Conceptual Query Expansion

F.A. Grootjen and Th. P. van der Weide

January 26, 2004

Abstract

Without detailed knowledge of a collection, most users find it difficult to formulate effective queries. In fact, as observed from Web search engines, users may spend large amounts of time reformulating their queries in order to satisfy their information need. A proven successful method to overcome this difficulty is to treat the query as an initial attempt to retrieve information and use it to construct a new, hopefully better query. Another way to improve a query is to use global (thesauri-like) information. In this article a new, hybrid approach is presented that projects the initial query result on the global information, leading to a local conceptual overview. The resulting concepts are candidates for query refinement.

To show its effectiveness, we show that the conceptual structure resulting after a typical short query (2 terms) contains refinements that perform as well as a most accurate query formulation.

Next we show that query by navigation is an effective mechanism that in most cases finds the optimal concept in a small number of steps. If the optimal concept is not found, then it still finds an acceptable sub-optimum. We show that the proposed method compares favorably to existing techniques.

1 Introduction

Formulating queries is not a simple task. Most searchers seem to have a searching strategy where the search engine's contribution is recall whereas the searcher is concerned with precision (see for example [3]). Typically searchers carefully select a small (3 or less) number of keywords to describe their information need. Such queries are referred to as short queries. For example, in [30] it is reported that for the Excite Web search engine the average number of terms in a query amounted to 2.4 over 1997, 2.4 over 1999 to 2.6 over 2002.

The searcher thus is challenged with finding a small combination of discriminating keywords. This is especially critical when the searcher is not familiar with the actual description of relevant documents. Another problem is that short queries usually have a large degree of ambiguity. The query result of an adequate short query is seen as an indication of the lowerbound on query formulation expressiveness.

Full text queries as used in TREC [12] ad-hoc evaluations (see appendix 1) may be seen as a most accurate formulation of the information need. They are balanced, and cover all kinds of details that are of relevance for the searcher. We interpret the results of these queries as an indication of the upperbound on query formulation expressiveness. In this paper we will try to bridge the gap between lowerbound and upperbound, and offer a mechanism to help the searcher to go beyond this upperbound on formulation expressiveness. In order to solve this problem, the query result is interpreted as a materialized meaning assignment to the query (see [10]). In case of ambiguity, the query result is a mix of the various interpretations of the query. For non-ambiguous queries, this may lead to a better formulation of the information need by refining the query with elements of this meaning. In case of ambiguity, the refinement process starts with the selection of the intended particular meaning variant.

Some directions have been chosen to handle the query refinement problem. We will shortly discuss the extensional approach, the intentional approach and the collaborative approach.

Weather related fatalities.
Document will report a type of weather event which has directly caused at least one fatality in some location.
A relevant document will include the number of people killed and injured by the weather event, as well as reporting the type of weather event and the location of the event.

Figure 1: TREC query 59

1. The extensional approach is based on the materialization of the information need in terms of documents. This query enrichment method may be done with user intervention, for example using relevance feedback, or without user intervention, by a so-called local analysis. In case of relevance feedback, the searcher may be asked to assign relevance judgements to a number of documents. Another approach is to ask the searcher to indicate the highest ranked relevant document.

2. The intentional approach is based on the meaning of the keywords. Thesauri are used to describe the meaning and relations of terms. By locating the short query within the thesaurus, a reformulation of the query can be obtained. The thesaurus may be based on the underlying collection (if any) or based on world knowledge (Wordnet). In the case of a restricted area of interest, a representative collection may serve as meaning framework for that area of interest.

3. In the collaborative approach the system tries to employ previous behavior of searchers to obtain a better idea of the intended meaning of a query. See for example [5, 16, 15].

In this paper we employ a mix between the extensional and the intentional approach. We discuss a hybrid form of query expansion called Conceptual Query Expansion. In this approach, both the local initial query result and the global information from the complete collection are used. Rather than using the conceptual structure resulting from global collection information, we project this structure onto the local initial query result. The resulting concepts may be seen as a structured overview of the various interpretations of the query.

