9	**read**
9	**sexual_and_gender_disorders**
9	**teaching**
9	**travel_medicine**
8	**undergrad**
8	**valve_diseases**
8	**schizophrenia**
8	**sleep_apnoea**
8	**smoking**
8	**prevention_and_health_promotion**
8	**postgrad:academic**
8	**obgyn:other**
8	**needs_assessment**
8	**neurological_injury**
8	**history**
8	**endocrine_system**
8	**epilepsy_and_seizures**
8	**gastroenterology:other**
8	**medical_careers:other**

8	**molecular_medicine**
8	**International_health:nonclinical**
8	**cancer_lung**
8	**AIDS**
8	**alcohol**
8	**drugs:infections**
8	**drugs:obstetrics_and_gynaecology**
8	**congenital_heart_disease**
7	**conflict**
7	**child_and_adolescent_psychiatry**
7	**clinical_research**
7	**drugs:psychiatry**
7	**adolescents**
7	**barker_hypothesis**
7	**Basic_sciences:nonclinical**
7	**cardiovascular_medicine:other**
7	**cervical_screening**
7	**cancer:breast**
7	**bayesian_statistics_descriptions**
7	**interventional_radiology**
7	**infectious_diseases:other**
7	**inflammatory_bowel_disease**
7	**information_in_practice**
7	**irritable_bowel_syndrome**
7	**medical_careers:continuous_professional_developmen**
7	**medical_careers:numbers**
7	**gastrointestinal_system**
7	**erectile_dysfunction**
7	**hypertension**
7	**neonates**
7	**oncology:other**
7	**other_rehabilitation_medicine**
7	**other_medical_informatics:other**
7	**osteoporosis**
7	**other_imaging_techniques**
7	**Osteoarthritis**
7	**psychiatry:other**
7	**palliative_medicine**
7	**sexually_transmitted_infections**
7	**somatoform_disorders**
7	**rheumatoid_arthritis**
7	**rheumatology:other**
7	**research_and_publication_ethics**
7	**research**
7	**radiological_diagnosis**
7	**vascular_surgery**
7	**Tuberculosis**
7	**systematic_reviews:statistics_descriptions**
6	**surgery:other**
6	**venous_thromboembolism**
6	**WWW**
6	**quality-improvement**
6	**psychogeriatrics**

6	**psychology**
6	**socioeconomic_determinants_of_health**
6	**Pancreas_and_biliary_tract**
6	**paediatric**
6	**patient_caregiver_communication**
6	**patients_other**
6	**postgrad:GP**
6	**pregnancy**
6	**orthopaedic_and_trauma_surgery**
6	**Organisation_of_health_care:nonclinical**
6	**other_geriatric_medicine**
6	**other_immunology**
6	**other_management**
6	**Other_respiratory_infections**
6	**obesity**
6	**neonatal**
6	**human_rights**
6	**impulse_control_disorders**
6	**guidelines**
6	**genetics**
6	**exams**
6	**gastrointestinal_surgery**
6	**environmental**
6	**medicine_in_developing_countries**
6	**liver**
6	**long_term_care**
6	**infants**
6	**informed_consent**
6	**chemical_pathology**
6	**anxiety**
6	**allergy**
6	**10_minute_consultations**
6	**adult**
6	**drugs_cardiovascular_system**
6	**drugs:immunological_products_and_vaccines**
6	**disability**
6	**doctor-doctor_communication**
6	**chronic_obstructive_airways**
6	**cytopathology**
5	**Culture**
5	**complementary_medicine**
5	**doctors_morale_and_well_being**
5	**drug_misuse**
5	**adverse_drug_reactions**
5	**anaesthesia:other**
5	**Cardiomyopathy**
5	**cancer:other**
5	**Cancer:prostate**
5	**cancer_gynaecological**
5	**ischaemic_heart_disease**
5	**medical_education:other**
5	**migraine**
5	**menopause**

5	**epidemiology:screening**
5	**eating_disorders**
5	**end_of_life_decisions**
5	**falling_sperm_counts**
5	**governments_non_uk**
5	**health_serv_reasearch**
5	**multiple_sclerosis**
5	**musculoskeletal_syndromes**
5	**neuromuscular_disease**
5	**nursing**
5	**other_sports_medicine**
5	**organ_donations**
5	**postgrad:residency**
5	**plastic_and_reconstructive_surgery**
5	**personality_disorders**
5	**patients_views**
5	**peer_review**
5	**patient_caregiver_relationships**
5	**parkinsons_disease**
5	**otolaryngology**
5	**paediatrics:other**
5	**sociology**
5	**statistics:other_descriptions**
5	**small_intestine**
5	**resource_allocation**
5	**respiratory_medicine:other**
5	**radiotherapy**
5	**reproductive_medicine**
5	**religion**
5	**randomised_controlled_trials_examples**
5	**stomach_and_duodenum**
5	**transplantation**
4	systematic_reviews:statistics_examples
4	uk_government
4	randomised_controlled_trials_descriptions
4	regulation
4	renal_medince
4	others_nutrition_and_metabolism
4	pathology:other
4	patients_cultures
4	pharmacology_and_toxicology
4	other_ommunication
4	neuropathology
4	ophthalmology
4	neurology:other
4	getting_and_changing_jobs
4	histopathology
4	family_planning
4	epidemiology:other
4	mens_health
4	medical_error_patient_safety
4	journalology:other
4	Lung_function

4	cancer:gastroenterological
4	cardiothoracic_surgery
4	chemotherapy
4	asthma
4	diabetes
4	drugs:central_nervous_system
4	drugs:respiratory_system
4	competing_interests
4	children
4	Connective_tissue_disease
4	dementia
3	Cystic_fibrosis
3	confidentiality
3	dermatology
3	authorship
3	abuse_child_partner_elder
3	adjustment_disorders
3	changing_physician_behaviour
3	liesbian_bisexual_gay_transgendered_health
3	medicine_and_the_law
3	motility_and_visceral_sensation
3	endocrinology:other
3	ethics:other
3	heart_failure
3	general_surgery
3	professional_conduct
3	pain
3	qualitative_research_examples
3	urology
3	stroke
3	surgical_oncology
2	neurosurgery
2	mad_cow
2	drugs:musculoskeletal_and_joint_diseases
2	chronic_diseases
1	delirium_amnestic_cognitive_disorders
1	Diffuse_parenchymal_lung_disease
1	bayesian_statistics_examples
1	Motor_neurone_disease
1	haematology
1	poisoning?page=3
1414	**Total**

8.10.1 Information sheet

AUT
AUCKLAND UNIVERSITY OF TECHNOLOGY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Information Sheet for usability testing of the Medical Searching system

Introduction

The purpose of this study is to evaluate the usefulness of a computer-based system for searching for medical information. Usability testing is a way of finding out how easy it is for users to use a computer system. A “test driver” or test user is asked to use the system and the researchers observe what happens. It is the software that is being tested and not the participant or his or her ability. The goal is to make these systems easier to use, and to show if they make the process easier. The Medical Searching system is based around the idea that if people can “pool “ the results of their searches then these searches will become better.

Invitation

You are invited to participate in this project. This should take no more than one hour of your time. You may withdraw from the study at any time before data collection is completed and up to the beginning of data analysis, and any questions are welcomed. Your input will greatly help the research process.

Procedure

Typically, you will be first asked to fill out a questionnaire prior to usability testing. You will then be given tasks to accomplish usually on a desktop PC or laptop system. Sometimes the software may be difficult to use, or even break down. However, you as the test user are helping to discover these problems and are under no pressure to complete these tasks. You may decide to stop and/or withdraw from testing at any time. There may be an observer taking notes and the computer itself may record what happens. You may be asked questions during the process and possibly again after the session.

Results

Your impressions of the system and use of it will be recorded via computer log and questionnaire. The effectiveness of your search using the system will be compared to a similar search using conventional tools. This research is part of David Parry’s PhD work at AUT, and anonymous results will be presented in the thesis. The researchers also hope to present this work in journal and/or conference papers. Material collected during the testing will only be used for the purpose of the study. This material will be securely stored. The material will be destroyed when no longer of use for its original purpose.

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor Professor Philip Sallis. Concerns regarding conduct should be notified to the Executive Secretary, AUTEC, Madeline Banda, madeline.banda@aut.ac.nz ,917 9999 ext 8044.

Private Bag 92006 Auckland 1020 New Zealand Facsimile 64-9-917-9944

TELEPHONE 64-9-917-9999 EXTENSION 8918 EMAIL DAVE.PARRY@AUT.AC.NZ

8.10.2 Consent form

AUT
AUCKLAND UNIVERSITY OF TECHNOLOGY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Consent to Participation in Research

This form is to be completed in conjunction with, and after reference to, the AUTEC Guidelines Version 1.4 (Revised July 1998).

Title of project: Medical Searching System Validation

Project Supervisors: Professor Philip Sallis

Researcher :David Parry

- I have read and understood the information provided about this research project as outlined in the information sheet .
- I have had an opportunity to ask questions and to have them answered.
- I understand that I may withdraw myself or any information that I have provided for this project at any time prior to completion of data collection, without being disadvantaged in any way. If I withdraw, I understand that all relevant tapes and transcripts, or parts thereof, will be destroyed.
- I grant permission for any information collected to be used for purposes as outlined in the information sheet .
- I agree to take part in this research.

