# A Review of Ontology based Query Expansion

**J. BHOGAL[1], A. MACFARLANE, P. SMITH**

*School of Informatics, City University,*
*Northampton Square, London , EC1V 0HB*
*[1]Dept. of Computing, University of Central England*
*Feeney Building, Perry Barr, Birmingham, B42 2SU*

ABSTRACT

This paper examines the meaning of context in relation to ontology based query expansion and contains a review of query expansion approaches. The various query expansion approaches include relevance feedback, corpus dependent knowledge models and corpus independent knowledge models. Case studies detailing query expansion using domain-specific and domain-independent ontologies are also included. The penultimate section attempts to synthesise the information obtained from the review and provide success factors in using an ontology for query expansion. Finally the area of further research in applying context from an ontology to query expansion within a newswire domain is described.

## 1. INTRODUCTION

Information space in online retrieval systems is comparatively larger than traditional information retrieval systems and combined with the ambiguity of the English language, a search query quite often results in a long list of results being returned, much of which are not always relevant to the user's information needs. The main difference in online retrieval systems and traditional information retrieval systems is that the former are usually web-based and as a result the document collection is more dynamic or fluid. To increase the number of relevant documents retrieved, queries need to be disambiguated by looking at their context. Query expansion techniques range from relevance feedback mechanisms to use of knowledge models such as ontologies to resolve ambiguities. This paper reviews the different approaches that have been used for query expansion and highlights important related issues. The structure of the paper is as follows. Section 2 gives a general introduction to context in information retrieval. Section 3 outlines the motivation for this research paper. The overall aim of query expansion and the benefits provided to the user are given in Section 4 which also looks at the different approaches used to derive context for the purposes of query expansion. Corpus dependent and Corpus independent knowledge models are described in sections 5 and 6 respectively. Section 7 highlights the success factors in using an ontology for query expansion. An evaluation of WordNet against these success factors is given in section 8. The area of further research is described in section 9 and a summary is given in section 10.

## 2. WHAT IS CONTEXT IN INFORMATION RETRIEVAL?

Context does not have a standard definition (Finkelstein et al 2002). The Merriam-Webster Online (2006) dictionary contains two main definitions of context. The first is a linguistic definition of context "the parts of a discourse that surround a word or passage and can throw light on its meaning". The second is a more general definition based on circumstances "the interrelated conditions in which something exists or occurs". Most researchers have interpreted the linguistic definition as being appropriate to information retrieval but the general definition is beginning to get more recognition in information retrieval. Context can be based on a particular knowledge domain or it can be related to a specific task. The following sections summarise the various IR approaches to context some of which overlap with the disciplines of artificial intelligence and statistics.

### 2.1 Personalisation

Personalisation is when the system records the history of queries and documents viewed and reuses this information in future searches. It involves learning short term and long-term user models ie user interest and preferences over time. The user models provide context for queries and other interactions with the information system. So two users with the same query could end up with different results if they have different profiles. Two examples of Personalisation are SearchPad (Bharat, 2000) and Remembrance Agent (Rhodes 2000).

Recently personalisation has moved towards user community based information examples of which are collaborative filtering and collaborative querying. Collaborative filtering is a process which looks for votes from users with the same voting patterns as a given user or active user and then attempts to automatically predict votes for the active user. Recommendation systems use the votes predicted by collaborative filtering to make automatic recommendations to the active user. The assumption is that those who agreed in the past tend to agree again in the future. For example a music collaborative filtering or recommendation system can recommend music that other users with similar taste liked. Collaborative querying reduces the time taken by users to formulate queries by recommending similar queries carried out by other users (Fu et al 2005b). Another interesting dimension in Recommender Systems is that of trust (O'Donovan & Smyth 2005). It may not be enough to just look at users with similar preferences but also look at the previous ratings history of each of these users and highlight those users that have been reliable in their ratings who thereby gain higher trust and thus higher weighting in the recommendation process. The collaborative approach can also be used for web page ranking and classification. Lifantsev (1998) (as cited by Huang 2000) describe the OpenGrid project which proposes to extend the HTML standard and allow users to submit their opinions about the pages they browse. These opinions are collected regularly and are used to influence page ranking and classification.

Early personalised systems had their limitations such as they were based on local context of a single user profile and required a high level of input from the user in order to mark relevant pages and delete obsolete queries (Bharat 2000).

The more recent approaches such as Collaborative filtering and Collaborative querying both aim to get the best out of sharing information between the user community and extracting the relevant parts to make specific recommendations to a user.

### 2.2 Link Analysis

Link analysis is another approach to modelling the context of an information space which could be a web page and its different types of hyperlinks to companies, homepages and user-groups. The content and structure of the information space surrounding documents is an important contextualising factor, which can be used to improve the accuracy of relevance rankings assigned to documents (Lowe 2000). Web pages can be amended by adding multi-destination links to them (El-Beltagy et al 2001). The user interests define which links are to be exported for that user and a document's content defines which of these links are to be rendered in that document. Linkages among documents may indicate the importance of documents on the basis that important documents are pointed to by many documents (Yu et al 2001).

### 2.3 Language Models

Language models can be used to represent context and support context-based techniques such as relevance feedback and query disambiguation ( Croft et al 2001). These are generative models of documents and queries which serve as a formal framework for describing a range of retrieval processes such as query expansion and cross lingual retrieval. Each document is represented using a document language model. Language models are based on statistical language modelling (SLM) such that a language model is a probability distribution that captures statistical regularities of natural language use. In other words, how likely the ith word in a sequence would occur given the identities of the preceding i-1 words. So in a collection of computing documents the probability of ‘computer’ following ‘apple’ is more probable than ‘pie’. The query is treated as sample of text from a language model. The query is assessed for ambiguity and the level of ambiguity is quantified. A clarity value (non-negative number) is assigned to the query based on how different its associated language model is from the corpus language model. If the query is ambiguous then probable contexts are identified. Sentences that are representative of each context (ie the sentences have high probabilities of generation in those contexts) are shown to the user for clarification. Documents are ranked according to the probability that the document language model could generate the query text. The language model requires a large collection of training data for translation probabilities and must rank documents efficiently. Passage level language model is beneficial for documents that are long or span several subjects (Liu & Croft 2002).

### 2.4 Ubiquitous Computing

Context is also dependent on the physical environment. In mobile computing environments the change in the physical environment implies that the context of user information needs will also change. Information retrieval is proactive, it looks for information which is likely to be of interest by making use of context data (Jones & Brown 2003).

### 2.5 User background

The assumption that for domain-specific search engines the context is implied by the chosen database which contains domain specific information might be an over generalisation because firstly the problem with this is that ambiguity might still exist within a domain (Krovetz & Croft 1992) and secondly, even within a particular domain users will differ in their information needs and information seeking behaviour. Differences in user levels can be addressed through the user interface design. Novice users need more support from the user-interface for navigating and querying than expert users (Torrissen 1998). Future knowledge resources will have these properties built in so that a user can select a knowledge resource that is relevant to his needs before carrying out a search.

### 2.6 Task-related context

Tasks can be categorised into work-tasks and social tasks (Ingwersen 2004). The former are related to the subject domain the user is working on and also the subject knowledge the has. Bauer & Leake (2001) state that the success of personal information agents (such as Calvin) depends on their ability to provide task-relevant information. Access profiles are built which identify terms occurring frequently in sequences of document accesses and which are expected to be useful for distinguishing sets of documents related to the same task. Similarly, it might be beneficial to have different user interfaces to suit different tasks (Toms et al 2004) and different retrieval strategies to suit different query classifications.

### 2.7 Context of Knowledge Resources

Knowlede resources are assumed to be context-free however Ruthven (2004) points out that whilst the knowledge resources are useful in deriving context, the knowledge resources themselves have a certain context attached based on how they were created and why they were created. To use these knowledge resources effectively, it is useful to understand the “built-in” context otherwise we will not be using the resources to their full potential. For example a mathematics knowledge resource that has been created by a school teacher will probably include information relevant to teaching mathematics at school level and is unlikely to include information that would be required by post-graduate students.

In this paper we are interested in the use of context for query expansion and we use the linguistic definition of context and the use of an ontology to provide our view of context.

This view is similar to collaborative filtering in the sense both use a community of users. An ontology is a conceptual view which is shared by a set of users whereas collaborative filtering is a view of searches shared by a user community.

## 3 MOTIVATION

It is clear that context is a growing area which researchers are looking to for providing improvements in query retrieval. Context in information retrieval covers a wide range of topics of which query expansion is one. It is difficult to express an information need using exact query terms. Query expansion allows the user to carry out searches on morphological variations of the original term(s) and/or include any other terms which result from word sense disambiguation. Various approaches exist for conducting query expansion. The most recent of these is to use ontologies to infer context for ambiguous queries. The concepts in the ontology can be used for word sense disambiguation and subsequent query expansion. Query expansion has been successful to a certain extent but there is still scope to improve the techniques, interfaces or algorithms used to infer context more accurately in order to improve the results even further. Issues to address in query expansion include designing algorithms for optimum parameter choice; techniques to identify when to expand; and dealing with document collections which do not have a controlled vocabulary and are not consistently written such as web pages.

The following section defines what is meant by query expansion and summarises the various information retrieval approaches to handling context within query expansion.

## 4. WHAT IS QUERY EXPANSION?

Much work has been done in the area of query expansion. Query expansion is needed due to the ambiguity of natural language and also the difficulty in using a single term to represent an information concept. Krovetz and Croft (1992) observed that most benefit is achieved with high-recall searches that depend on matches of single concepts.