In our experiments on the TREC Associated Press collection, we use automatically generated index expressions to describe the contents of documents. This mechanism is better suitable than using word combinations. The reason is that it finds word combinations that go beyond word position in the sentence. Besides, index expressions use a connector to clarify the relation between the combined words. This will lead to a better quality of the generated concepts. Index expressions are effectively obtained from a text with a high level of precision by the grammarless parser technique [8].

Our assumption is that the best ranked documents of the initial query result form a proper basis to generate a local thesaurus. Even when a top ranked document is not relevant for the searcher, it will still address the topics around the query. The top-ranked documents thus may be used to map out the meaning structure associated with the query to some reasonable extent. In order to investigate this principle and to get an impression of its potentials, we will restrict ourselves in this paper to the benefits of a single document, the top-ranked document. In our experiments, we will also consider the effect of using the highest ranked relevant document instead.

After generating the local thesaurus by projecting the global thesaurus onto this top-ranked document, we obtain a conceptual (sub)structure. We argue that each concept in this structure relates to a specific and meaningful query expansion. The problem is to find the best concept, and then use the related query expansion.

An important result of our experiments is that this local thesaurus contains concepts that approximate the full text formulation (the upperbound) very well (see section 4). As a consequence, the local thesaurus is a reasonable framework for query expansion.

However, a simple heuristic to locate a (nearly) optimal concept is not obvious. In practice, a searcher may navigate through this local thesaurus (Query by Navigation, see [4]) to locate the best suitable concept. To illustrate the power of this mechanism, we let a simulated searcher perform this process of Query by Navigation. Our experiments show that this searcher will effectively locate the best concept in only a few navigation steps. Usually the searcher takes a straight course. In some case, however, finding this concepts takes after a bit of wandering around. The average navigation path in all cases is very limited.

The structure of this paper is as follows. In section 2 we discuss related work. In section 3 the proposed model is described. In section 4 we describe the experiments that show the presence of good quality concepts. In section 5 the experimental results for query by navigation are presented and discussed. Section 6 contains conclusions and further directions for research.

2 Query expansion

In this section we discuss in some more detail the query expansion techniques hat will be used for comparison.

2.1 User Relevance Feedback

Probably the most popular query reformulation technique is Relevance Feedback. Using an initial retrieval result the user is asked to mark relevant documents in a list of 10 (or 20) ranked documents. Early experiments [25] have shown good improvements in precision for small test collections using relevance feedback. Although relevance feedback is relatively easy to implement, in practice it seems to be very difficult to persuade a searcher to tediously work through a list of documents and marking them relevant. At best, a user may be asked for selecting or a relevant document when presented a list of document excerpts. Information like this is easy to collect, since this is exactly what we do when using a search engine like for example Google.

If the initial query result is expected to be good, one might consider an alternative form of relevance feedback, assuming for example that the top ranked document is relevant. This mechanism is called pseudo relevance feedback.

There are several ways to calculate the new query. For the vector model, a single relevant document and plain positive feedback strategy (no non-relevant documents selected by user), Rocchio [24] and IDE [13] provide the same formula for the modified $\vec{q_m}$.

$$\vec{q}_m = \alpha\vec{q} + \beta\vec{d}$$

where $\vec{q}$ is the original query, and $\vec{d}$ the (pseudo) relevant document, and α and β are tuning parameters. Assuming that both the query and the documents are normalized, taking $\alpha = 1$ and $\beta = 1$ seems to be a reasonable choice.

2.2 Global query expansion

Another way of query expansion is is adding words (synonyms) or related words with respect to the original query. By doing this the knowledge stored in a thesaurus or other (global) information source can be used to increase recall. Thesauri have frequently been incorporated in information retrieval systems as a device for the recognition of synonymous expressions and linguistic entities that are semantically similar but superficially distinct.

Automatic query expansion using thesauri has been the target of research for nearly four decades, and a lot of methods have been proposed. [19] presents an concise overview of these methods and distinguishes 3 categories:

- Hand-crafted thesauri
- Co-occurrence based thesauri

- Head modifier based thesauri

Query expansion based on hand-crafted thesauri is only successful if a the thesaurus is domain-specific and correspond closely to to the domain-specific document collection [6]. The use of general purpose hand-crafted thesauri for automatic query expansion has not been very succesfull [31, 29]. Experiments with co-occurence based thesauri show a gain of 20% in retrieval performance [21] on small test collections, but is less effective on larger collections [27]. More linguistically motivated approaches like head modifier based thesauri show similar results [7, 14]. As shown in [19], a combination of query expansion techniques yields better performance than the techniques on their own.