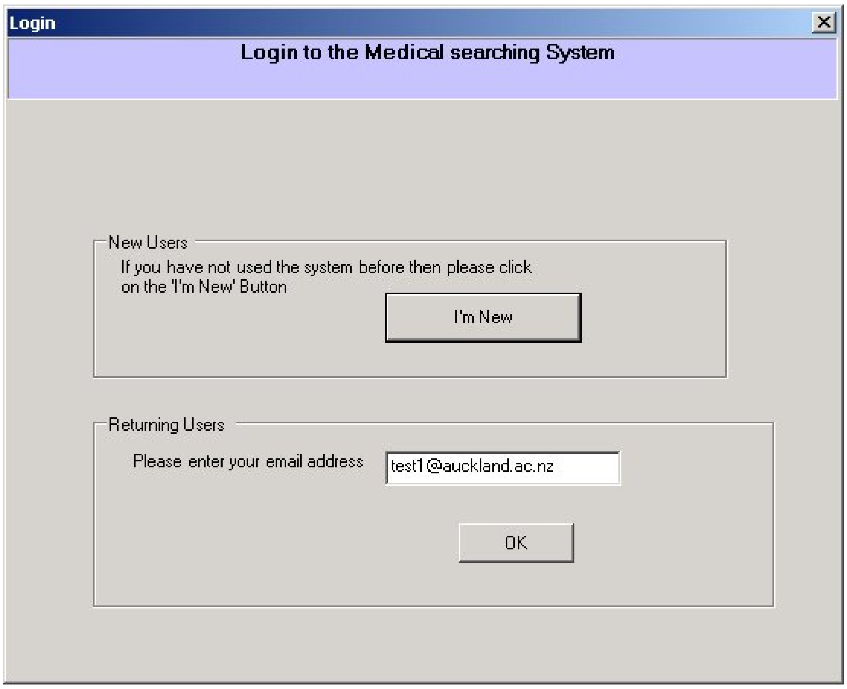
Participant signature: ....

Participant name: ....

Date: ....

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor. Concerns regarding conduct should be notified to the Executive Secretary, AUTEC, Madeline Banda, madeline.banda@aut.ac.nz ,917 9999 ext 8044.

8.11 Screens



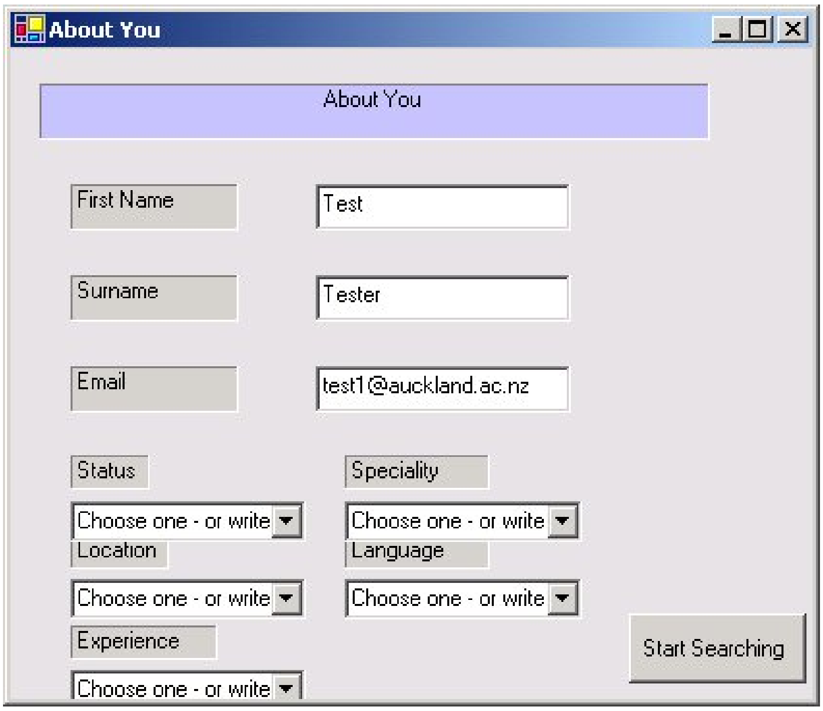
Login to the Medical searching System

If you have not used the system before then please click on the 'I'm New' Button

Please enter your email address test1@auckland.ac.nz

Figure 73: The Login screen

The login screen (Figure 73) uses the users email address as the unique identifier. There is no password. The user information screen is then displayed, to confirm details if the user is returning or to enter details if they are new. The Radio Frequency Identification (RFID) option (described in section 8.5.1) bypasses this screen.



About You

About You

First Name Test

Surname Tester

Email test1@auckland.ac.nz

Status Choose one - or write

Speciality Choose one - or write

Location Choose one - or write

Language Choose one - or write

Experience Choose one - or write

Figure 74: The User Details screen

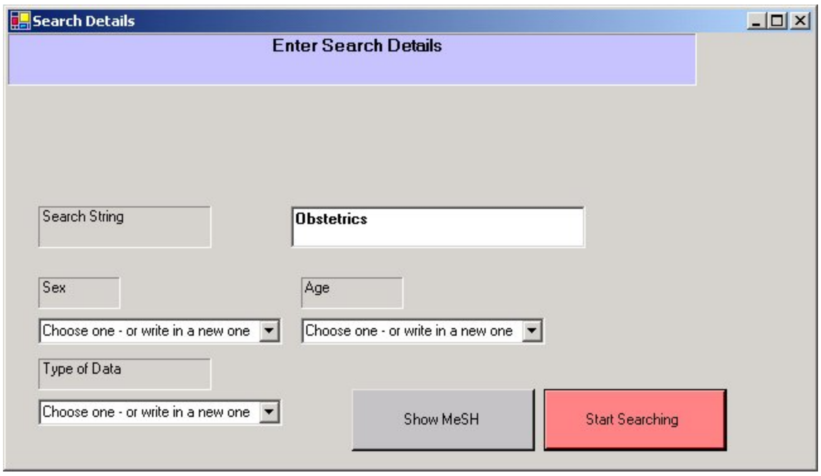
The user details screen (Figure 74) allows the user to self identify the professional group they belong to, and their status within that group – for example a Registrar (status) in obstetrics and gynaecology (speciality). They are also asked about their preferred language, their location – which country they are in - and their experience of using computer based systems for searching, classified using Dreyfus's classification:

1. *Novice. Operates by consciously learnt context-free rules. Lacks any sense of the overall task.*
2. *Advanced beginner. Uses more sophisticated rules, which refer to situational elements as well as context-free ones. These situational elements are features such as the pattern of behaviour, which distinguishes a drunken from a sober driver. They are learnt by experience, and the advanced beginner can't formalise them.*
3. *Competent. Has now learnt to recognise many context-free and situational elements. Still lacks any sense of their overall importance to the task, and rapidly becomes overwhelmed. Tries to overcome this by hierarchical goal-based planning. This hierarchical decomposition of the task means that, at any time, the competent pays attention only to that small number of features relevant to a particular sub goal, thus avoiding being overwhelmed.*
4. *Proficient. Most of the time, now performs his task intuitively, without analytical thought. But this deep involvement in the task will be broken when*

certain elements present themselves as particularly important. The proficient then stops and thinks analytically about what to do next.

5. *Expert. Performs his task intuitively, almost all the time. Occasionally has to stop and deliberate, but this involves critical reflection on his intuitions, rather than goal-based planning.*

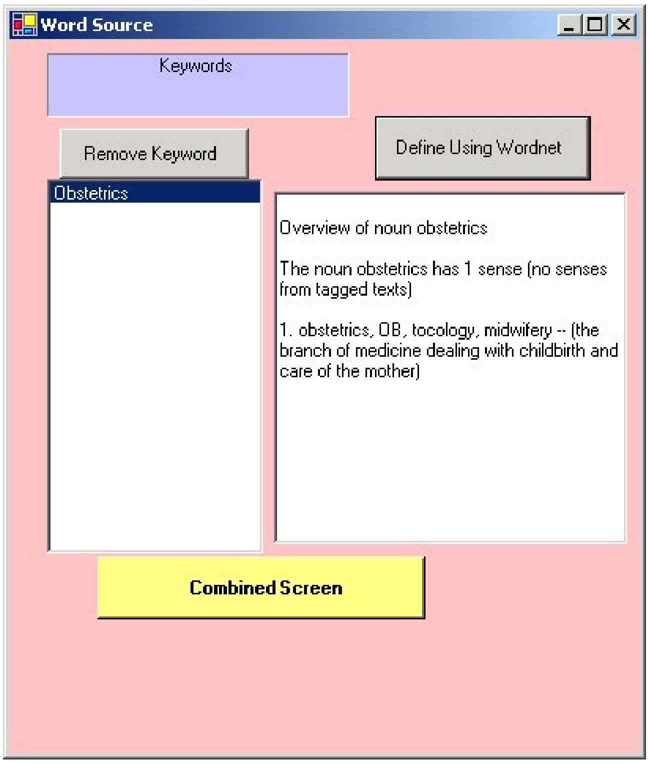
(Source (Dreyfus, Dreyfus, & Athanasiou, 1986), page 22).