With query expansion, the user is guided to formulate queries which enable useful results to be obtained. The main aim of query expansion (also known as query augmentation) is to add new meaningful terms to the initial query. This process of adding terms can either be manual, automatic or user-assisted. Manual query expansion relies on user expertise to make decisions on which terms to include in the new query. In the case of automatic query expansion, weightings are calculated for all terms and the terms which have the highest weighting are added to the initial query. Different weighting functions produce different results, therefore retrieval performance depends on how the weightings have been calculated. With user-assisted query expansion, the system generates possible query expansion terms and the user selects which of these to include.

The new terms resulting from the chosen term selection method should provide contextual information for the initial query with a view to improving the retrieval results. The contextual information can be acquired from relevance feedback, term co-occurrence and more recently it has been derived from knowledge models such as ontologies. The

following sections explain how each approach works and includes some relevant sample case studies.

## 4.1 QUERY EXPANSION USING RELEVANCE FEEDBACK

Relevance feedback is a fairly established technique for modification of the initial query using words from top-ranked or identified relevant documents (Salton & McGill 1983). It is an easier way of improving the retrieved document set as opposed to the user having to construct a new query.

The relevance feedback loop requires the user to enter an initial query which results in a display of ranked documents (usually titles/abstracts). From this display, the user makes relevance judgements and selects the relevant documents. The relevant terms from these documents are added to the initial query. An alternative to this is pseudo relevance feedback where the top ranked n documents are assumed to be relevant. Terms from these documents are selected and used for expanding the query. Whether pseudo-relevance feedback is used or traditional relevance feedback, the term selection method is a key factor in the performance of expanded queries. It needs to consider how to weight the new terms, whether to exclude the original query terms, whether to include all of the new terms or just some of them and if so how many new terms to include. These issues are described below.

### 4.1.1 SELECTION OF NEW TERMS

This point examines how to combine the initial query terms and the expanded terms. It is possible that the query expansion process generates such a large number of candidate terms that it might not be practical to use all of these terms. Some research has been carried out on the optimum number of terms to include and there are differing viewpoints ranging from a one-third of the terms as suggested by Robertson and Willett (1993), 20 terms (Harman 1992) to massive query expansion (Buckley et al 1995). In the latter, 300-530 terms were added to the original query. The terms came from known relevant documents / top retrieved documents and seemed to improve effectiveness from 7%-25%. Buckley concluded that massive query expansion is effective for routing ie moving the query vector towards the centroid of relevant documents and away from non-relevant documents. A contrasting view provided by Sihvonen and Vakkari (2004) is that the number of terms used for query expansion is less important than the type and quality of the terms chosen. Billerbeck and Zobel (2004) systematically studied the effect of the number of query expansion terms on performance. They concluded that one size does not fit all because the optimal number varies from query to query.

Collaborative learning is a more recent term selection approach whereby term concepts learned by other queries can be used for query expansion (Klink 2002). The relevant document set is an accumulation of the relevant document sets for all queries which contain a term from the original query. Terms which are most similar to the concept of individual query terms are selected as opposed to selecting terms that are similar to the entire query. Results of this approach look positive.

### 4.1.2 WEIGHTING OF NEW TERMS

In answer to the question whether all terms in the expanded query should have equal weighting or whether the new terms should have a higher/lower weighting, Voorhees (1994) found that assigning lower weights to added concepts enhances retrieval accuracy. She used a factor between 0 and 1 for weighting added terms. The original f4.5 formula (Sparck-Jones & Robertson 1976) was modified and renamed f4modified (Robertson 1986). The f4modified formula took into account the addition of new terms to the original query and achieved the desired effect of bringing terms with low frequency to the top of the ranked list. In 1990, Robertson developed a new ranking algorithm which was based on the Probability Ranking principle (Robertson 1990). This principle requires for each document an initial relevance judgement from which the probability of a document being relevant to a query can be calculated. According to this principle, the retrieved documents are presented in decreasing probability of relevance to the user who submitted the query.

The inclusion of term t in the search formulation with weight $w_t$ will increase the effectiveness of retrieval by

$$WPQ_t = w_t(p_t - q_t)$$

where, $w_t$ is a weighting function, which in this case is the F4.5 formula; $p_t$ is the probability of term t occurring in a relevant document; and $q_t$ is the probability of a term t occurring in a non-relevant document.

$$WPQ_t = \log \frac{(r + .5)(N - n - R + r + .5)}{(n - r + .5)(R - r + .5)} \cdot \left( \frac{r}{R} - \frac{n - r}{N - R} \right)$$

where N is the total number of documents in the collection; R is the sample of relevant documents as defined by the user's feedback; n is the number of documents indexed by term t; r is the number of relevant documents (from the sample R) assigned to term t.

Rocchio's method (1971) used an information retrieval system based on the vector-space model whereby documents are represented as vectors in the information space and so are queries. The vectors are weighted, with higher weightings being given to relevant vectors. After relevance feedback the weights are adjusted and only those new terms that occur in the initial query or occur in at least half of the relevant documents are added to the original query. This produced positive results.

An extension of this approach is based on predictive algorithms for collaborative filtering (Hoashi et al 2001). The query-document similarity is used for calculating term weights during query expansion. The user database of votes is viewed as a set of vectors where each vector expresses a user and the elements of each vector express the votes of the active user. The main aim of the collaborative filtering process is to predict the values of

empty elements of the active user's vector. Such predictive algorithms are showing signs of effectiveness.

### 4.1.3 SAMPLE SIZE OF RELEVANT DOCUMENTS

Efthimiadis (1996) describes the various approaches adopted by different researchers with respect to sample size. Okapi uses a sample size of 3 documents, some recommend a sample size of 5 documents and others use a larger sample size. There is no clear effect of the sample size of relevant documents on the search performance and the area of selecting the optimal sample size is still an open area of IR research. In the absence of relevance judgements, a common alternative is to have a cut-off point in the form of 'top X documents' which are all treated as relevant. The sample size of relevant documents can be selected from the documents retrieved prior to the cut-off point.

## 4.2 CASE STUDIES OF QUERY EXPANSION USING RELEVANCE FEEDBACK

Harman (1988) used the Cranfield test collection to discover techniques to select candidate terms for effective query expansion. Non-common words are extracted from the top ten relevant documents. The terms are sorted using a statistical technique. The top 20 terms from the sorted list are used to expand the query. The results showed a significant improvement. Another approach is to take ambiguous words and derive the context based on each word sense. An ambiguous word can have several different meanings. Word sense disambiguation helps us to select the meaning which makes the most sense in a given context. Sanderson (1994) describes various techniques used for Word sense disambiguation. Query expansion techniques can subsequently be used to add the words from the chosen context to the original query.

Tombros and Sanderson (1998) compared the use of normal summaries with query biased summaries and found that the latter improve accuracy and speed of user relevance judgements. A query biased summary consists of sentences that contain the original query terms. The underlying assumption is that if users can see the original query terms in the summaries they are in a better position of judging more accurately whether the document is relevant or not. However, from the experiments carried out in 2001, Tombros et al found that good quality expansion terms can only be generated if the original document collection contains a large number of relevant documents. If the set of relevant documents is small then the choice of query expansion terms can be poor. An additional factor which affects the quality of expansion terms produced is that the results will vary depending on which summarization technique is used. An ideal solution should be able to produce good results and not be dependent on the document collection or the choice of summarization technique.

Global techniques and local techniques can be used for automatic query expansion. A comparison of the two techniques was carried out by Xu and Croft (2000). The problem with global techniques is that of performance because of the time taken to analyze the whole collection to discover word relationships. The local techniques focus on the most

frequently occurring terms in the top ranked documents. Xu and Croft found that local techniques only work if all of the top-ranked documents are relevant. The performance improvement by local context analysis over local feedback is due to a better metric for selecting expansion terms.

Vakkari et al (2003) investigated the expression of search facets and relevance feedback and compared its success in interactive query expansion with automatic query expansion. They found that interactive query expansion is better if all retrieved relevant items were counted otherwise it made no difference if only those items recognised relevant by users were observed.

Croft, Cronen-Townsend and Lavrenko (2001) state that relevance feedback is not very successful in search engines with the closest feature being "find more like this". The main problem is trying to get users to provide relevance information. Simply indicating 'Relevant/Not relevant' does not give users enough incentive. Instead user feedback can be of different types (Spink 1997). Users can give feedback on the content of the retrieved documents; state which of the relevance feedback terms are relevant; give feedback on the magnitude of the retrieved set as to whether it is too large, too small or just right; and give feedback on terminology by judging the relevance of the terms in the inverted file. Robertson (1990) also found that when a user is asked to make a relevance judgement this is slightly ambiguous. For a given relevance judgement, it is important to distinguish between whether it is relevant to the topic and/or relevant to the user. Some research has been done on passive measures such as time spent browsing a page or number of links followed from a page. If the user can indicate relevant sections or even phrases in a document, relevance feedback is more accurate. This implies we need more feedback from users not less because term suggestion increases the recall, while interaction enables the user to increase precision. Successful application of relevance feedback involves sophisticated interface design and good algorithms for inferring context.

Ruthven & Lalmas (2003) state that domain specific collections perform better with relevance feedback than domain independent collections because it is easier to select good expansion terms from this type of a collection or because the ambiguity of search terms is less significant.

In conclusion, effectiveness of query expansion using relevance feedback can vary depending on many factors such as choice of parameters in the term weighting process, number of relevant documents in the document collection, facilities provided for users to give good quality relevance feedback with ease and finally whether the collection is domain specific or domain independent.