Note that this technique does not require a relevant document.

2.3 Related work

The use of concept lattices for the construction of knowledge bases has been recognized before, see for example [26]. The effects of term dependence structure have been studied before, with spanning trees as underlying data structures. The results were not positive (see [1, 23]). In [22] a term similarity thesaurus is used, and shown to lead to a significant retrieval improvement on small collections. In [28] an client side web agent (ARCH) uses domain knowledge from web based classification hierarchies such as Yahoo combined with user profile information. In this paper we focus on conceptual query expansion using dynamically created thesauri.

3 Conceptual query expansion

Obviously, the words in a (pseudo) relevant document are good candidates for query expansion. However, the question is: which words (or better, combinations of words) are the most suitable? Even when a document contains only a couple of hundreds of words, there are many possible combinations. Most of them will be meaningless. We will use *term* and *descriptor* as generic terms for *word*.

The technique proposed in this article called Conceptual Query Expansion uses a special notion to drastically limit the number of possible term combinations: the notion of formal concepts. The key thought is to consider only those combination of terms that make sense in the collection, that is: only consider combinations of terms that form a formal concept.

3.1 Formal Concept Analysis

Before continuing, we will shortly discuss the elements of *Formal Concept Analysis* [32, 11].

3.1.1 Context

Suppose we have a collection $\mathcal{D}$ of documents. Individual members of this collection (documents) are written with small letters like d, d_1, d_2, while subsets are written in capitals (D, D_1, D_2). During the indexing process, descriptors (attributes) are attached to documents. We write $\mathcal{A}$ to denote the set of all attributes, a, a_1, a_2 for individual attributes and A, A_1, A_2 for attribute sets (subsets of $\mathcal{A}$). The result of indexing process is reflected in the binary relation $\sim$: we write $a \sim d$ iff attribute a describes document d. The tuple $(\mathcal{D}, \mathcal{A}, \sim)$ is called a *context*.

The context relation $\sim$ is overloaded to cover set arguments in the following way:

$$a \sim D \equiv \forall_{d \in D} [a \sim d]$$

$$A \sim d \equiv \forall_{a \in A} [a \sim d]$$

$$A \sim D \equiv \forall_{a \in A, d \in D} [a \sim d]$$

3.1.2 Properties of contexts

Using the context relation a classification of documents and attributes can be generated such that each class can be seen as a *c*oncept in terms of properties of the associated documents and attributes. In our interpretation, documents and attributes assign meaning to each other via the context relation: within the limits of this view, we can not distinguish between document with identical properties, while attributes having the same extensionality are assumed to be identical. Sharing document meaning thus can be seen as sharing attributes:

Definition 1
The common attributes of a set of documents are found by the right polar function **ComAttr** : $\mathcal{P}(\mathcal{D}) \rightarrow \mathcal{P}(\mathcal{A})$ defined as follows:

$$\textbf{ComAttr}\,(D) = \{a \in \mathcal{A} \mid a \sim D\}$$

Documents may also be shared by attributes:

Definition 2
The documents sharing properties are captured by the left polar function **ComDocs** : $\mathcal{P}(\mathcal{A}) \rightarrow \mathcal{P}(\mathcal{D})$ defined by:

$$\textbf{ComDocs}\,(A) = \{d \in \mathcal{D} \mid A \sim d\}$$

3.1.3 Concepts

A special situation is when the duality of meaning between a set D of documents and a set A of attributes is symmetric: $A \sim D$. It is easily verified that $D \subseteq$ **ComDocs** (A) and $A \subseteq$ **ComAttr** (D). If the sets D and A are maximal, then this combination is referred to as a concept:

Definition 3
A *concept* is a pair $(D, A) \in \mathcal{P}(\mathcal{D}) \times \mathcal{P}(\mathcal{A})$ such that D and A are their mutual meaning:

$$\begin{aligned} \textbf{ComAttr}\,(D) &= A \\ \textbf{ComDocs}\,(A) &= D \end{aligned}$$

Obviously not *every* set of documents (attributes) forms a concept. But when it does at, most one concept can be associated with it. So a concept is uniquely identified by its set of documents or its set of attributes.