Search Details
Enter Search Details
Sex
Age
Choose one - or write in a new one
Choose one - or write in a new one
Type of Data
Choose one - or write in a new one
Show MeSH
Start Searching

Search String Obstetrics

Figure 75: The Search Details Screen

The search details screen (Figure 75) allows the user to select criteria for the search including the search term or phrase, the age range, the gender of the subjects of the documents and the type of documents. These are derived from those discussed in section 1.3. The “Show MeSH” button takes the user directly to the alphabetical MeSH listing Figure 80, whereas the “Start Searching” button allows the search string to be parsed and then brings up the Word Source/Identified Keywords window.



Obstetrics

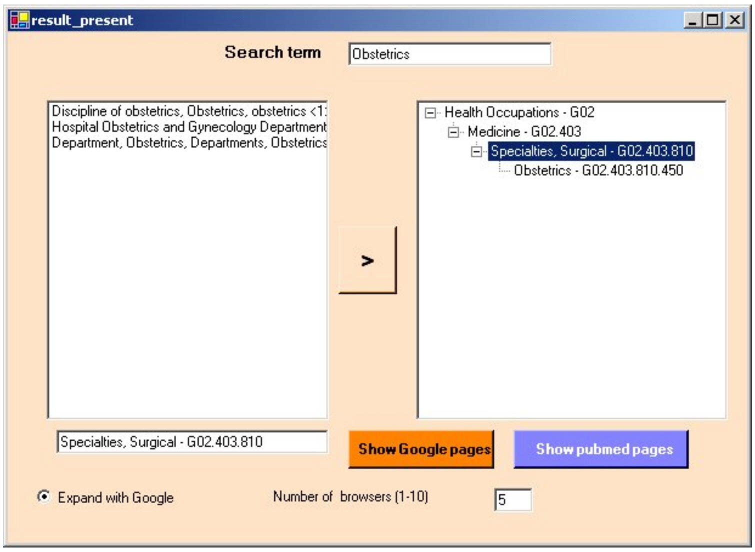
Overview of noun obstetrics

The noun obstetrics has 1 sense (no senses from tagged texts)

1. obstetrics, OB, tocology, midwifery -- (the branch of medicine dealing with childbirth and care of the mother)

Figure 76: The Identified Keywords screen

The identified keywords screen (Figure 76) lists the keywords and phrases identified from the search sting along with any unknown words. Items can be removed from this list using the "remove keyword" button. The WordNet programme (see section 5.2.6) can be called from this screen in order to define the selected term in the left hand list. "Combined screen" then takes the user to the results presentation screen (Figure 77).



Search term: Obstetrics

Discipline of obstetrics, Obstetrics, obstetrics <1:
Hospital Obstetrics and Gynecology Department
Department, Obstetrics, Departments, Obstetrics

Health Occupations - G02
Medicine - G02.403
Specialties, Surgical - G02.403.810
Obstetrics - G02.403.810.450

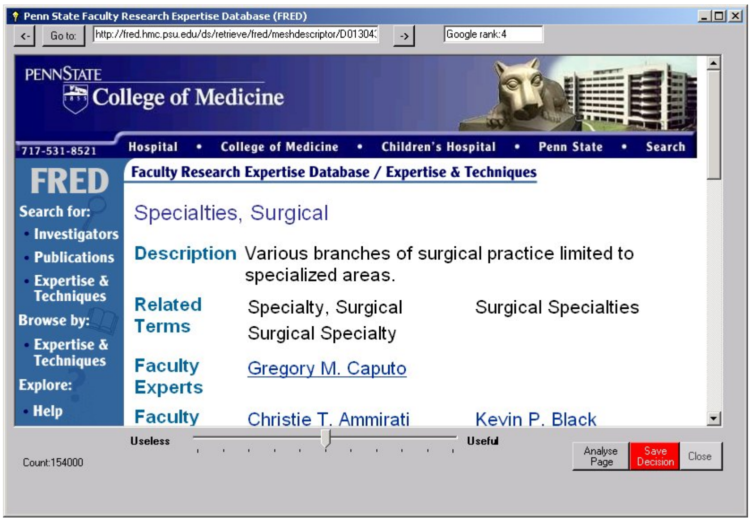
Specialties, Surgical - G02.403.810

● Expand with Google

Number of browsers (1-10): 5

Figure 77: The Results Presentation Screen

The left hand list box contains those concepts identified from the UMLS that may be related to the keyword list previously displayed. The list box on the right displays the Parts of the MeSH ontology related to those terms including locations where the concepts occur along with parents, children and siblings of the first of these concepts. The “>” key expands an item selected from the left hand box. To continue the search, the user selects an item from the right hand box, or types in a value directly into the textbox at the lower right hand side. The user then selects the Google, or PubMed search. The aim of this step is to explicitly identify related terms that the user may wish to select in preference to the originals.



Penn State Faculty Research Expertise Database (FRED)

Go to: http://fred.hmc.psu.edu/ds/retrieve/fred/meshdescriptor/D01304:

Google rank:4

Faculty Research Expertise Database / Expertise & Techniques

Specialties, Surgical

Description Various branches of surgical practice limited to specialized areas.

Related Terms Specialty, Surgical Surgical Specialties Surgical Specialty

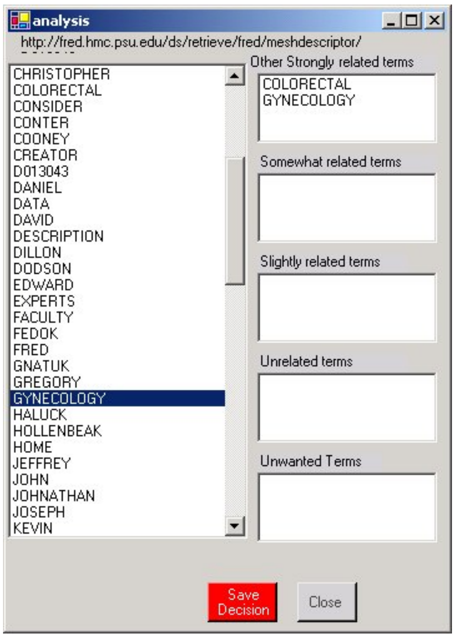
Faculty Experts Gregory M. Caputo

Faculty Christie T. Ammirati Kevin P. Black

Count:154000

Figure 78: The Google Browser Screen

Figure 78 shows one of the screens loaded after an item has been selected and the “Google” button pressed. Up to 10 of these browsers are displayed, the number being selected via the results presentation screen. The PubMed display is similar (see Figure 51). This screen acts as an Internet browser, links can be followed, URLs can be entered and forward and back buttons are provided. The user can rate the page currently being displayed using a visual analogue scale provided as a slider component. This approach to eliciting feedback has been used previously (Lenert LA, 2001) and shown to be acceptable. The design emphasises simplicity of feedback – the user is not expected to rate on multiple scales. If the user does not want to rate a page, they do not have to. Each time a user rates a page that value is updated in the databas,e that is only the most recent rating is retianed. If the user chooses to analyse a page then the analysis page (Figure 79) is displayed.



analysis

http://fred.hmc.psu.edu/ds/retrieve/fred/meshdescriptor/

CHRISTOPHER
COLORECTAL
CONSIDER
CONTER
COONEY
CREATOR
D013043
DANIEL
DATA
DAVID
DESCRIPTION
DILLON
DODSON
EDWARD
EXPERTS
FACULTY
FEDOK
FRED
GNATUK
GREGORY
GYNECOLOGY
HALUCK
HOLLENBEAK
HOME
JEFFREY
JOHN
JOHNATHAN
JOSEPH
KEVIN

Other Strongly related terms

COLORECTAL
GYNECOLOGY

Somewhat related terms

Slightly related terms

Unrelated terms

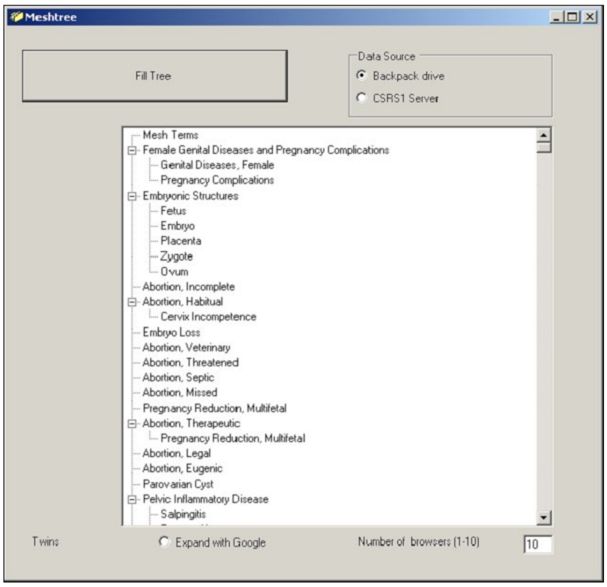
Unwanted Terms

Figure 79: The analysis screen

The analysis pages retrieves all words except stopwords from the page being viewed. The words are sorted alphabetically and duplicates removed. Words can be selected and then dragged across to the appropriate category boxes. The category boxes will scroll if necessary. This page acts as a floating window above the browser, to retain the association with the original page.

8.11.1 Query construction

The construction of the query is the most complicated aspect of the user interface of the system. The original approach of making “everything visible” as shown in Figure 80 resulted in too much information being displayed. This screen occurs at point MeSH/FO in the diagram (Figure 49).