## 5. QUERY EXPANSION USING CORPUS DEPENDENT KNOWLEDGE MODELS

Previous research concentrated on obtaining context from the document collection using techniques such as stemming, clustering and term co-occurrence. Stemming is a process whereby all variations of term are generated by adding/removing prefixes and suffixes as appropriate. During the late 60s and early 70s, query expansion using term clustering was investigated by many researchers. Similar documents are placed in a cluster. It is assumed that similar documents are relevant to the same requests. On this basis, if the query terms mapped onto one or more clusters the terms from the cluster(s) would be used for expanding the query. The usefulness of this technique was found to be marginal due to poor clusters resulting from small document collections (Lesk 1969) or insufficient differences in vocabulary between relevant and non-relevant documents (Sparck-Jones 1973). Another problem with clustering is that it works on the assumption that a term can only belong to one cluster. For ambiguous terms this may not be the case. Deerwester et al (1990) also recognised the difficulty in representing document clusters and term clusters together at the same time. Term co-occurrence has also been an area of investigation to assist in inferring context.

## 5.1 PHRASES BASED ON TERM CO-OCCURRENCE

Term co-occurrence refers to two or more terms that are situated next/near to each each other in the source document. Smeaton and Van Rijsbergen (1983) ran experiments which used new terms generated from sources such as maximum spanning trees and found very little improvement. Peat and Willett's explanation of this is that similar terms have comparable frequencies, query terms have high collection frequencies and as a result so will the candidate terms (Peat & Willett 1991). Since high frequency terms do not discriminate between relevant and non relevant documents the addition of these terms for query expansion is ineffective.

Schatz et al (1996) used two complementary term suggestors, subject thesauri and co-occurrence lists. The thesauri are generated by human indexers who based on their subject knowledge decide where in a subject hierarchy a term should be placed. The co-occurrence lists are computer generated and terms are placed in frequency order of co-occurrence. The overall conclusion was that multiple views results in a better quality search. This is supported by Mandala, Takenobou and Tanaka (1999) who also suggest that it is better to use a combination of query expansion techniques than to use a single technique. Similar findings were made by Huang et al 2005 as a result of conducting experiments which combined conceptual indexing and keyword indexing.

Ogawa et al (2000) describe a Japanese text retrieval system to which an English tokenizer has been added to process English text. A stemmer is used to process the natural language in the documents and produce word tokens for indexing. The indexing system stores term positions, to provide proximity based search to enable the user to specify the distance between query terms. The query is constructed by parsing each topic and extracting words into appropriate fields. The constructed query is fed into the ranking system. The ranking system uses a modified version of Okapi to assign term weights to query terms. It scores each document and returns the set of top-ranking documents. Query expansion is based on feedback from the pseudo relevant documents. The words

from the pseudo relevant documents are collected and ranked. The highest ranking words are added to the query, and new term weights are reassigned. The experiments contained 8 runs on TREC-7 and TREC-8: Title only versus title and description; phrasal terms versus non phrasal; expansion and non expansion. For the TREC-7 topics, there was a 28% improvement on precision in the 'title and description' run when phrasal terms and queries are expanded. If the stemmer over-stems the words this can have an a negative impact on overall performance. Although results of preliminary experiments have been stated as the reason for deciding to give phrasal terms a lower weighting than single terms, further details explaining this strategy would be beneficial. Certain words which would not benefit from expansion such as common words or those that are too specific are excluded. This process should be accurate to avoid excluding words that might be relevant. The choice of parameters used in the modified Okapi formula also affects the performance of the retrieval system.

Chu et al (2002) present a knowledge-based query expansion technique to improve document retrieval effectiveness. The level of relevancy of a specific term in the resulting query is determined by its co-occurrence with the general concept term, which can be mined from the corpus. The general concept terms in a query are substituted by a set of specific concept terms used in the corpus that co-occur with the key query concept. Since the expanded query matches with the document index terms much better, experimental results reveal that such query expansion produces better retrieval effectiveness than the unexpanded ones. The downside of this approach is that specific terms reduce recall. This approach is only suited to situations when precision is of more importance than recall.

Vechtomova et al (2003) present two novel approaches to query expansion with long-span collocates. This is where words significantly co-occur in topic-size windows with query terms. The first approach is global collocation analysis where collocates of query terms are extracted from the entire collection. The second approach, local collocation analysis extracts terms from a subset of retrieved documents. The experiments showed that global collocation analysis performed worse than unexpanded queries. This may be due to the fact that terms extracted from the global collection are too general and they need to have a more specific context. The local collocation experiments produced better results. However if the top retrieved documents are not rich enough they may exclude documents which are still relevant to the query.

Another property which affects the complexity of the query expansion algorithm is whether the phrase is compositional or non-compositional. With compositional phrases each term in the phrase can be expanded using substitute terms and the final expanded phrase will retain its meaning. Cui et al (2003) identify phrases using n-grams from query logs. They filter out those phrases that do not appear in the documents. The general findings are that short phrases are a more accurate representation of information requirements. An example of a compositional phrase is 'search engine'. Also phrases have a higher inverse-document frequency because terms are more common in document collections than phrases. Eguchi (2005) states phrases together with pseudo relevance feedback work better than phrase expansion alone.

Sometimes referred to as idiomatic phrases, non-compositional phrases are phrases which take on meanings that go beyond the meanings of their parts. Lin (1999) defines a non-compositional phrase as a phrase where 'its mutual information differs significantly from the mutual information of phrases obtained by substituting one of the word in the phrase with a similar word'. For example, the phrase 'red tape' can be used to refer to bureaucracy. If we replace the colour red with another colour eg 'yellow tape', this does not achieve the same meaning as the original phrase.

## 5.2 CONCEPT NODE STRUCTURES

Lexical networks are another important source for deriving context. They contain domain-specific vocabularies and relationships which have been automatically extracted from the collection. Text analysis tools can be used to develop the vocabulary for the lexical network. Lexical networks play an important role in the navigation process. Lexical relationships between terms are utilised to suggest additional terms. Some systems indicate the strength of the relationship between two terms. The general tendency is to disambiguate terms during the search process and not store the disambiguated terms prior to the search.

In the early sixties, word sense was frozen into the lexicon. However in the early nineties, researchers discovered that full lexical knowledge comes from the texts themselves. Pustejovsky (1995) proposed the use of a generative lexicon to disambiguate word sense. A generative lexicon uses machine-readable dictionaries and large text corpora to construct a core lexical engine which acquires new lexical entries and refines existing ones through statistically based corpus acquisition methods. Coates-Stephens (1991) established an algorithm for obtaining the meaning of proper nouns from the text.

Callan et al (1992) describe a retrieval system (INQUERY) that provides support for complex query formulation. The system is based on a type of probabilistic retrieval model called the inference net. The inference net is similar to a bayes net. The inference net has two component networks, one for the document collection and one for each query. Probability can be represented by weights on arcs. Query nodes are linked to concept nodes and these in turn are linked to the document nodes. The concept-document node link can be true or false depending on whether the concept is observed in the document. Therefore relevant documents can be retrieved by following the arcs from the concept nodes which have a true value to the associated document nodes. INQUERY accepts the root node of the query net and evaluates it, returning a single node which contains a belief list. A belief list is a structure containing documents and their corresponding beliefs of meeting information needs. Proposed extensions to the INQUERY retrieval system were to incorporate morphological processing and the ability to handle phrases.

Concept hierarchies can also be used to carry out query expansion techniques. (Sanderson 2004). With these, instead of string based searches, concept based searches are carried out which generally produce more useful results than the former (Jarvelin et al 2001). The concept hierarchies are automatically generated from the document collection by

extracting salient words from the top-ranked documents and organizing them hierarchically using a subsumption function to determine not just the fact that two concepts are related but also how they are related ie if one concept is said to subsume another concept then the former should be placed as the parent in the ontology hierarchy (Sanderson & Lawrie 2000). Terms that co-occur are grouped together. Similar term clustering is used by Alta Vista in its Refine option. The inverse document frequency measure was used to indicate whether a term was general or specific. In the concept hierarchy, a child can have more than one parent (Directed Acyclic graph DAG). Ambiguous terms would have a separate entry in the hierarchy, one for each word sense. Experiments that were carried out were designed to test the effectiveness of the hierarcy. They indicate that a hierarchy is a more efficient means of locating relevant documents than scanning a ranked list. For static collections this is a reasonable approach to use otherwise the concept hierarchies have to be generated repeatedly, which is a time consuming activity. Another issue to be resolved in relation to this is whether to archive previous versions of concept hierarchies and their document collections in case the user wants to return to older queries.

Query expansion using corpus dependent knowledge models is more suitable for static document collections. For web collections, the knowledge models would have to be constantly updated/regenerated because the collections on the web are more dynamic/fluid in nature. If phrases offer more selectivity due to higher inverse document frequency then, the area of phrase based query expansion has further scope for investigation.

## 6. QUERY EXPANSION USING CORPUS INDEPENDENT KNOWLEDGE MODELS

The problem with traditional relevance feedback techniques and corpus dependent query expansion is that they are content driven. The corpus content is analysed to extract candidate terms for query expansion. This can only work if there are sufficient relevant documents to work with and also that these documents contain a reasonable set of terms that represent the subject area for the query. Corpus independent knowledge models do not suffer from this drawback.

Corpus independent knowledge models can be in the form of a thesaurus or an ontology although according to some, the distinction between the two is blurred. Ontologies provide consistent vocabularies and world representations necessary for clear communication within knowledge domains (Leroy et al 2000). Bates (2002) suggests an ontology is more accurately described as “a classification, thesaurus or a set of concept clusters”. Another definition of ontologies is ‘classifications, lists of indexing terms, or concept term clusters’ (Communications of the ACM, 2002). An ontology is a model of reality, it is not reality itself. Bates (2002) suggests that instead of calling it an "ontology," the system should be labelled using a more accurate description– a classification, thesaurus or a set of concept clusters(see also Soergel, 1999). Ontologies improve the accuracy in fuzzy information search and facilitate mono- and multi-lingual human-computer dialogues by paraphrasing the query of the user through context

identification and disambiguation (Leger et al 2001). Gruber (1993) defines an ontology to be a 'specification of a conceptualisation'. and explains that ontologies were first used in philosophy then Artificial Intelligence. An ontology is a model of reality, it is not reality itself. In a natural language, a word may have multiple meanings depending on the applicable context. The purpose of an ontology is to provide a context for the vocabulary it contains. In a computer system, context may be represented and constrained by an ontology. Therefore, an ontological model can effectively disambiguate meanings of words from free text sentences.