Definition 4
Let $c = (D, A)$ be a concept, we will write $\delta(c)$ to denote its extensionality D and $\alpha(c)$ for its intention A.

3.1.4 The concept lattice

Let $\mathcal{C}$ be the set of all concepts that can be derived from the set of documents $\mathcal{D}$ an the set of attributes $\mathcal{A}$ and their relation $\sim$. These concepts are ordered in the following way:

Definition 5
A concept c_1 is more specific than concept c_2 if it has a restricted extensional meaning:

$$c_1 \subseteq c_2 \equiv \delta(c_1) \subseteq \delta(c_2)$$

Having a restricted extensional meaning is equivalent with having an augmented intentional meaning: $c_1 \subseteq c_2 \Longleftrightarrow \alpha(c_1) \supseteq \alpha(c_2)$. The fact that $(\mathcal{C}, \subseteq)$ is a partial order follows directly from the fact that $(\mathcal{P}(\mathcal{D})$ is a partial order.

Let C be a set of concepts. A lower bound of C is a common subconcept. If there exists a greatest element in the set of lower bounds of C, then this element is called *greatest lower bound*,

and denoted as $\wedge(C)$. Likewise the smallest element in the set of upper bounds is called *smallest upper bound*, denoted as $\vee(C)$.

It can be proven that for each set of concepts C:

$$\begin{array}{lcllcl} \delta(\wedge(C)) & = & \bigcap_{c\in C}\delta(c) & \alpha(\wedge(C)) & = & \textbf{ComAttr}\,(\delta(\wedge(C))) \\ \alpha(\vee(C)) & = & \bigcap_{c\in C}\alpha(c) & \delta(\wedge(C)) & = & \textbf{ComDocs}\,(\alpha(\vee(C))) \end{array}$$

As (in our case) each set of concepts has a unique lower and upper bound, the resulting lattice $(\mathcal{C}, \subseteq)$ is a complete lattice. This property is important when generating concept lattices as we will see in section 4.5.

4 Evaluating the Expressiveness of the Concept Lattice

In order to test the expressiveness of the concept lattice, we investigate the quality of the concepts as possible query expansions bridging the gap between the lower boud and the upperbound. Our intention is to show that the concept lattice contains concepts that approximate the full text query reasonably well.

To test the different query expansion techniques we ran a number of experiments on the Associated Press collection used in TREC competitions. This collection is approximately 800Mb big, contains 250,000 documents and is accompanied by 50 queries with their relevance judgements. It consists of more than 100,000,000 words of which were 300,000 unique. All tests were done by BRIGHT, a SMART like vector model based tool using tf-idf document weighting.

BRIGHT uses linguistic stemming (lemmatizing the words to their base form) without stopword removal. The indexer is capable of generating both single word descriptors as well as index expressions (with lenght 2) which improves retrieval results (see [9]). Index expressions go beyond linguistic head-noun modifiers (see [2, 17]). Furthermore, BRIGHT contains a concept lattice builder.

The retrieval results are measured on recall levels 0.0, 0.1 ...1.0 averaged over all queries. Finally we calculate an 11-point and a 3-point average precision value for each run.

Our experiments are based on queries 51-75. A special case is query 65, as it has no relevant document in the AP collection. In the next subsections we describe the various elements of our experiments.

type	11-pt average		3-pt average	
full	0.1462	(100%)	0.1348	(100%)
2-word	0.0990	(68%)	0.0878	(65%)
wordnet1	0.0714	(49%)	0.0625	(46%)
wordnet2	0.0619	(42%)	0.0540	(40%)
rftop	0.1038	(71%)	0.0918	(68%)
rfrel	0.2956	(202%)	0.2512	(186%)
cqetop	0.1377	(94%)	0.1296	(96%)
cqerel	0.3434	(235%)	0.3072	(228%)
navtop	0.1282	(88%)	0.1207	(90%)
navrel	0.3064	(210%)	0.2714	(203%)

Table 1: Result overview

4.1 The full query retrieval experiment

For each selected query the retrieval result is determined on the original full text query. This run will probably yield the best retrieval results since all information in the original query is used.