Data Source: ● Backpack drive ○ CSRS1 Server

- Mesh Terms
- Female Genital Diseases and Pregnancy Complications
  - Genital Diseases, Female
  - Pregnancy Complications
- Embryonic Structures
  - Fetus
  - Embryo
  - Placenta
  - Zygote
  - Ovum
- Abortion, Incomplete
- Abortion, Habitual
  - Cervix Incompetence
- Embryo Loss
- Abortion, Veterinary
- Abortion, Threatened
- Abortion, Septic
- Abortion, Missed
- Pregnancy Reduction, Multifetal
- Abortion, Therapeutic
  - Pregnancy Reduction, Multifetal
- Abortion, Legal
- Abortion, Eugenic
- Parovarian Cyst
- Pelvic Inflammatory Disease
  - Salpingitis

○ Expand with Google

Number of browsers (1-10) 10

Figure 80: The scrolling list amounts to up to 1000 top-level words

This interface attempted to represent every item in the MeSH hierarchy simultaneously, with all items at the 3rd and 2nd level at the first level of the tree. These levels were chosen in order to allow users to see terms that they may themselves recognise. Lower items were shown when the + button was pressed. However the following issues were not successfully addressed:

1. Browsing is not easily supported, with the majority of potential terms concealed by the scroll.
2. There is not really a natural order to this scheme for searching behaviour. In fact the terms were ordered by their MeSH codes, which were unlikely to be known by the users. However any ordering scheme seems unlikely to succeed given such a small proportion of visible terms and the wide range of potential homonyms. For example an alphabetical ordering would not have “Embryo Loss” near “Abortion, Spontaneous”, and this is likely to result in a poor selection of terms by the user. User comments (see Chapter 6) however suggested that alphabetical ordering was at least comprehensible and this was adopted.

This “browsing” method is also complimented by a keyword search method. The user is invited to enter a keyword, along with some limits to the search. The following processes happen at this point, generally based around the database;

- The entered string is checked to see if it is a substring of the descriptions contained in the MRCXT table, broadened to include alternative spellings. Any strings or phrases discovered in this way are added to the keyword confirm screen list, and the concept ID added to a hidden field associated with the string. This process continues until there are no more results.

- The entered string is then parsed to discover any words that also occur in the MRCXT table which are not also contained in the stopword table (see section 6.1.2 for details of the stopword table). This process continues until there are no more results.

The example below shows the process:

User enters "Obstetrics and Gynaecology"
"Obstetics and Gynecology…" discovered in MRCXT – spelling of Gynaecology/ Gynecology confirmed using cross reference table.
Obstetrics and Gynecology added
…
Searches for "obstetrics" and "Gynecology"(mapped by the cross referencing) are performed – "and " removed by the stoplist process.
For each distinct conceptID in MRCXT that contains either "obstetrics" and or "Gynecology", the appropriate concept ID is added to the list along with single, visible example of "obstetrics" and "Gynecology".

After the list has been generated the user has the ability to remove items from the keyword list and/or have them defined using WordNet (D.Slomin & Tengi, 2003).

8.12 The WWW fuzzy ontology site

This site is located at: http://csrs2.aut.ac.nz/fuzzont/default.html. The use can select terms, which are keywords, or text words derived from the Reuters-21578 training corpus (see section 6.5). See Figure 81.

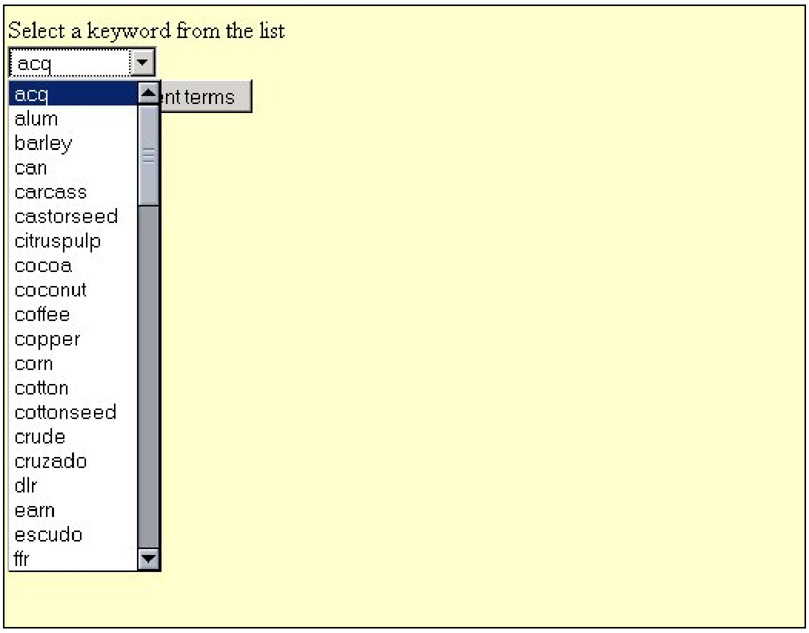
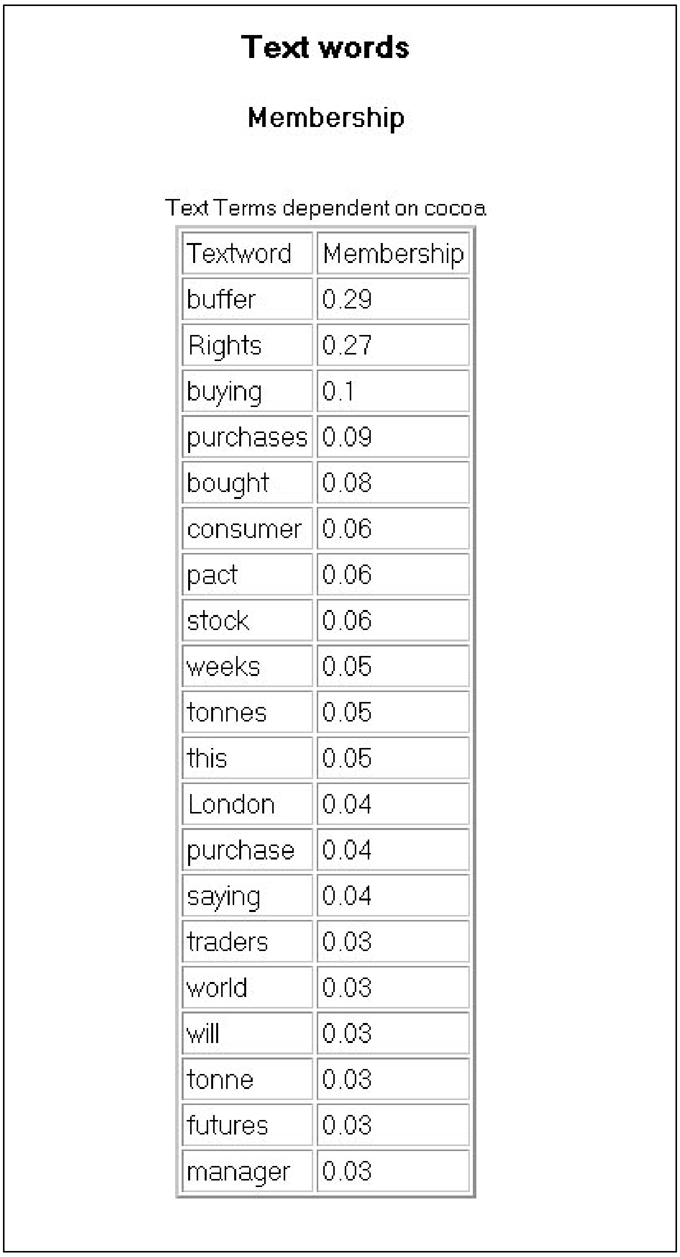


Figure 81: Choice of keyword term

When the user has selected a text word or keyword, the terms related to that term are displayed along with the membership value μ of the relationship between the selected term and the displayed term. For clarity only the top 20 terms that are related are displayed, and they are displayed in descending value of μ . See Figure 82



Text words

Membership

Text Terms dependent on cocoa

Textword	Membership
buffer	0.29
Rights	0.27
buying	0.1
purchases	0.09
bought	0.08
consumer	0.06
pact	0.06
stock	0.06
weeks	0.05
tonnes	0.05
this	0.05
London	0.04
purchase	0.04
saying	0.04
traders	0.03
world	0.03
will	0.03
tonne	0.03
futures	0.03
manager	0.03

Figure 82: The output of the fuzzy linkages WWW page

8.13 The Usability Questionnaire

Questionnaire System Evaluation

Based on: Davis, F. D. (1989) *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology.* **MIS Quarterly**, 13:3, 319-340.

About question.cgi

Please rate the usefulness and ease of use of Medical Searching System.

- Try to respond to all the items.
- For items that are not applicable, use: **N/A**
- Add a comment about an item by clicking on its icon, or add comment fields for all items by clicking on **Comment All**.

System:

Medical Searching System

Optionally provide comments and your email address in the box.