Ontologies range from general to domain-specific. WordNet, EuroWordNet and Cyc are examples of a general ontology. Many domain-specific ontologies exist for example in the medical and legal domains. (Buckland 2003) distinguishes between three different kinds of ontology. Axiomatic ontologies contain abstract concepts that facilitate reasoning. Terminological Ontologies composed of structures of lexicalized concepts and Domain ontologies which represent the knowledge organization systems used for documentary purposes such as Dewey Decimal Classification system which uses a numerical notation to denote an unlimited number of topics and has an English language index to the classification numbers.

Advantages offered by ontologies are that they are readily available(Bateman 2005); they routinely include proper nouns: personal names and place names; many software tools exist to automate the creation and support the evolution of ontologies; finally most ontologies have been defined in a portable language such as XML (eXtensible Markup Language). Also the extensible features of XML can be used to capture changes in the evolving ontologies.

However, ontologies are not without their problems. The first issue is related to vocabulary mismatch between the query terms and the concepts in the ontology. A mapping process needs to take place to overcome this problem. Secondly, if an ontology for a particular domain does not exist then a lot of effort is required to construct ontologies from scratch not just from a technical point of view but more importantly the process of knowledge extraction from domain experts and arriving at a consensus view. The design and construction of domain ontologies is labour intensive, time consuming and difficult (Kashyap 2001). Kashyap proposes a methodology for creating an ontology with minimal involvement of the domain expert by reusing readily available information such as schemas, queries,data dictionaries and thesauri. Hwang (1999) proposed one method for automatic generation of ontology started from the seed-words suggested by domain experts. This system collected relevant documents from the Web, extracted phrases containing seed-words, generated corresponding concept terms and located them in the 'right' place of the ontology. Several kinds of relations are extracted: is-a, part-of, manufactured-by or owned-by etc. It also collects "context lines" for each concept generated, showing how the concept was used in the text, as well as frequency and co-occurrence statistics for word association discovery and data mining. The drawback is that it fully depends on the seedwords provided by the domain experts. (Lame 2003) presents a method to identify ontologies' components. The method relies on text analysis to extract concepts and relations among these concepts. This method is applied in the

legal field to build an ontology dedicated to information retrieval. Legal texts on which the method is performed are carefully chosen as describing and conceptualizing the legal field. It is assumed that the method can help legal ontology designers and may be used while building ontologies dedicated to tasks other than information retrieval. (Saias 2002) states that web legal information retrieval systems need the capability to reason with the knowledge modelled by legal ontologies. Using this knowledge it is possible to represent and to make inferences about the semantic content of legal documents.

## 6.1 **USING ONTOLOGIES FOR INFORMATION RETRIEVAL TASKS**

Ontologies are applicable to domain independent retrieval such as web information retrieval and even more useful in specialized information retrieval tasks. Three common classifications of queries are navigational, information and transactional (Broder 2002). Kang and Kim (2003) also produce similar query classifications. Navigational queries typically are used to find homepages. Information queries are used to find information relevant to a given topic. Transactional queries are just that, they allow a user to locate a website offering a service such as shopping and enable the user to carry out a purchase transaction. According to Broder, the best query type for expansion is the information query. The reason for this might be that the information query is less specific than the other query types. Various information retrieval strategies exist (Bhavnani et al 2001). Locating is a strategy used to find specific types of information. Sampling is a strategy which selects a web directory or subject tree to find few sources of high quality information. Deep searching is a strategy which makes use of search features such as Boolean operators, truncation and proximity operators to perform an effective search. Bhavnani, Drabenstott & Radev define a framework for IR tasks and strategies and propose that the choice of retrieval strategy depends on task type.

Ontologies have been used for a range of information retrieval tasks such as thematic summarization; word sense disambiguation; indexing; text classification; image retrieval cross language information retrieval, and for query expansion. In this paper the focus is on query expansion (as detailed in section 6.2 onwards).

### 6.1.1 Thematic summarization

Loukachevitch and Dobrov (2000) describe the technique of construction of a structural thematic summary (STS) for a Russian-English Thesaurus on Sociopolitical life. An STS represents text content by indicating main theme and sub-themes. Each term in the text is linked to a thematic node. The suggestion is that summaries help users to determine relevance of text to users information needs. STS involves thematic analysis of texts including term disambiguation and analyses cohesion relations.With a conventional thesaurus, humans use their domain, common sense and grammatical knowledge to index documents. So conventional thesauri are hard to use in an automatic indexing environment (Salton 1989). The Sociopolitical thesaurus differs from linguistic resources such as WordNet and EuroWordNet. The Sociopolitical thesaurus only includes specific domain words not terms used in common language; ambiguous terms are clustered into a single concept (Pustejovsky 1995); encyclopaedic information is included for terms such

as possible situations, reasons, results, participants and properties. Cohesion involves relations between words that connect different parts of the text eg lexical cohesion is expressed by repetitions, synonyms and hyponyms or by words connected with other semantic relations such as whole-part, situation-participant and object-property. Cohesion relations connect sentences of a text and therefore connect the main theme with sub-themes. A lexical chain is a chain of words in which the criterion for inclusion of a word is some kind of cohesive relationship to a word that is already on the chain (Morris & Hirst 1991). Barzilay and Elhadad (1997) construct lexical chains based on WordNet relations. The authors claim that WordNet was not created as a tool for automatic text processing because it lacks necessary information such as relations between different parts of speech and conceptual relations such as situation-participant, situations-domain or object-property.

### 6.1.2 Word sense disambiguation

Gonzalo et al (1998) use a manually disambiguated test collection of queried and documents derived from the SEMCOR semantic concordance. Their experiment covers three types of index spaces: original terms; word senses derived from manual disambiguation and finally WordNet synsets. The authors observe that if queries are not disambiguated, indexing by synsets performs only as good as standard word indexing.

Stairmand (1997) describes QUESCOT, a program which analyses and quantifies textual contexts in documents with reference to the WordNet database and ascertains the dominance of topics in a document. Results suggest QUESCOT is not suitable for word sense disambiguation and text segmentation, but more promising as an indexing component. QUESCOT is used to generate a resource (based on collocation relations between concepts in a document collection) to supplement WordNet. A concept is based on WordNet synsets. It is recognised that WordNet has some limitations such as it doesn't deal with collocation information and proper names. Stairmand suggests incorporating the work of Coates-Stephens (1991) as a preliminary stage. Coates-Stephens derives the meanings of proper nouns from the document text itself. This view concurs with Pustejovsky (1995) who states that word sense is not frozen into the lexicon but full lexical knowledge comes from the texts themselves.

### 6.1.3 Indexing

Davies and Weeks (2004) describe an information-seeking system which combines traditional keyword querying of WWW resources with the ability to browse and query against RDF annotations of those resources. RDF schemas and RDF are used to specify and populate an ontology and the resultant RDF annotations are then indexed along with the full text of the annotated resources. The resultant index allows both keyword querying against the full text of the document and the literal values occurring in the RDF annotations, along with the ability to browse and query the ontology. This approach is seen to be a key enabler for fully exploiting the Semantic Web in the area of knowledge management and argue that the ability to combine searching and browsing behaviours more fully supports a typical information-seeking task.

Khan and Wang (2002) index images according to meanings rather than objects that appear in images. To convert objects to meanings a meaning based index structure has been created through the design and implementation of a concept-based model using domain dependent ontologies. In order to make this approach scalable, a mechanism that can generate ontologies automatically is proposed. This is based on clustering and a vector space model. To calculate the object similarity measure, colour and shape are considered together. The system poses the ability to automatically select concepts using a disambiguation algorithm that prune irrelevant concepts and allows relevant ones to be associated with images. The system uses a neural network to successfully identify objects present in images. Once identified by the neural network, the objects are fed into the domain-dependent ontologies for high precision classification of the image based on its contents.

### 6.1.4 Text classification

Scott and Matwin (1998) describe experiments in machine learning for text classification using a new representation of text based on WordNet hypernyms. Performance of discrimination rules based on hypernym density representation is compared with rules produced using the 'bag-of-word' representation. The hypernym density representation leads to more accurate and comprehensible rules. Future work aims to incorporate more of the relations from WordNet such as meronymy to give even more semantic character.

### 6.1.5 Query formulation

Suomela and Kekalaninen (2005) look at 16 real users' use of the CIRI (Concept based Information Retrieval Interface) ontology to help formulate queries compared with queries formulated without the ontology. Search results were evaluated with generalised precision and recall in addition to personal assessments of precision. The CIRI query expansion feature resulted in higher number of search terms compared to direct interface queries nevertheless direct interface queries performed better. The experiments rely on the user being familiar with the ontology and being able to navigate the ontology with a good level of confidence. Also the initial query formulation starts within the ontology, so there is an increased likelihood of the user losing their sense of direction or being distracted by the different number of paths during the navigation process and the time taken to traverse those paths. The paper states that eleven of the users found the CIRI interface more complex to use. All queries are expanded using the ontology regardless of whether they are ambiguous or not.