The results of this retrieval experiment will be used as a baseline for comparison with the query expansion runs (see table 1).

#	query	#	query	#	query
51	airbus, government	52	Africa, sanction	53	leveraged, buyout
54	satellite, launch	55	insider, profit	56	prime, rate
57	MCI, financial	58	rail, strike	59	weather, fatality
60	merit, seniority	61	Israel, affair	62	military, coup
63	machine, translation	64	hostage, political	66	language, processing
67	disturbance, political	68	fiber, hazard	69	SALT, revive
70	surrogate, motherhood	71	border, incursion	72	U.S., movement
73	country, movement	74	conflicting, policy	75	automation, cost

Table 2: 2-word queries

4.2 The 2-word query experiment

Since full text queries are not so common, we manually produced for each of the 24 queries a 2-word query alternative (see table 2). The selection of two keywords is straightforward for most queries. The difficulty of describing the information need in two words becomes apparent while performing this exercise. Although the choice for some of keywords seems to be arbitrary, it is the goal of the experiment to see what happens in different situations.

The retrieval performance of this sophisticated searcher using 2-word queries is summarized in table 1. In Figure 2 lowerbound and upperbound are presented in a recall-precision graph.

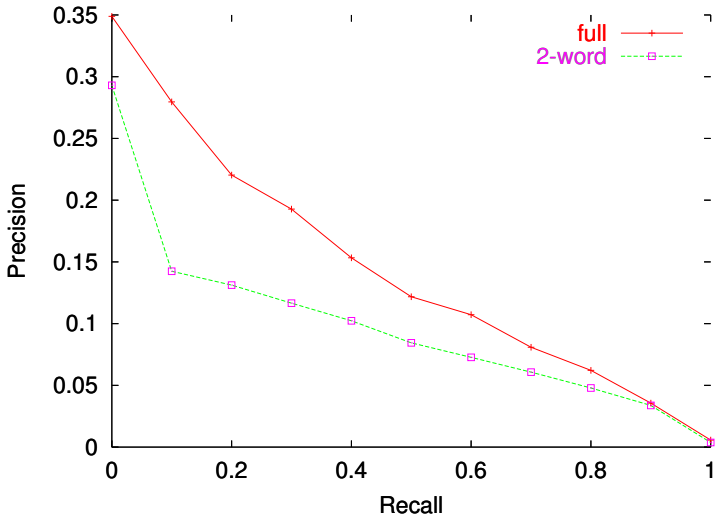

Figure 2: Comparing lowerbound and upperbound

4.3 Global query expansion

In this sub-experiment the 2-word query is expanded with global information. We use Wordnet [20] to expand the query in two ways:

1. Adding all synonyms (words from the same Wordnet-sense)
2. Adding all synonyms and related words

The effects on the retrieval performance for global query expansion using Wordnet synonyms (wordnet1) and both synonyms and related words (wordnet2) are shown in table 1.

As can be seen in figure 3 the results for global query expansion are poor. This is consistent with other research. Still, the experiment is included to make the contrast between feedback mechanisms on the same document collection more explicit.

Query expansion based on hand-crafted thesauri only succeeds if a domain-specific thesaurus is used which corresponds closely to domain-specific document collection [6]. The results for general-purpose hand-crafted thesauri are disappointing [31, 29]. However some good results are obtained by automatically created thesauri [19]. For a detailled study why Wordnet expansion is not working see [18].