PERCEIVED USEFULNESS		**-2**	**-1**	**0**	**1**	**2**		**N/A**
1. Using Medical Searching System in my job would enable me to accomplish tasks more quickly	DISAGREE	○	○	○	○	○	AGREE	○
2. Using Medical Searching System would improve my job performance	DISAGREE	○	○	○	○	○	AGREE	○
3. Using Medical Searching System in my job would increase my productivity	DISAGREE	○	○	○	○	○	AGREE	○
4. Using Medical Searching System	DISAGREE	○	○	○	○	○	AGREE	○

	would enhance my effectiveness on the job								
5.	Using Medical Searching System would make it easier to do my job	DISAGREE	○	○	○	○	○	AGREE	○
6.	I would find Medical Searching System useful in my job	DISAGREE	○	○	○	○	○	AGREE	○
PERCEIVED EASE OF USE			**-2**	**-1**	**0**	**1**	**2**		**N/A**
7.	Learning to operate Medical Searching System would be easy for me	DISAGREE	○	○	○	○	○	AGREE	○
8.	I would find it easy to get Medical Searching System to do what I want it to do	DISAGREE	○	○	○	○	○	AGREE	○
9.	My interaction with Medical Searching System would be clear and understandable	DISAGREE	○	○	○	○	○	AGREE	○
10.	I would find Medical Searching System to be flexible to interact with	DISAGREE	○	○	○	○	○	AGREE	○
11.	It would be easy for me to become skillful at using Medical Searching System	DISAGREE	○	○	○	○	○	AGREE	○
12.	I would find Medical Searching System easy to use	DISAGREE	○	○	○	○	○	AGREE	○
			-2	**-1**	**0**	**1**	**2**		**N/A**

List the most **negative** aspect(s):

1. ____________________

2. ____________________

3. ____________________

List the most **positive** aspect(s):

1.

2.

3.

4.

5.

References

Abramatic, J.-F. (1997, 1997). *W3C - World Wide Web Consortium*. Retrieved 1st May 2005, from www.w3.org/Talks/9704WWW6-Chairman/slide18.htm

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665-695.

Akkermans, H., Baida, Z., Gordijn, J., Perea, N., Altuna, A., & Laresgoiti, I. (2004). Value Webs: Using Ontologies to Bundle Real-World Services. *Intelligent Systems, IEEE 19*(4), 57-66.

Alani, H., K., S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., et al. (2003). Automatic ontology-based knowledge extraction from Web documents. *Intelligent Systems, IEEE, 18*(1), 14-21.

Alborz, A., & McNally, R. (2004). Developing methods for systematic reviewing in health services delivery and organization: an example from a review of access to health care for people with learning disabilities. Part2. Evaluation of the literature ;a practical guide. *Health Information and Libraries Journal, 21*(4), 227-236.

Alexander, G., Hauser, S., Steely, K., Ford, G., & Demner-Fushman, D. (2004). *A Usability Study of the PubMed on Tap User Interface for PDAs.* Paper presented at the Medinfo 2004, San Fransisco.

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., et al. (2003). Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002 *SIGIR Forum 37* (1), 31-47

Altman, D. G., & Andersen, P. K. (1999). Calculating the number needed to treat for trials where the outcome is time to an event. *British Medical Journal, 319*(7223), 1492-1495.

Ammasai, D. (2004, Updated 8 Jun 2001). *Layman's SOAP*. Retrieved 9th August 2004, from http://www.codeproject.com/soap/laymansoap.asp

Apache Software Foundation. (2000). Xerces Java Parser (Version 1.4.3).

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2--43.

Bachmann, L. M., Coray, R., Estermann, P., & ter Riet, G. (2002). Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read. *J Am Med Inform Assoc, 9*(6), 653-658.

Belew, R. K. (2000). *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.

Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language Trees and Zipping. *Physical Review Letters, 88*(4), 048702-048701 to 048702-048704.

Berland, G. K., Elliott, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., et al. (2001). Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *JAMA, 285*(20), 2612-2621.

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

Berque, D., Johnson, D. K., & Jovanovic, L. (2001). *Teaching theory of computation using pen-based computers and an electronic whiteboard.* Paper presented at the

Proceedings of the 6th annual conference on Innovation and technology in computer science education, Canterbury, United Kingdom.

Bjorneborn, L. (2001). Small-world linkage and co-linkage. 133--137.

Bloom, P., & Keil, F. C. (2001). Thinking Through Language. *Mind and Language, 16*(4), 351-367.

Bodenreider, O. (2001). *Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention.* Paper presented at the AMIA Annual Symposium 2001.

Bodoff, D. (2004). Relevance models to help estimate document and query parameters http://doi.acm.org.ezproxy.aut.ac.nz/10.1145/1010614.1010615 *ACM Trans. Inf. Syst. , 22* (3), 357-380

Bordogna, G., Carrara, P., & Pasi, G. (1995). Fuzzy Approaches to Extend Boolean Approaches to Information Retrieval. In J. Kacprzyk & P. Bosc (Eds.), *Fuzziness in database management systems* (pp. 231-274). Heidelberg: Physica-Verlag.

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

Borgman, C. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Sciences, 37*(6), 387-400.

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

Bosc, P., Galibourg, M., & Hamon, G. (1988). Fuzzy querying with SQL: extensions and implementation aspects. *Fuzzy Sets Syst., 28*(3), 333-349.

Boxwala, A., Ogunyemi, O., & Zeng, Q. (2003). *Computing for Biomedical Scientists - Part of the MIT OpenCourseWare series.* Retrieved October 2003, from http://ocw.mit.edu/NR/rdonlyres/Health-Sciences-and-Technology/HST-952Computing-for-Biomedical-ScientistsFall2002/

Brem, S., & Boyes, A. (2000). Using critical thinking to conduct effective searches of online resources. *Practical Assessment, Research & Evaluation, 7*(7).

Brewington, B., & Cybenko, G. (2000). *How dynamic is the web?* Paper presented at the Proceedings of the Ninth International World Wide Web Conference, Amsterdam.

Brewster, C., O'Hara, K., Fuller, S., Wilks, Y., Franconi, E., Musen, M. A., et al. (2004). Knowledge representation with ontologies: the present and future. *Intelligent Systems, IEEE 19*(1), 72-81.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3-10.

Burdette, S., Herchline, T., & Richardson, W. S. (2004). Killing Bugs at the Bedside: A prospective hospital survey of how frequently personal digital assistants provide expert recommendations in the treatment of infectious diseases. *Annals of Clinical Microbiology and Antimicrobials, 3*(1), 22.

C. Apté, F. Damerau, & S. Weiss. (1994). Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst., 12*(3), 233-251.

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

Caterinicchia, D. (2003, August 11 2003). *DOD's misguided bet on the future*. Retrieved 17/12/2004, from http://www.fcw.com/fcw/articles/2003/0811/pol-misguided-08-11-03.asp

Chen, M.-S., Han, J., & Yu, P. (1996). Data Mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883.

Cohen, A., Stavri, P., & Hersh, W. (2004). A categorization and analysis of the criticisms of Evidence-Based Medicine. *International journal of medical informatics., 73*(1), 35-43.

Coiera, E., & Tombs, V. (1998). Communication behaviours in a hospital setting: an observational study. *British Medical Journal, 316*(7132), 673-676.

Crestani, F., Lalmas, M., van Rijesbergen, C., & Campbell, I. (1998). "Is this document relevant ?.. Probably": A survey of Probabilistic Methods in Information Retrieval. *ACM Computing Surveys, 30*(4), 528-552.

Cullen, R. J. (2002). In search of evidence: family practitioners' use of the Internet for clinical information. *Journal of the Medical Librarians association, 90*(4), 370-379.

D.Slomin, & Tengi, R. (2003). Wordnet (Version 2.0): Princeton University Cognitive Science Lab.

Davidoff, F., Haynes, B., Sackett, D., & Smith, R. (1995). Evidence based medicine *British Medical Journal, 310*(6987), 1085-1086.

Davis, F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319-340.

Dhyani, D., Ng, W. K., & Bhowmick, S. S. (2002). A survey of Web metrics. *ACM Computing Surveys, 34*(4), 469-503.

Doucet, A., & Ahonen-Myka, H. (2002). Naive clustering of a large XML document collection. *Initiative for the Evaluation of XML retrieval*.

Dray, S., Siegel, D., Feldman, E., & Potenza, M. (2002). Why do version 1.0 and not release it?: Conducting field trials of the tablet PC. *interactions, 9*(2), 11--16.

Dreyfus, H. L., Dreyfus, S. E., & Athanasiou, T. (1986). *Mind over machine : the power of human intuition and expertise in the era of the computer*. Oxford: Blackwell.

Eastman, C. M. (2002). 30,000 hits may be better than 300: precision anomalies in internet searches
http://dx.doi.org.ezproxy.aut.ac.nz/10.1002/asi.10115 *J. Am. Soc. Inf. Sci. Technol. , 53* (11), 879-882

Eastman, C. M., & Jansen, B. J. (2003). Coverage, relevance, and ranking: The impact of query operators on Web search engine results. *ACM Trans. Inf. Syst., 21*(4), 383-411.

Ensing, M., Paton, R., Speel, P. H., & Rada, R. (1994). An object-oriented approach to knowledge representation in a biomedical domain. *Artificial Intelligence in Medicine, 6*, 459-482.

Eysenbach, G. (2002). Infodemiology: the epidemiology of (mis)information. *The American Journal of Medicine, 113*(9), 763-765.

Eysenbach, G. (2004). Websites on screening for breast cancer: "Infodemiology" studies have surely had their day. *British Medical Journal, 328*(7442), 769-b-.