Hyvonen (2003) considers the situation when a user is faced with an image repository whose content is complicated and semantically unknown. This work shows how ontologies can then be of help to the user in formulating the information need, the query, and the answers. This problem is approached through a case study based on an image collection on promotion ceremonies of the Helsinki University Museum.

### 6.1.6 Cross Lingual Information Retrieval (CLIR)

CLIR is when the user wishes to find a document written in a different language to the one used in the query. Multilingual resources such as Euronet are used by (Pavel et al 2003). Finally, the use of ontology aims also to solve the cross-language problem as there is not always 1-to-1 correspondence of terms between different cultures. The solution of mapping terms from one language to another should be as natural as possible to the user (Leger et al 2001). Grover et al (2003) combine the use of an ontology for cross lingual information retrieval with a personalisation agent in order to present results according to user preferences.

### 6.1.7 Concept mapping

(Leroy et al 2000) describe the development and testing of a Medical Concept Mapper as an aid to providing synonyms and semantically related concepts to improve searching. All terms are related to the user-query and fit into the query context. The system is unique because its five components combine human-created and computer-generated elements. The Arizona Noun Phraser extracts phrases from natural language user queries. WordNet and the UMLS (Unified Medical Language System) Metathesaurus are used to provide synonyms. The Arizona Concept Space generates conceptually related terms. Semantic relationships between queries and concepts are established using the UMLS Semantic Net.

## 6.2 COMPARISON OF ONTOLOGY-BASED QUERY EXPANSION TECHNIQUES WITH RELEVANCE FEEDBACK TECHNIQUES

When we compare ontology based query expansion techniques with those that do not rely on knowledge bases there are some similarities. Both sets of techniques still need to acquire feedback whether it be manual or automatic. The difference is that the ontology based query expansion techniques will obtain term suggestions from the knowledge model whereas relevance feedback techniques rely on having a reasonable set of relevant documents from which to suggest suitable terms. This is a point worth making because much debate has taken place on selecting the optimal sample size of relevant documents in relevance feedback techniques. With the ontology-based query expansion approach, sample size is not needed. Another similarity between the two sets of approaches is that the newly suggested terms still need to have weightings attached for the ranking algorithm. Finally the recent trend in collaborative filtering also bears similarity with ontology based query expansion in the sense that the former is based on getting information from a community of users who share the same interest as the searcher and an ontology is usually a “collective” representation of a domain which has been derived from a community of domain-specialist users.

## 6.3 CASE STUDIES OF QUERY EXPANSION USING WORDNET

Ontologies have also been used to aid query expansion since the early nineties with mixed success. WordNet has been a popular general ontology used in the area of query expansion as the following works show. (Gonzalo et al 1998) use a manually disambiguated test collection of queries and documents derived from the SEMCOR semantic concordance. Their experiment covers three types of index spaces: original

terms; word senses derived from manual disambiguation and finally WordNet synsets. The authors observe that if queries are not disambiguated, indexing by synsets performs only as good as standard word indexing. According to Gonzalo, indexing with word sense improves information retrieval by more than 29%.

Voorhees (1993) carried out experiments to exploit the semantics contained within WordNet to improve retrieval effectiveness by indexing with word senses instead of word stems. The results showed that the effectiveness of the vectors produced by this disambiguation technique was worse than word stem vectors for all five collections. The findings indicate that short query statements can be difficult to disambiguate because the IS-A hierarchy is not sufficient to reliably select the correct sense of the noun. Incorrect sense resolution or query expansion using automatically generated synsets resulted in missing correct matches which in turn have a deteriorating effect on retrieval performance than using spurious matches. Voorhees (1994) used WordNet to conduct experiments on small single domain TREC collections. The results suggest that query expansion can improve problems of mismatched vocabularies especially in cases where the terms that are expanded are lexically related to the query terms. In contrast, query expansion makes little difference in retrieval effectiveness for long queries because they usually contain a full description of the information request. Voorhees found that some long queries could benefit from other techniques such as relevance feedback.

Hearst (1992) describes a pattern-matching method for the automatic acquisition of the hyponymy lexical relation from unrestricted text. For example “Bruises, wounds, broken bones or other injuries …” the method infers that each of the items on the list are hyponyms of the term injury. The hyponymy relation is also used to identify the general meaning of unfamiliar noun phrases eg hyponym ("broken bone", "injury"). The unfamiliar noun phrase “broken bone” is unlikely to appear in a dictionary but due to the hyponym relation, “broken bone” can be understood at some level as an injury without determining the correct senses of the component words. This approach avoids the use of pre-coded knowledge and can be applied across a wide range of text. According to Hearts, list items are usually similar so they are used to indicate synonyms and can be used for synonym expansion. Instances of the hyponymy relation ("is-a") that are found using this approach are compared with the relations in existing lexicons such as WordNet’s noun hierarchy. The results show a high level of similarity between Hearst’s approach and WordNet’s noun hierarchy.

Finkelstein et al (2002) describe Intellizap which is a context based search system. Query ambiguity is eliminated by deriving the context from the text surrounding the marked query in a given document. Keywords are extracted from the surrounding text to augment the query. This involves semantic keyword extraction and clustering to generate new queries. The expanded query is then submitted to various search engines. The results are then re-ranked. The system uses a semantic network for measuring the distances between words. It uses the vector space model to represent 27 domains. Linguistic information (such as hypernym and meronym) is obtained from the WordNet dictionary. The system combines the WordNet metric and the correlation metric to statistically

analyse the relationship between words. The results show that using context to guide a user's search process offers definite improvements.

Navigli and Velardi (2003) use sense information and ontologies for query expansion. They argue that expanding with synonyms and hyperonyms has a limited effect on web information retrieval performance. They suggest that other types of semantic information derivable from an ontology is more effective such as gloss words and common nodes. This is because words in the same semantic domain and same level of generality are best candidates for expansion. The ontology is used to extract the semantic domain of a word and then the query is expanded further using co-occurring words. Effectiveness of using ontologies to improve retrieved results depends on the type of task (ie subject finding vs. site retrieval) and query length. They concur with the view that query expansion is suitable for short queries. Their experiments used TREC 2001 web track, WordNet 1.6 for the ontology and Google. A semantic network is created for each word sense. The relevant semantic networks are then intersected pair-wise and a score is assigned to the configuration based on the number of common nodes where common nodes are those nodes in the semantic network which can be reached by both semantic network centres through directed paths. The experiment results showed a systematic improvement over the unexpanded query.

Baziz et al (2005) state ontology based information retrieval is promising in increasing the quality of responses since document semantics are captured. They used a small document collection and document content is represented using an optimum semantic network called document semantic core. WordNet concepts (which include words and phrases) are extracted and then globally disambiguated with reference to document terms to produce the optimum semantic network. Similarity measures between connected nodes weight the links. Four similarity measures are used, of which the first three are 'is-a' based and the fourth one is based on gloss overlaps. Scored concepts are used for document conceptual indexing. In automated word sense disambiguation, the challenge is to keep relevant concepts and discard irrelevant concepts.

Jones et al (1995) analysed the INSPEC thesaurus and used 8 relational database tables to store thesaurus information such as terms, equivalence, hierarchical, associative, components, classes, facets and words. Their experiments found no correspondence between the number of terms chosen and the query performance. Also the other main finding was that the number of terms selected is dependent on the distance from the original node. In other words there was a tendency for nodes with a short distance from the original node to be selected. Jones et al state that one of the main success factors in using a thesaurus for query expansion is to have a good match in the thesaurus to start with. For this reason terms should be expanded before carrying out the search of the document database to overcome the fact that there might not be an exact match of query terms with a WordNet node ( Jones 1993). Contrary to Efthimiadis (1996) who suggests that query expansion using knowledge models may be done without term re-weighting, Jones et al suggests that term weighting is a good idea. A reference is made to Shoval who bases the weightings on criteria such as number of connections; type of relationship; number of co-occurrences and path length (Shoval 1985). A further reference is made to

Chen who suggests that users should be able to adjust the term weightings (Chen et al 1993). Jones found that quality of the thesaurus is paramount. A thesaurus which has greater coverage, depth and accuracy has improved chances of producing better results.

Grootjen and van der Weide (2006) generate a local thesaurus by projecting a global thesaurus onto the top ranked documents resulting from an initial short two word query. A concept lattice is produced for the local thesaurus, however to solve the problem of finding a good starting concept, the user has to navigate the lattice. Even if the top ranked documents are not relevant they still cover topics related to the query. According to Mandala et al (1998), WordNet has two major weaknesses, namely the inadequate proper nouns and secondly the inability to form relations between terms that belong to different parts of speech because words in WordNet are organized on the basis of part-of speech. In other words we cannot form relationships between an adjective term and a noun term even though such a relationship might exist in reality. Also the meronym relationship is too narrow in its interpretation. To overcome this, Grootjen and van der Weider use a hybrid approach consisting of a global thesaurus and a dynamically created local thesaurus.

## 6.4 CASE STUDIES OF QUERY EXPANSION USING DOMAIN-SPECIFIC ONTOLOGIES

The problem with domain-independent ontologies such as WordNet is that because they have a broad coverage, ambiguous terms within the ontology can be problematic. For narrower search tasks, domain-specific ontologies are the preferred choice. A domain-specific ontology models terms and concepts which are specifically used in a given domain. Domain-specific ontologies have been constructed in many different application areas such as law, medicine, archaeology, agriculture, geography, multimedia, business, economics, history, and even the news domain to name but a few.

Fu et al (2005a) present query expansion techniques based on both a domain and a geographical ontology. In their work, a query is expanded by derivation of its geographical footprint. Spatial terms such as place names are modeled in the geographical ontology and non-spatial terms such as ‘near’ are encoded in a tourism domain ontology. The experiments showed that this method results in improved searches.