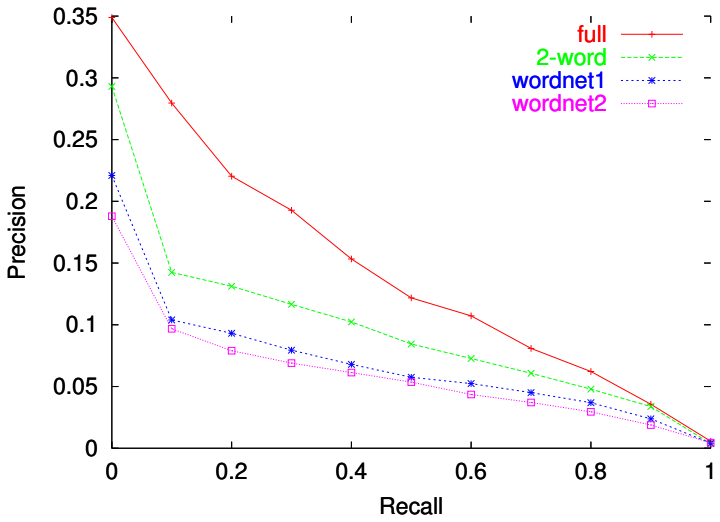

Figure 3: Performance after global query expansion

4.4 Relevance feedback

This experiment uses classical user relevance feedback to expand the query. The query is expanded according to the Rocchio method in two ways:

1. Assuming the top ranked document is relevant (pseudo relevance feedback)
2. Using the top relevant document from the retrieval result (user relevance feedback)

The outcome is summarized in table 1, and displayed in Figure 4.

4.5 Conceptual Query Expansion

The use of Formal Concept Analysis for Information Retrieval purposes looks appealing, but due to the size of collections a straightforward calculation of concept lattices is impossible. Even with the fastest algorithms known today, calculating the complete lattice for the Associated Press context, spanning 250,000 documents and 4,000,000 attributes is out of the question.

But in some cases, there is no need to calculate the complete lattice. Sometimes a sublattice, calculated on the fly, may suit its purpose.

Suppose that the initial query produced a relevant (or pseudo relevant) document. The terms in this document are probable candidates for query expansion. By calculating the sublattice generated from these terms as attributes, we find concepts (that is combinations of terms) that have a conceptual meaning in terms of the collection. The calculation of this sublattice is easy, and may be split into two steps:

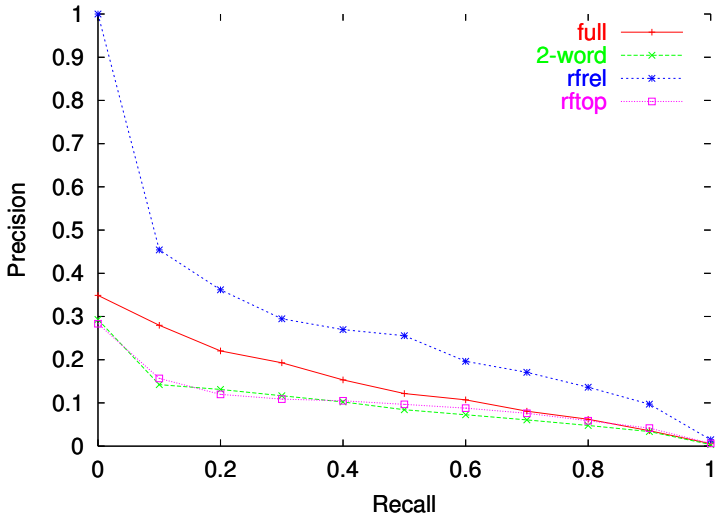

Figure 4: Performance after incorporating Rocchio relevance feedback

1. Base concepts: For each attribute a calculate the concept (**ComDocs** (**ComAttr** $(\{a\})), \{a\})$.
2. Compound concepts: Take two concepts c_1 and c_2, and generate their join $c_1 \vee c_2$.
3. Repeat step 2 until while new concepts can be generated.

The concept lattice generated by the attributes of a document from the Associated Press collection typically contains a few hundred concepts. In this experiment, we evaluate all concepts in the lattice and use the concept with the best 11 point average precision/recall value.

The folling table show some example expansions generated for both top ranked as top relevant documents.

query	original query	top relevant	top ranked
q51	airbus, government	germany, spain	germany
q53	leveraged, buyout	repay	use
q54	satellite, launch	commercial, the company, the department, government, licence, launch licence, rocket, transportation	rocket
q58	rail, strike	commuter	commuter
q59	weather, fatality	central, destroy, flood, home, kill, more, people, province, the storm	-
q72	movement, u.s.	bureau, census bureau	edge
q73	movement, country	emigration	here

The power of conceptual query expansion is illustrated by the fact that even non-relevant documents may lead to good expansion: both the relevant document `ap900914-0105` as the non-relevant document `ap890203-0058` expand `rail` and `strike` to `commuter` for query 58. The difference in expansion for query 72 and 73 is also remarkable.