Eysenbach, G., & Diepgen, T. L. (1998a). Labeling and filtering of medical information on the Internet. *Methods of Information in Medicine, 38*, 80-88.

Eysenbach, G., & Diepgen, T. L. (1998b). Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. *British Medical Journal, 317*(7171), 1496-1500.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E. (2002). Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. *JAMA, 287*(20), 2691-2700.

Fåhræus, E. R., Bridgeman, N., Rugelj, J., Chamberlain, B., & Fuller, U. (1999). *Teaching with Electronic Collaborative Learning Groups.* Paper presented at the

Annual Joint Conference Integrating Technology into Computer Science Education - Working group reports from ITiCSE on Innovation and technology in computer science education, Cracow, Poland.

Farahat, A., Nunberg, G., Chen, F., & Heylighen, F. (2002). AuGEAS: authoritativeness grading, estimation, and sorting: Collective Intelligence and its Implementation. *Computational & Mathematical Organization Theory, 5*(3), 194--202.

Fayyad, U. (1998). Diving into Databases. *Database Programming and Design*(March), 24-31.

Fayyad, U., & Piatetsky-Shapiro, G. (1996). From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery in Databases and Data Mining* (pp. 1-34).

Fragoudis, D., & Likothanassis, S. D. (1999). *Retriever: an agent for intelligent information recovery.* Paper presented at the Information Intelligence and Systems, 1999. Proceedings. 1999 International Conference on.

Fu, Y., & Mostafa, J. (2004). Toward information retrieval web services for digital libraries In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (pp. 370-371). Tuscon, AZ, USA ACM Press.

Gardner, M. (1997). "Information retrieval at the point of care". *British Medical Journal, 314*, 950-953.

Garg, A., & Turtle, K. M. (2003). Effectiveness of training health professionals in literature search skills using electronic health databases;a critical appraisal. *Health Information and Libraries Journal, 20*(1), 33-41.

Gerwe, P., & Viles, C. L. (2000). User effort in query construction and interface selection In *Proceedings of the fifth ACM conference on Digital libraries* (pp. 246-247). San Antonio, Texas, United States ACM Press.

GlaxoSmithKline. (2003). *Prescription Medicine List*. Retrieved 20th August 2004, from http://www.gsk.com/products/prescriptionmedicines.jsp

Glover, E., Lawrence, S., Birmingham, W., & Giles, C. L. (1999). *Architecture of a Metasearch Engine that Supports User Information Needs.* Paper presented at the Eighth International Conference on Information and Knowledge Management (CIKM'99).

Glover, E., Lawrence, S., Gordon, M. D., Birmingham, W., & Giles, C. L. (2001). Web Search---Your Way. *Communications of the ACM, 44*(12), 97 - 102.

Goldstein, M., Nyberg, M., & Anneroth, M. (2003). Providing proper affordances when transferring source metaphors from information appliances to a 3G mobile multipurpose handset. *Personal Ubiquitous Comput., 7*(6), 372-380.

Gori, M., & Witten, I. (2005). The bubble of web visibility. *Communications of the ACM, 48*(3), 115-117.

Gosling, A. S., Westbrook, J. I., & Coiera, E. W. (2003). Variation in the use of online clinical evidence: a qualitative analysis. *International Journal of Medical Informatics, 69*(1), 1-16.

Green, M. L., & Ruff, T. R. (2005). Why Do Residents Fail to Answer Their Clinical Questions? A Qualitative Study of Barriers to Practicing Evidence-Based Medicine

Acad Med, 80(2), 176-182.

Grimes, D. A., & Schulz, K. F. (2002). An overview of clinical research: the lay of the land. *The Lancet, 359*(9300), 57-62.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5*(2), 199-220.

Gruppen, L. D., Rana, G. K., & Arndt, T. S. (2005). A controlled comparison study of the efficacy of training medical students in evidence-based medicine literature searching skills. *Academic Medicine, 80*(10), 940-944.

Grutter, R., Eikemeier, C., & Steurer, J. (2001). *Up-scaling a semantic navigation of an evidence-based medical information service on the Internet to data intensive extranets.* Paper presented at the User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on, Inst. for Media & Commun. Manage., Univ. of St. Gallen, Switzerland.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002 -a). Cluster validity methods: part I *SIGMOD Rec. , 31* (2), 40-45

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002 -b). Clustering validity checking methods: part II *SIGMOD Rec. , 31* (3), 19-27

Haramundanis, K. (2001). *Learnability in information design.* Paper presented at the Proceedings of the 19th annual international conference on Computer documentation, Sante Fe, New Mexico, USA.

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *British Medical Journal, 323*(11 August), 334-336.

Hardwick, J. C. R., & MacKenzie, F. M. (2003). Information contained in miscarriage-related websites and the predictive value of website scoring systems. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 106*(1), 60-63.

Hassoun, A., Vellozzi, E. M., & Smith, M. A. (2004). Colonization of personal digital assistants carried by healthcare professionals. *Infect Control Hosp Epidemiol., 25*(11), 1000-1001.

Haynes, R. B., & Wilczynski, N. L. (2004). Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *British Medical Journal, 328*(7447), 1040-1040.

Health on the Net Foundation. (2003). *HON code of conduct*. Retrieved 31/12/2003, from http://www.hon.ch/honcode/conduct.html.

Hertzberg, S., & Rudner, L. (1999). The Quality of Researchers' Searches of the ERIC Database. *Education Policy Analysis Archives, 7*(25).

Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. *Nature, 430*(6998), 453-456.

Hilbert, D., & Trevor, J. (2004). Personalizing shared ubiquitous devices. *interactions, 11*(3), 34-43.

Hogg, K., Chilcott, P., Nolan, M., & Srinivasan, B. (2004). *An evaluation of Web services in the design of a B2B application.* Paper presented at the The 27th conference on Australasian computer science, Dunedin.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264-323.

Jansen, B., & Spink, A. (2005). An analysis of web searching by European All the Web.com users. *Inf. Process. Manage., 41*(2), 361-381.

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7,* 217-240.

Jin, R., Chai, J. Y., & Si, L. (2004). An automatic weighting scheme for collaborative filtering. In *Proceedings of the 27th annual international conference on Research and development in information retrieval* (pp. 337-344). Sheffield, United Kingdom: ACM Press.

K. F. Eustice, T. J. Lehman, A. Morales, M. C. Munson, S. Edlund, & Guillen, M. (1999). A Universal Information Appliance. *IBM Systems Journal, 38*(4), 575-601.

Kacprzyk, J., & Bosc, P. (1995). *Fuzziness in database management systems.* Heidelberg: Physica-Verlag.

Kacprzyk, J., & Zadrozny, S. (2003). Internet as a challenge to fuzzy querying. In *Intelligent exploration of the web* (pp. 74-95): Physica-Verlag GmbH.

Kacprzyk, J., & Zilkowski, A. (1986). Database queries with fuzzy linguistic quantifiers. *IEEE Trans. Syst. Man Cybern., 16*(3), 474-479.

Kagolovsky, Y. M., JR. (2001). *A New Approach to the Concept of "Relevance" in Information Retrieval (IR).* Paper presented at the Medinfo 2001, London.

Karp, P., Chaudhri, V., & Thomere, J. (2004). *XOL - Ontology Exchange Language*. Retrieved 1st September 2002, from http://www.ai.sri.com/pkarp/xol/xol.html

Kasabov, N. (2002). *Evolving Connectionist Systems*. London: Springer-Verlag.

Kidd, A. (1994). *The Marks are on The Knowledge Worker.* Paper presented at the Human factors in computing systems:celebrating interdependence, Boston.

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). Evidence-Based Software Engineering. 273-281.

Kodratoff, Y. (2001). Comparing Machine Learning and Knowledge Discovery in Databases. In G. Paliouras, V. Karkaletsis & C. Spyropoulos (Eds.), *Machine Learning and Its Applications* (pp. 1-21). Heidelgerg: Springer.

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

Kyung-Sam Choi, Chi-Hoon Lee, & Phill-Kyu Rhee. (2000). *Document ontology based personalized filtering system.* Paper presented at the International Multimedia Conference: Proceedings of the eighth ACM international conference on Multimedia, Marina del Rey, California, United States.

Lederer, A. L., Maupin, D. J., Sena, M. P., & Zhuang, Y. (2000). The technology acceptance model and the World Wide Web. *Decision Support Systems, 29*(3), 269-282.

Lenert LA, S. A. (2001, 2001). *Acceptability of computerized visual analog scale, time trade-off and standard gamble rating methods in patients and the public.* Paper presented at the Proceedings of the Annual AMIA Symposium 2001.

Littlejohns, P., Wyatt, J. C., & Garvican, L. (2003). Evaluating computerised health information systems: hard lessons still to be learnt. *British Medical Journal, 326*(7394), 860-863.

Liu, H., Lieberman, H., & Selker, T. (2002). *GOOSE: A Goal-Oriented Search Engine with Commonsense.* Paper presented at the Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems,.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2,* 159-165.

Mackinlay, J. D., & Zellweger, P. T. (1995). *Browsing vs. search: can we find a synergy? (panel session).* Paper presented at the Conference companion on Human factors in computing systems, Denver, Colorado, United States,.

Maedche, A. M., B.; Stojanovic, L.; Studer, R.; Volz, R. (2003). Ontologies for enterprise knowledge management. *Intelligent Systems, IEEE 18*(2), 26-33.