In the TREC Genomics Track, Hersh et al (2003 ) ran one experiment using phrases based on gene name synonyms and another experiment assessed query expansion using external knowledge resources. The results for the first experiment were better than the results for the query expansion experiment. Hersh, Bhupatiraju & Price conclude that the query expansion results could improve if the query is for a specific task.

Nilsson et al (2005) use a domain specific ontology based on Stockholm University Information System (SUiS) to carry out query expansion. SUis differs from other question answering systems because it does not allow free-form questions. The question types are restricted to who, what, when and where. Instead of expanding queries with all semantic relationships provided by an ontology such as WordNet, only synonyms and

hyponyms are used to increase precision. The experiments have shown an improvement in results.

Huang (2000) describes hierarchical directories as ontologies of the web. Each node on the hierarchy acts as a portal the contents of which are related to a particular category. The user does not have to search the entire web, the hierarchical directory allows the user to search the portal which is relevant to that category. A popular web directory is Yahoo!. The other advantage of such hierarchies is that they can be searched to provide a list of topic paths first and if the user is satisfied with the topic path only then the documents for that topic path are retrieved. This means that a user does not have to search through pages of retrieved results to find relevant documents, any ambiguities can be resolved prior to retrieving a document results set. However, such hierarchies need to be kept uptodate and an automatic updating process is preferred due to the continuous growth of the Web compared to updates being carried out manually. Agrawal et al (1998) (as cited by Huang 2000) describe TAPER as an example of an authomatic classification system.

To conclude this section, ontologies have been used for a wide range of information retrieval tasks. As mentioned in 2.6, tasks can be categorised as either work-tasks or social tasks. Domain specific ontologies are more suitable for work-tasks. The terminology in these ontologies is less ambiguous therefore short queries can be expanded with a higher chance of accuracy. General ontologies would be suitable for information type broad queries however the query expansion process may need some guidance or interaction from the user. The trend now is to use multiple ontologies to satisfy a user search query ( Mena et al 2000). Magnini and Speranza (2002 ) state that "Linguistic ontologies encompass ontological and lexical information so partly overcome other limitations such as insufficient broad coverage and the need to be constantly updated". They merge a global ontology with a specialized linguistic ontology such as an economics ontology which includes specific terms and common terms. O'Sullivan et al (1995) merged WordNet with an domain specific ontology created for word processing software. Bao et al (2004) integrate a domain-specific ontology and a domain-independent ontology for colonoscopy video database annotation. The domain specific ontology contains information on colonoscopy and the domain-independent ontology contains general information on properties of videos.

## 7. SUCCESS FACTORS IN USING AN ONTOLOGY FOR QUERY EXPANSION

Ontologies seem to be a promising way forward in query expansion. They improve the accuracy in fuzzy information search by paraphrasing the query of the user through context identification and disambiguation (Leger et al 2001). The success of using an ontology for query expansion depend on various factors. These are described below.

### 7.1 KNOWLEDGE MODEL QUALITY

According to (Cheng & Pan 2004), "the quality of the interpretation of free text is strongly dependent on the quality of the model. Coherence, stability, and resistance to inconsistency and ambiguity are desirable ontological model characteristics". This is

supported by Jones (1993) who states that the quality of a knowledge model or thesaurus is of paramount importance. The model must be accurate, stable, comprehensive and up-to-date. If a data model does not cover the subject area in a comprehensive way then queries which are relevant to a subject area will not get any results because the model is suffering from some omissions.

## 7.2 KNOWLEDGE MODEL FAMILIARITY

Suoemela and Kekalaninen (2005) argue that the search process has a higher chance of success if the user is familiar with the knowledge model. This viewpoint is shared by Sihvonen and Vakkari (2004) who state that query expansion using a thesaurus, is only beneficial if the searcher is familiar with the search topic.

## 7.3 NAVIGABILITY OF KNOWLEDGE MODEL

If a user can navigate a knowledge model with ease, this increases its effectiveness. Some ontologies are hundreds of megabytes in size so suitable mechanisms should be used to allow large ontologies to fit onto one screen otherwise users may 'lose' their way in the vast information space and have difficulty in navigating large knowledge models. To overcome the difficulties users have in navigating ontologies, a mixed approach might be better whereby the system automatically searches the ontology for expansion terms which are suggested to the user who will then interact with the system by selecting the relevant terms. Term suggestion is used to enhance recall while user interactions enables precision to be maintained. This is supported by Efthimiadis (1992) who carried out experiments to study the behaviour of ranking algorithms for query expansion and also end-users during the process of query expansion especially how they select terms. The results provided evidence for effectiveness of interactive query expansion.

The above three success factors are related to the properties of the ontology. In addition to these, other factors however also influence the effectiveness of the ontology for conducting query expansion. Firstly query terms need to be mapped onto corresponding ontology concepts. If an exact match is not found then the mapping process must find the 'next best' match. The entry point into the ontology forms the basis of any subsequent expansion so it is crucial to get this process right. Secondly query length determines whether there will be any resulting benefit from conducting query expansion. It is widely argued that shorter queries are ideal candidates for query expansion because they tend to be more ambiguous. However the query expansion method used must expand the query accurately otherwise a degradation in performance will be evident (Voorhees 1993). Accurate query expansion is when a higher percentage of the suggested terms are relevant to the query and thus produce more relevant documents in the output. Thirdly, as argued by Broder (2002), general, broader information queries benefit more from query expansion than navigational or transactional queries. Finally, using combined query expansion techniques with an ontology produces better results than using a single technique. Schatz et al (1996) used two complementary term suggestors, subject thesauri and co-occurrence lists. Thesauri help to give precision and are generated by human

indexers who place selected terms in a subject hierarchy. Co-occurrence lists are computer generated and place all terms in frequency order of occurrence together. Co-occurrence lists can be used to aid recall because they show the context in which terms are used in the document collection. Mandala et al (1999) identify three types of thesaurus: hand-crafted general purpose thesaurus; co-occurrence based automatically constructed thesaurus and head-modifier based automatically constructed thesaurus. They argue that each type of thesaurus has different advantages and disadvantages. Their experiments show that using of three types of thesauri gives better retrieval results than using just one type of thesaurus. Huang, Huang and Wen (2005) use a dual indexing approach to separately index the collection at document level and passage level. Query terms are searched in each of these indexes and the results are combined using three different merge functions. The dual index model significantly improves the retrieval performance.

This section has tried to identify the properties of an ontology which assist in successful query expansion. In addition to these inherent properties, there are some external factors which also influence the effective use of an ontology during query expansion.

## 8. EVALUATING WORDNET

If we take the first three factors into account, WordNet is based on established lexical rules thus offering a sound foundation for the ontology. Nouns, verbs, adverbs and adjectives are organised by semantic relations into synonym sets (synsets), each representing one concept. Examples of semantic relations used by WordNet are synonymy, antonymy (opposites), hyponymy (is-a) and meronymy (whole-part). Nouns are organised as topical hierarchies and verbs are organised by a variety of relations. Familiarity of a word influences many performance variables such as speed of reading, comprehension, ease of recall and probability of use. Consequently, instead of using frequency of occurrence as an index of familiarity, WordNet uses polysemy. For every word form, an integer is assigned that represents the number of wordsenses that word form has when it is used as a noun or a verb. The words with higher index of familiarity is used for searching. Also another dimension of the familiarity factor, is that this is less problematic because users don't have to have specific knowledge of a particular domain or specific domain terminology.

The user interface design for WordNet is simple but effective thus making it easier for the user to navigate. The interface can be adapted to suit the user's needs through a set of features which are available for selection.

WordNet would not be suitable for automatic text processing because it lacks information such as relations between different parts of speech and conceptual relations such as situation – participant, situations-domain, or object-property (Climent et al 1996). In WordNet there is no distinction between concept and individual (Gangemi et al 2001). For example Palestine and Trust_Territory are hyponyms of Territorial_Domain where the former is specific instance of Trust_Territory. There is limited handling of phrases in

WordNet. Words forming a phrase are represented as a single unit in the hierarchy. Fellbaum (1997) distinguishes between different types of idioms and puts forward some proposals for including idioms into WordNet. WordNet needs to be able to deal with compositional phrases. Bentivogli and Pianta (2004) propose to add phrasets which are sets of free combinations of words which are recurrently used to express a concept. Whilst short query statements are best for query expansion, these can be difficult to disambiguate because the noun is-a hierarchy is not sufficient to reliably select the correct sense of the noun. Navigli and Velardi (2003) suggest gloss words are more useful than synonyms and hypernyms. The similarity between a concepts is based on calculating a gloss overlap measure because the gloss contains more contextual information. Harabagiu et al(1999) enhance WordNet synsets by parsing the gloss information and adding part-of-speech tags and phrasal brackets so that improved use can be made of WordNet as a linguistic tool.

## 9. FURTHER RESEARCH

One of the ways of representing and constraining context is to use an ontology. Therefore, an ontological model can effectively disambiguate meanings of words from free text sentences (Buckland 2003). Gonzalo et al 1998 state that "Ontologies have been used to aid query expansion since the early nineties with mixed success". A detailed investigation into query expansion using a ontologies is needed to study the reasons for their success/failure. A highly effective retrieval model is the probabilistic retrieval model, a detailed description of which can be found in Sparck-Jones et al (2000). Research experiments are needed to examine the effects of using an ontology for query expansion in the newswire domain. Compound words add complexity to the query expansion process however, further research is needed on the effective deployment of compound words in query expansion. Jones et al (1995) note further work is required with compound words which are prevalent in a thesaurus. Their work only records whether a word is part of a compound or not, it does not make any further use of this. Qiu and Frei (1993), argue that selecting query expansion terms based on relatedness to the whole query is more effective. Mandala, Tokunaga & Tanaka also handle any ambiguity in polysemous words by only expanding those terms that have a high similarity to the entire query. Okapi uses a dictionary table of synonyms (Walker and Jones 1987) and the terms in this list are weighted using the BM25 formula. Okapi expands the query by selecting the best terms from the list. This research seeks to enrich the dictionary table of synonyms using ontological information. The user-assisted method of query expansion is proposed for use because it enables users to retain control of the search process and as found by Hancock-Beaulieu and Walker (1992) the user interaction facility of Okapi encourages users to use its query expansion function.