Mamdani, E. H., & Bonissone, P. (2004). *Soft computing as a tool.* Paper presented at the Proceedings of the IEEE International Conference on Fuzzy Systems.

Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N. L., et al. (2003). Estrogen plus Progestin and the Risk of Coronary Heart Disease. *N Engl J Med, 349*(6), 523-534.

Mayer, M., Darmoni, S. J., Fiene, M., Kohler, C., Roth-Berghofer, T. R., & Eysenbach, G. (2003). MedCIRCLE: collaboration for Internet rating, certification, labelling and evaluation of health information on the World-Wide-Web. *Stud Health Technol Inform, 95.,* 667-672.

McCahan, J. (1998). Turf Wars and Curriculum Change [Letter]. *Academic Medicine, 73*(3), 221.

Middleton, S. E., Shadbolt, N. R., & Roure, D. C. D. (2004). Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst., 22*(1), 54--88.

Mihalcea, R. F., & Mihalcea, S. I. (2001). *Word semantics for information retrieval: moving one step closer to the Semantic Web.* Paper presented at the Tools with Artificial Intelligence, Proceedings of the 13th International Conference on, Southern Methodist Univ., Dallas, TX, USA.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic Personalization based on Web Usage Mining. *Communications of the ACM, 43*(8), 142-151.

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

Mosteller F., & Wallice D. (1964). *Applied Bayesian and Classical Inference: the case of the Federalist Papers*: Addison-Wesley.

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

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

National Library of Medicine. (2002). *Pubmed*. Retrieved 1/05/2002, from http://pubmed.gov

National Library of Medicine. (2003, 21 July 2003). *Esearch Entrez Utility*. Retrieved 2003, from http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esearch_help.html

Neilson, J. (2000). *Designing web usability*. Indianapolis, Ind.: New Riders.

Neilson, J. (2005, October 3rd 2005). *Top Ten Web Design Mistakes of 2005*. Retrieved 1st December 2005, from http://www.useit.com/alertbox/designmistakes.html

Nelson, S., Schopen, M., J., S., & N., A. (2001). *An Interlingual Database of MeSH Translations*. Retrieved October, from http://www.nlm.nih.gov/mesh/intlmesh.html

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). *UCI Repository of machine learning databases*. Retrieved 1st December 2005, from http://www.ics.uci.edu/~mlearn/MLRepository.html

Northwestern Mutual Life Insurance Company. (2004). *The Longevity Game!* Retrieved 17/12/04, from http://www.nmfn.com/tn/learnctr--lifeevents--longevity

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

Noy, N. F., & McGuinness, D. (2001). *Ontology Development 101: A guide to Creating your First Ontology*: Stanford University.

Noy, N. F., & Musen, M. A. (1999, 1999). *SMART:Automated support for Ontology Merging and Alignment*. Retrieved 3rd October 2003, from http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-1999-0813.pdf

Nygren, E., M. Lind, M. Johnson, and B. Sandblad. (1992). The Art of the Obvious. In *Proceedings ACM Conf. Human Factors in Computing Systems (CHI '92).* (pp. 235-239). New York: ACM.

Olston, C., & Chi, E. H. (2003). ScentTrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction., 10*(3), 177--197.

Oman, P. W., & Cook, C. R. (1989). Programming style authorship analysis. 320--326.

Oren Zamir, & Etzioni, O. (1999). *Grouper: A Dynamic Clustering Interface to Web Search Results.* Paper presented at the 8th World Wide Web Conference., Toronto, Canada.

Oxford University Computing Services. (2001). British National Corpus (Version 2). Oxford.

Padden, M. (1999). HELLP Syndrome: Recognition and Perinatal Management. *American Family Physician, 60*(3), 829-836, 839.

Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G., & Cho, J. (2000). Beyond document similarity: understanding value-based search and browsing technologies. *SIGMOD Rec., 29*(1), 80-92.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*: Stamford University.

Palen, L., & Salzman, M. (2002). Beyond the handset: designing for wireless communications usability. *ACM Transactions on Computer-Human Interaction, 9*(2), 125-151.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*: Morgan Kaufmann Publishers Inc.

Perlman, G. (2001). *Web-Based User Interface Evaluation with Questionnaires*. Retrieved 1/3/2003, from http://www.acm.org/~perlman/

Pirolli, P., Card, S. K., & Wege, M. M. V. D. (2003). The effects of information scent on visual search in the hyperbolic tree browser. *ACM Transactions on Computer-Human Interaction, 10*(1), 20--53.

Plain English Campaign. (2003, 1 /12/2003). *"The Foot in Mouth award"*. Retrieved 31/12/2003

Powell, J. A., Lowe, P., Griffiths, F. E., & Thorogood, M. (2005). A critical analysis of the literature on the Internet and consumer health information. *Journal of Telemedicine and Telecare, 11*(Supp 1.), 41-44.

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction Design : Beyond human-computer interaction*. New York, NY: J. Wiley & Sons.

Puglisi, A., Benedetto, D., Caglioti, E., Loreto, V., & Vulpiani, A. (2003). Data compression and learning in time sequences analysis. *Physica D: Nonlinear Phenomena, 180*(1-2), 92-107.

Quinlan, J. R. (1993). *C4.5:programs for machine learning*. San Francisco: Morgan Kaufman.

Quinlan, R. J. (1991). Knowledge Acquisition from Structured Data. *IEEE Expert*(December), 32-37.

Quintana, Y. (1998). Intelligent medical information filtering. *International Journal of Medical Informatics, 51*((2-3)), 197-204.

R. L. Sribnick, & W. B. Sribnick. (1994). *Smart Patient, Good Medicine: Working With Your Doctor to Get the Best Medical Care*. New York: Walker and Company.

Rauber, A., & Mller-Kgler, A. (2001). Integrating automatic genre analysis into digital libraries. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries* (pp. 1-10). Roanoke, Virginia, United States: ACM Press.

Raza, N., Bradshaw, V., & Hague, M. (1999). *Applications of RFID technology*. Paper presented at the IEE Colloquium on RFID Technology,.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186). Chapel Hill, North Carolina, United States ACM Press.

Retrieval, O. T. (2003). *Stop Word List 1*. Retrieved 30/4/2004, from http://www.lextek.com/manuals/onix/stopwords1.html

Reuters. (1987, 16 Feb 1999). *Reuters-21578 Text Categorization Collection*. Retrieved 1/12/2004, from http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Richardson, C., Resnick, P., Hansen, D., & Rideout, V. (2002). Does Pornography-Blocking Software Block Access to Health Information on the Internet? *JAMA, 288*(22), 2887-2894.

Rideout, V. (2003). Internet filters block valuable data, too. *USA Today*.

Ritter, H., & Kohonen, T. (1989). Self-Organizing Semantic Maps. *Biological Cybernetics, 61*, 241-254.

Sackett, D., Richardson, W., Rosenberg, W., & Haynes, B. R. (1997). *Evidence Based Medicine - How to Practice and Teach EBM*: Churchill Livingstone.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't

British Medical Journal, 312(7023), 71-72.

Sallis, A., Carran, G., & Bygrave, J. (2000). *The Development of a Collaborative Learning Environment: Supporting the Traditional Classroom.* Paper presented at the WWW9, Netherlands.

Sallis, P., & Kasabova, D. (2000). Computer-Mediated Communication, Experiments with e-mail readability. *Information Sciences*(123), 43-53.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Sanderson, M., & Croft, B. (1999). *Deriving concept hierarchies from text.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States.

Saracevic, T. (1975). Relevance: A review of and a framework for thinking. *Journal of the American Society for Information Sciences, 26*, 321-343.

Sculpher, M., Drummond, M., & O'Brien, B. (2001). Effectiveness, efficiency, and NICE. *British Medical Journal, 322*(7292), 943-944.

Sedelow, S. Y. (1970). The Computer in the Humanities and Fine Arts. *ACM Computing Surveys (CSUR), 2*(2), 89-110.

Shannon. (1948). Mathematical Theory of Communication. In *Bell Systems Technical Journal*.

Sharon, T., Lieberman, H., & Selker, T. (2003). *A zero-input interface for leveraging group experience in web browsing.* Paper presented at the Proceedings of the 8th international conference on Intelligent user interfaces, Miami, Florida, USA.

Shaughnessy, A. F., & Slawson, D. C. (1999). Are we providing doctors with the training and tools for lifelong learning? *British Medical Journal, 319*(7220), 1280-.

Shaughnessy, A. F., & Slawson, D. C. (2003). What happened to the valid POEMs? A survey of review articles on the treatment of type 2 diabetes

10.1136/bmj.327.7409.266. *British Medical Journal, 327*(7409), 266-.

Shneiderman, B. (1998). *Designing the user interface : strategies for effective human-computer interaction*. Reading, Mass.: Addison Wesley.

Silberg, W. M., Lundberg, G. D., & Musacchio, R. A. (1997). Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. *JAMA: The Journal Of The American Medical Association, 277*(15), 1244-1245.

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

Singhe, S. T., F.J. (1995). Neural networks and disputed authorship: new challenges *Artificial Neural Networks, 1995., Fourth International Conference on* (pp. 24-28).

Sirin, E., Parsia, B., & Hendler, J. (2004). Filtering and Selecting Semantic Web Services with Interactive Composition Techniques. *Intelligent Systems, IEEE 19*(4), 42-49.

Slawson, D. C., & Shaughnessy, A. F. (1997). Obtaining useful information from expert based sources. *British Medical Journal, 314,* 947-949.

Slawson, D. C., & Shaughnessy, A. F. (2005). Teaching Evidence-Based Medicine: Should We Be Teaching Information Management Instead? *Academic Medicine, 80*(7), 685-689.

Slaytor, E. K., & Ward, J. E. (1998). How risks of breast cancer and benefits of screening are communicated to women: analysis of 58 pamphlets. *British Medical Journal, 317*(7153), 263-264.

Smith, M. K., Welty, C., & Deborah L. McGuinness. (2004, 10/02/2004). *OWL Web Ontology Language Guide*. Retrieved 1st March 2004, from http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OntologyMapping

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28,* 111-121.

Spidlen, J., Hanzlicek, P., Riha, A., & Zvarova, J. (2005). Flexible information storage in MUDR(II) EHR. *International journal of medical informatics.*

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer, 35*(3), 107-109.

Srihari, S. N., & Lee, S. (2002). *Automatic handwriting recognition and writer matching on anthrax-related handwritten mail.* Paper presented at the Eighth International Workshop on Frontiers in Handwriting Recognition.

Staab, S., & Maedche, A. (2001). Knowledge Portals:Ontologies at Work. *AI Magazine, 22*, 63-85.

Stanford University Knowledge Systems Laboratory. (2001). *Ontolingua Ontology Editor*. Retrieved 1/3/2004, from http://www-ksl-svc.stanford.edu:5915/&service=frame-editor

Stanford, V. (2003). Pervasive computing goes the last hundred feet with RFID systems. *Pervasive Computing, IEEE, 2*(2), 9-14.

Stavri, P. (2001). *Personal Health Information-Seeking: A qualitative review of the literature.* Paper presented at the Medinfo 2001, London.

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

Stienbach, M., Karypis, G., & Kumar, V. (2000). *A Comparison of Document Clustering Techniques* (Technical report No. 00-034): University of Minnesota.

STN. (2003). *STN easy list of stop words*. Retrieved 30th June 2003, from http://stneasy.cas.org/html/english/helps/2search/2B4stopw.htm

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*: Doubleday.

Swanson, D. R. (1988). Historical Note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39,* 92-98.

Takagi, T., Kawase, K., Otsuka, K., & Yamaguchi, T. (2000). *Data retrieval using conceptual fuzzy sets.* Paper presented at the Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on, Dept. of Comput. Sci., Meiji Univ., Kanagawa, Japan.

Talburt, J. (1985, 1986). *The Flesch index: An easily programmable readability analysis algorithm.* Paper presented at the Annual ACM Conference on Systems Documentation. Proceedings of the Fourth International Conference on Systems documentation,.

Taylor, J. R. (2003). *Linguistic Categorization* (Third ed.). Oxford: Oxford University Press.

Thunell, L., Milsom, I., Schmidt, J., & Mattsson, L. (2006). Scientific evidence changes prescribing practice—a comparison of the management of the climacteric and use of hormone replacement therapy among Swedish gynaecologists in 1996 and 2003. *British Journal of Obstetrics and Gynaecology, 113*(1), 15-20.

van Rijsbergen, C. J. (1979). *Information Retrieval* (Second ed.).

Vel, O. d., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record, 30*(4), 55-64.

W3C. (1999, 24th December 1999). *HTML 4.01 Specification*. Retrieved 15th August 2004, from http://www.w3.org/TR/html4/

W3C. (2000, 6 October 2000). *Extensible Markup Language (XML) 1.0 (Second Edition)*. Retrieved 11 January 2001, from http://www.w3.org/TR/2000/REC-xml-20001006

W3C. (2004a, 14/02/2004). *OWL Web Ontology Language Overview -W3C Recommendation 10 February 2004*. Retrieved 16/02/2004

W3C. (2004b). *Web Services*. Retrieved 9th August 2004, from http://www.w3.org/2002/ws/

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. *Information Processing & Management, 36*(2), 229-251.

Wang, S.-C., Yu, C.-Y., Liu, K.-P., & Li, S.-P. (2004). *A Web-Based Political Exchange for Election Outcome Predictions.* Paper presented at the Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on.

Want, R., Fishkin, K. P., Gujar, A., & Harrison, B. L. (1999). Bridging physical and virtual worlds with electronic tags. 370--377.

Weiss, A. (2001). Microsoft's .NET: platform in the clouds. *netWorker, 5*(4), 26-31.

Weiss, A. (2003). Me and my shadow. *netWorker, 7*(3), 24--30.

Westbrook, J., Gosling, A. S., & Coiera, E. (2004). Do Clinicians Use Online Evidence to Support Patient Care? A Study of 55,000 Clinicians. *Journal of the American Medical Informatics Association., 11*(2), 113-121.

White, H. D., Lin, X., Buzydlowski, J. W., & Chen, C. (2004). User-controlled mapping of significant literatures. *PNAS, 101*(suppl_1), 5297-5302.

White, R. W., Jose, J. M., & Ruthven, I. (2003). An approach for implicitly detecting information needs. 504--507.

Widyantoro, D. H. Y., J. (2001). *Using fuzzy ontology for query refinement in a personalized abstract search engine.* Paper presented at the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 2001.

Wilczynski, N. L., & Haynes, R. B. (2004). Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. *BMC Medicine, 2*(23).

Wilczynski, N. L., McKibbon, K. A., & Haynes, R. B. (2001). *Enhancing Retrieval of Best Evidence for Health Care from Bibliographic Databases: Calibration of the Hand Search of the Literature.* Paper presented at the Medinfo 2001, London.

Williamson, K. (2005). Where One Size Does Not Fit All: Understanding the Needs of Potential Users of a Portal to Breast Cancer Knowledge Online. *Journal of Health Communication, 10*(6), 567 - 580.

Wilson, D., & Sperber, D. (2002). Relevance Theory. In G. Ward & L. Horn (Eds.), *Handbook of Pragmatics* (pp. 249-287).

Wilson, R. (2002). The "look and feel" of an ebook: considerations in interface design In *Proceedings of the 2002 ACM symposium on Applied computing* (pp. 530-534). Madrid, Spain ACM Press.

Witten, I., & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufmann.

Wolfram, D., Spink, A., Jansen, B., & Saracevic, T. (2001). "Vox populi: The public searching of the web". *Journal of the American Society of Information Science, 52*(12), 1073-1074.

Wollersheim, D., & Rahayu, W. (2002). *Methodology for creating a sample subset of dynamic taxonomy to use in navigating medical text databases.* Paper presented at the Database Engineering and Applications Symposium, 2002. Proceedings. International.

Wolpert, D., & Tumer, K. (2000). *An introduction to Collective intelligence*: NASA Ames Research Centre.

Wood, F. B., Benson, D., LaCroix, E., Siegel, E. R., & Fariss, S. (2005). Use of Internet Audience Measurement Data to Gauge Market Share for Online Health Information Services. *Journal of Medical Internet Research, 7*(3), e31.

Wozar, J., & Worona, P. (2003). The use of online information resources by nurses. *J Med Libr Assoc, 91*(2), 216-221.

Wyatt, J. C. (1997). Commentary: measuring quality and impact of the World Wide Web. *British Medical Journal (Clinical Research Ed.), 314*(7098), 1879-1881.

Wyatt, J. C., & Sullivan, F. (2005a). eHealth and the future: promise or peril?
10.1136/bmj.331.7529.1391. *British Medical Journal, 331*(7529), 1391-1393.

Wyatt, J. C., & Sullivan, F. (2005b). Keeping up: learning in the workplace
10.1136/bmj.331.7525.1129. *British Medical Journal, 331*(7525), 1129-1132.

Xiang, Y., Wong, S. K. M., & Cercone, N. (1995). Quantification of uncertainty in classification rules discovered from databases. *Computational.Intelligence., 11*, 427-441.

Xiaolan Zhu, S. G., Lutz Gerhard, Nicholas Kral, Alexander Pretschner. (1999). Ontology-Based Web Site Mapping for Information Exploration. In *CIKM* (pp. 188-194).

Y. Kagolovsky, & J.R. Moehr. (2000). *Evaluation of Information Retrieval: Old Problems And New Perspectives.* Paper presented at the 8th International Congress on Medical Librarianship, London.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval, 1*(1/2), 69-90.

Young, P. M. C., Leung, R. M. W., Ho, L. M., & McGhee, S. M. (2001). An evaluation of the use of hand-held computers for bedside nursing care. *International Journal of Medical Informatics, 62*(2-3), 189-193.

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

Zadrozny, S., & Kacprzyk, J. (1996). FQUERY for Access: towards human consistent querying user interface. In *Proceedings of the 1996 ACM symposium on Applied Computing* (pp. 532-536). Philadelphia, Pennsylvania, United States: ACM Pres.

Zipf, G. (1949). *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Massachusetts: Addison Wesley.

Ziv, J. L., A. (1977). A universal algorithm for sequential data compression. *Information Theory, IEEE Transactions on, 23*(0018-9448), 337-343.

Zweigenbaum, P., Jacquemart, P., Grabar, N., & Habert, B. (2001). *Building a text corpus for representing the variety of medical knowledge.* Paper presented at the Medinfo 2001, London.