## 10. SUMMARY

This paper looks at the wide range of definitions of context within the field of information retrieval and then focuses on the use of context for query expansion. It analyses the use of relevance feedback, corpus dependent knowledge models and corpus

independent model as ways of handling context within query expansion. A section on the use of ontologies for a range of information retrieval tasks is presented and in particular their use in the area of query expansion. Researchers have reported mixed success in ontology based query expansion. This paper analyses case studies on domain-independent and domain-specific ontology based query expansion in order to examine the reasons for the success or failure of ontology based query expansion. Finally further research is outlined for the use of ontology based query expansion in the area of newswire domain.

## REFERENCES


Agrawal, R., et al. (1998) Scalable Feature Selection, Classification and Signature Generation for Organizing Large Text Databases into Hierarchical Topic Taxonomies., *VLDB J.,* 7 (3) pp.163-178.

Bao, J., et al. (2004) Integration of Domain-Specific and Domain-Independent Ontologies for Colonoscopy Video Database Annotation Proceedings of the International Conference on Information and Knowledge Engineering, June 21-24, CSREA Press, pp 82-90

Barzilay, R. & Elhadad, M. (1997) Using lexical chains for text summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop,ISTS'97

Bateman, J. John Bateman's ontology Portal [online] Available from: http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/ontology/ontology-root.htm [20/07/05]

Bates, M. (2002) After the Dot-Bomb: getting information retrieval right this time, *First Monday - peer reviewed journal*

Bauer, T. & Leake, D. (2001) Real Time User Context Modelling for information Retrieval agents, *CIKM 2001*

Baziz, M., et al. (2005) Conceptual Indexing Based on Document Content Representation Information Context: Nature, Impact, and Role: 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005, June 4-8, p. 171

Bentivogli, L. & Pianta, E. (2004) Extending WordNet with Syntagmatic Information Proceedings of the Second Global WordNet Conference, January 20-23, 2004, pp 47-53

Bharat , K. (2000) SearchPad:explicit capture of search context to support web search, *WWW9/Computer Networks,* 33 (1-6) pp.493-501

Bhavnani, S., et al. (2001) Towards a unified framework of IR Tasks and Strategies Proceedings of ASIST, pp 340-354

Billerbeck, B. & Zobel, J. (2004) Questioning Query Expansion: An Examination of Behaviour and Parameters. Proceedings of the Fifteenth Australasian

Database Conference, ADC 2004, CRPIT 18-22 January 2004, Australian Computer Society
Broder, A. (2002) A taxonomy of Web search., *SIGIR Forum,* 36 (2) pp.3-10
Buckland, M. (2003) Translingual information management using domain ontologies. [online] Available from: http://metadata.sims.berkeley.edu/GrantSupported/tides.html
Buckley, C., et al. (1995) Automatic query expansion using SMART: TREC 3 Proceedings of the Third Text Retrieval Conference, NIST Special Publication 500-225, 69-80
Callan, J., et al. (1992) The INQUERY Retrieval System DEXA, pp 78-83
Chen, H., et al. (1993) Generating, integrating and activating thesauri for concept-based document retrieval, *IEEE Expert,* pp.25-34
Cheng, C. & Pan, X. (2004) Using perception in managing unstructured documents, *Crossroads archive,* Volume 10 (2) p 5
Chu, W., et al. (2002) Textual Document Indexing and Retrieval via Knowledge sources and data mining Communication of the Institute of Information and Computing Machinery(CIICM)
Climent, S., et al. (1996) Definitions of the links and subsets for Nouns of the EuroWordNet Project.In EuroWordNet, E.C. LE2-4003 Deliverable 005 Universitat de Barcelona, Universitat Politècnica de Catalunya, Universidad Nacional de Educación a Distancia. Barcelona, Madrid
Coates-Stephens, S. (1991) Analysis and Acquisition of Proper Names for Robust Text Understanding, Computer Science, City University, London
Communications of the ACM (2002) Special Issue: Ontology Applications and Design, *Communications of the ACM,* 45 (2) pp.39-65
Croft, B., et al. (2001) Relevance Feedback and Personalization: A Language Modelling Perspective DELOS Workshop: Personalisation and Recommended Systems in Diglibs
Cui, H., et al. (2003) Query Expansion by Mining User Logs, *IEEE Transactions on Knowledge and Data Engineering,* 15 (4)
Davies, J. & Weeks, R. (2004) QuizRDF: Search Technology for the Semantic Web Proceedings of the 37th Hawaii International Conference on System Sciences
Deerwester, S., et al. (1990) Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science,* 41 (6) pp.391-407
Efthimiadis, E. (1992) Interactive Query Expansion and Relevance Feedback for Document Retrieval Systems, London, City University
Efthimiadis, E. (1996) Query Expansion, In *Annual Review of Information Systems and Technology (ARIST)*, Vol. 31 (Ed, Williams, M. E.) Information Today, pp. 121-187.
Eguchi, K. (2005) Query Expansion Experiments using Term Dependence Models Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access
El-Beltagy, S., et al. (2001) Linking in Context, *Journal of Digital Information,* 2 (3)

Fellbaum, C. (1997) WordNet: An electronic lexical database and some of its applications, MIT Press, Cambridge, MA.
Finkelstein, L., et al. (2002) Placing search in context: the concept revisited, *TOIS,* 20 (1) pp.116-131
Fu, G., et al. (2005a) Ontology-Based Spatial Query Expansion in Information Retrieval ODBASE: OTM Confederated International Conferences, 4 November 2005
Fu, L., et al. (2005) Evaluating the effectiveness of a collaborative querying environment. Proceedings of the 8th International Conference on Asian Digital Libraries ICADL, Lecture notes in computer science, pp 342-351
Gangemi, A., et al. (2001) Conceptual analysis of lexical taxonomies: the case of WordNet top-level. Proceedings ACM 2nd International Conference on Formal Ontology in Information Systems, FOIS 2001,October 17-19, pp 285-296
Gonzalo, J., et al. (1998) Indexing with WordNet synsets can improve text retrieval Coling-ACL 98
Grootjen, F. & van der Weide, T. (2006) Conceptual query expansion, *Data Knowl. Eng,* 56 (2) pp.174-193
Grover, C., et al. (2003) Use of Ontologies for Cross-lingual Information Management in the Web Proceedings of the Ontologies and Information Extraction International Workshop held as part of the EUROLAN 2003, July 28 - August 8, 2003
Gruber, T. (1993) A translation approach to portable ontologies, *Knowledge Acquisition,* 5 (2) pp.199-220
Hancock-Beaulieu, M. W., S (1992) An evaluation of Automatic Query Expansion in an Online Library Catalogue, *Journal of Documentation,* 48 (4) pp.406-421
Harabagiu, S., et al. (1999) WordNet 2 A Morphologically and Semantically Enhanced Resource Proceedings of SIGLEX-99, June 1999, pp 1-8
Harman, D. K. (1988) Towards Interactive Query Expansion Annual ACM Conference on Research and Development in Information Retrieval archive Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, pp 321-331
Harman, D. K. (1992) Relevance Feedback Revisited ACM SIGIR 15th Conference on Research and Development in Information Retrieval, June 21-24, ACM Press, pp 1-10
Hearst, M. (1992) Automatic acquisition of hyponyms from large text corpora, *14th International Conference on Computational Linguistics, France*.
Hersh, W., et al. (2003) Phrases, Boosting, and Query Expansion Using External Knowledge Resources for Genomic Information Retrieval TREC, pp 503-509
Hoashi, K., et al. (2001) Query Expansion Based on Predictive Algorithms for Collaborative Filtering, *SIGIR,* pp.414-415
Huang, L. (2000) A Survey on Web Information Retrieval Technologies.In ECSL, State University of New York, New York
Huang, X., et al. (2005) A dual index model for contextual information retrieval Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp 613-614

Hwang, C. H. (1999) Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), July 29-30
Hyvonen, E. (2003) Ontology Based Image Retrieval, *www2003,* pp.199
Ingwersen, P. & Jarvelin, K. (2004) Information seeking research needs extension towards tasks and technology, *Information Research,* 10 (1)
Jarvelin, K., et al. (2001) ExpansionTool: Concept-Based Query Expansion and Construction, *Information Retrieval,* (4) pp.231-255
Jones, G. J. F. & Brown, P. J. (2003) Context-Aware Retrieval for Ubiquitous Computing Environments., *Mobile HCI Workshop on Mobile and Ubiquitous Information Access,* pp.227-243
Jones, S. (1993) A thesaurus data model for an intelligent retrieval system, *Journal of Information Science,* 19 pp.167-178
Jones, S., et al. (1995) Interactive Thesaurus Navigation: Intelligence Rules OK?, *Journal of the American Society for Information Science,* 46 (1) pp.52-59
Kang, I. & Kim, G. (2003) Query type classification for web document retrieval Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Annual ACM Conference on Research and Development in Information Retrieval, ACM Press, pp 64-71
Kashyap, V. (2001) Design and Creation of Ontologies for Environmental Information Retrieval, *AOS Workshop, Rome, November 2001*
Khan, L. & Wang, L. (2002) Automatic Ontology Derivation using Clustering for Image classification 8th Intl w/s on Multimedia Information Systems, Oct, pp 56-65
Klink, S. (2004) Improving Document Transformation Techniques with Collaborative Learned Term-based Concepts, In *Reading and Learning: Adaptive Content Recognition*, Vol. 2956 Springer-Verlag, Berlin, Heidelberg, New York, pp. 281-305
Krovetz, R. & Croft., W. B. (1992) Lexical ambiguity in information retrieval., *ACM Transactions on Information Systems,,* 10 (2) pp.115–141
Lame, G. (2003) Using text analysis techniques to identify legal ontologie's components Legont2003, ICAIL 2003 w/s Legal ontologies and web base legal information management
Leger, A., et al. (2001) MKBEEM — Developing Multilingual Knowledge-Based Marketplace [online] (46) Available from: http://mkbeem.elibel.tm.fr/ [20th Dec]
Leroy, G., et al. (2000) Customizable and Ontology-Enhanced Medical Information Retrieval Interfaces, *Methods of Info in Medicine*
Lesk, M. (1969) Word-Word Associations in Document Retrieval Systems, *American Documentation,* 20 (1) pp.27-38
Lifantsev, M. (1998) The Open GRiD Project [online] Available from: http://www.ecsl.cs.sunysb.edu/~maxim/OpenGRiD/ [Feb 2006]
Lin, D. (1999) Automatic identification of non-compositional phrases Proceedings of ACL, 1999, pp.317-324
Liu, X. & Croft, B. (2002) Passage Retrieval Based on Language Models, *CIKM 2002*

Loukachevitch, N. V. & Dobrov, B. (2000) Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems, *Machine Translation Review,* (11) pp.10-20
Lowe, D. (2000) Improving web search relevance: using navigational structures to provide search context [online] Available from: http://ausweb.scu.edu.au/aw2k/papers/lowe/paper.html
Magnini, b. & Speranza, m. (2002) Merging Global and Specialized Linguistic Ontologies Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, LREC-2002, pp 43-48
Mandala, R., et al. (1998) The Use of WordNet in Information Retrieval Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference, pp 191-197
Mandala, R., et al. (1999) Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion, *Research and Development in Information Retrieval*
Mena, E., et al. (2000) Imprecise Answers In Distributed Environments: Estimation Of Information Loss For Multi-Ontology Based Query Processing, *International Journal of Cooperative Information Systems*
Merriam-Webster Online dictionary [online] Available from: http://www.m-w.com/dictionary/context [02/01/06]
Morris, J. & Hirst, G. (1991) Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics,* 17, pp. 21-43
Morris, J. & Hirst, G. (1991) Lexical cohesion, the thesaurus, and the structure of text., *Computational linguistics,* 17 (1) pp.21-48
Navigli, R. & Velardi, P. (2003) An analysis of Ontology-based query expansion strategies Workshop on Adaptive Text Extraction and Mining (ATEM 2003), in the 14th European Conference on Machine Learning (ECML 2003),September 22-26th
Nilsson, K., et al. (2005) SUiS – cross-language ontology-driven information retrieval in a restricted domain Proceedings of the 15th NODALIDA conference
O'Donovan, J. & Smyth, B. (2005) Trust in recommender systems Proceedings of the 10th international conference on Intelligent user interfaces, pp 167-174
Ogawa, Y., et al. (2000) Structuring and expanding queries in the probabilistic model TREC 2000, Nov 13-16
O'Sullivan, D., et al. (1995) Augmenting the Princeton WordNet with a Domain Specific Ontology Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Aug. 19-20
Pavel, S., et al. (2003) Thesauri and Ontologies for Digital Libraries Proceedings of the 5th Russian Conference on Digital Libraries RCDL
Peat, H. & Willett, P. (1991) The limitations of term co-occurrence data for query expansion in document retrieval systems, *American Society for Information Science and Technology (JASIST),* 42 (5) pp.378-383
Pustejovsky, J. (1995) The core lexical engine: The contextual determination of word sense [online] Available from: http://www.cs.tufts.edu/~jacob/isgw/Pustejovsky.html

Qui, Y. & Frei, H. (1993) Concept Base Query Expansion Proceedings of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp 160-169
Rhodes, B. & Maes, P. (2000) Just-in-time information retrieval agents, *IBM Systems Journal,* 39 (3 & 4)
Robertson, A. & Willett, P. (1993) A comparison of Spelling-Correction Methods for the Identification of Word Forms in Historical Text Databases, *Literary & Linguistic Computing,* 8 (3) pp.143-152
Robertson, S. (1986) On Relevance Weight Estimation and Query Expansion, *Journal of Documentation.,* 42 (3) pp.182-188
Robertson, S. (1990) On term selection for query expansion, *Journal of Documentation,* 46 (4) pp.359-364
Robertson, S. E. & Sparck-Jones, K. (1976) Relevance weighting of search terms., *Journal of the American Society of Information Science.,* 27 pp.129-146
Rocchio, R. (Ed.) (1971) Relevance Feedback in Information Retrieval*,* Prentice-Hall, Englewood Cliffs NJ
Ruthven, I. (2004) "... and this set of words represents the user's context ..." Sigir, Information retrieval in context workshop, ACM
Ruthven, I. & Lalmas, M. (2003) A survey on the use of relevance feedback for information access systems, *The Knowledge Engineering Review archive,* 18 (2) pp.95 - 145
Saias, J. (2002) Semantic Enrichment of a Web Legal Information Retrieval System Jurix 2002, Conference on Legal knowledge and IS, Dec 16-17, IOS press, pp 11-20
Salton, G. (1989) Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*.,* Addison-Wesley, Reading, MA.
Salton, G. & Mcgill, J. (1983) Introduction to Modern Information Retrieval*,* Mc_graw Hill, New York
Sanderson, M. (1994) Word sense disambiguation and Information Retrieval Proceedings of the 17th ACM SIGIR Conference, pp 142-151
Sanderson, M. (2004) A study of user interaction with a concept based interactive query expansion support tool (CiQuest) which is integrated into Okapi. Advances in Information Retrieval: 26th European Conference on IR Research, ECIR 2004, April 5-7, DOI: 10.1007/b96895, Lecture Notes in Computer Science, Springer-Verlag Heidelberg, pp. 42 - 56
Sanderson, M. & Lawrie, D. (Eds.) (2000) Building, Testing and applying concept hierarchies*,* Kluwer Academic Publishers.
Schatz, B. R., et al. (1996) Interactive Term Suggestion for Users of Digital Libraries ACM Digital Library Conference
Scott, S. & Matwin, S. (1998) Text Classification using WordNet Hypernyms Coling-ACL Workshop, Usage of WordNet in NLP Systems
Shoval, P. (1985) Principles, procedures and rules in an expert system for information retrieval, *Information Processing and Management,* 21 pp.475-487
Sihvonen, A. & Vakkari, P. (2004) Subject knowledge improves interactive query expansion assisted by a thesaurus., *Journal of Documentation,* 60 (6) pp.673-690

Smeaton, A. & Van Rijsbergen, C. J. (1983) The retrieval effects of query expansion on a feedback document retrieval system, *Computer Journal,* 26 (3) pp.239-246
Soergel, D. (1999) The rise of ontologies or the reinvention of classification, *Journal of the American Society for Information Science,* 50 (12) pp.1119-1120
Sparck -Jones, K. (1973) An evaluation of Query Expansion by addition of clustered terms for a document retrieval system, *Information Storage and Retrieval,* 9 (6) pp.339
Sparck-Jones, K., et al. (2000) A probabilistic model of information retrieval: development and status, *Information Processing and Management,* 36 (6) pp.809-840
Spink, A. (1997) Study of Interactive Feedback During Mediated Information Retrieval, *JASIS,* 48 (5) pp.382-394
Stairmand, M. (1997) Textual Context Analysis for Information Retrieval, *SIGIR,* pp.140-147
Suomela, S. & Kekalainen, J. (2005) Ontology as a search-tool: A study of real users' query formulation with and without conceptual support ECIR 2005, Springer-Verlag Berlin Heidelberg 2005, pp 315-329
Tombros, A., et al. (2001) A study on the use of summaries and summary based query expansion for a question answering task, *23rd BCS European Annual Colloquium on IR Research (ECIR)*
Tombros, A. & Sanderson, M. (1998) the advantages of query-biased summaries in information retrieval, *21st annual international ACM SIGIR,* pp.2-10
Toms, E., et al. (2004) Identifying the significant contextual factors of search, SIGIR, Information Retrieval in Context Workshop
Torrissen, B. C. (1998) Dewey goes surfing: Agent-based information retrieval and classfication support, norwegian university of science and technology
Vakkari, P., et al. (2003) Query Exhaustivity, relevance feedback and search success in automatic and interactive query expansion, *American Society for Information Science and Technology (JASIST),* 60 (2)
Vechtomova, O., et al. (2003) Query expansion with long-span collocates Information Retrieval, Kluwer academic publishers, pp 159-288
Voorhees, E. (1993) Using wordnet to disambiguate word senses for text retrieval, *ACM SIGIR,* pp.171-180
Voorhees, E. (1994) Query expansion using Lexical-semantic relations Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval table of contents, pp 61-69
Walker, S. & Jones, R. (1987) Improving Subject Retrieval in Online Catalogues: 1. Stemming, Automatic Spelling Correction and Cross-Reference Tables.In British Library, London
Xu, J. & Croft, B. W. (2000) Improving Effectiveness of Information Retrieval with Local Context Analysis, *ACM Transactions on Information Systems,* 18 (1) pp.79-112
Yu, C., et al. (2001) Efficient and effective metasearch for text databases incorporating linkages among documents, *SIGMOD*