The results of the conceptual query expansion are presented in table 1 and in figure 4.5

5 Navigation experiment

From the previous section we know that the generated concept lattice contains concepts that can be used to create appropriate query expansion terms. The results show that both in the case of selecting the top ranked and top relevant document the lattice contains such optimal concepts.

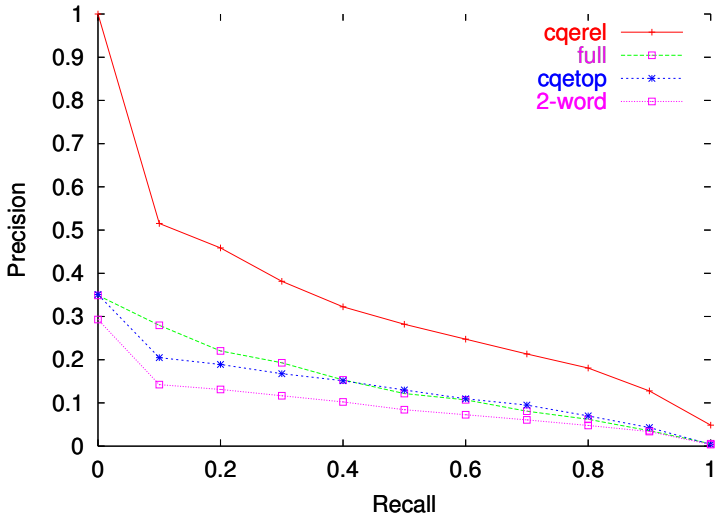

Figure 5: Conceptual Query Expansion

A good heuristic to find an optimal concept is not obvious. This section discusses how to find good concepts in the lattice by navigation.

5.1 Distribution

Figure 6 shows (for query 51) the distribution of the concepts according to their 11-points average precision they would produce if they were used for query expansion. It is clear that some con-

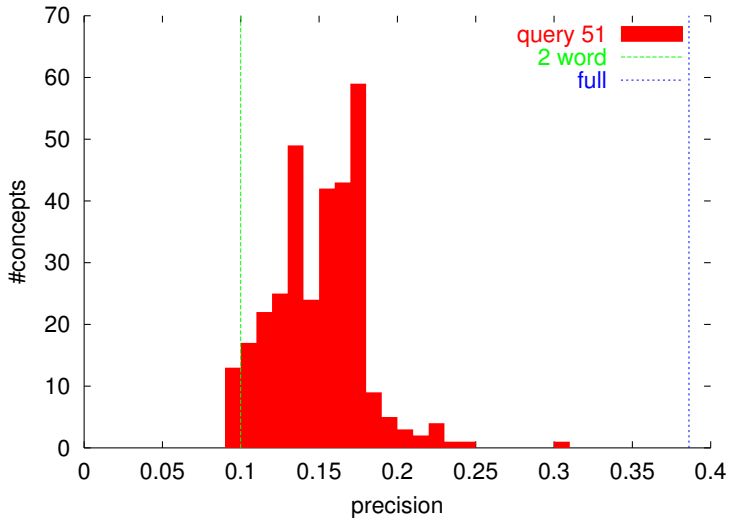

Figure 6: Concept distribution for query 51

cepts degrade retrieval performance, while others improve it. The question is, how do we find the best concept?

5.2 Navigation

Concept lattices are structured; the concepts in the lattice form a partial order. This partial order can be used for navigation: a user may select a subconcept and navigate down in the lattice making his query more specific by adding terms, or loosing terms by navigating up to a superconcept and making the query more general. We will call the current concept the user's focus. The process of navigating down is called refinement, and navigating up is called enlargement.

We will illustrate the navigation process for query 59 (see figure 7). Navigation starts at the top node (empty expansion). The user selects `flood` as expansion candidate. By doing so the 11-point average precision rises from 0.1107 to 0.1497. Subsequently the next step adds both the terms `central` and `storm`, with the accompanying score of 0.1788. Finally the end concept is reached by adding (in one single step) the terms `destroy`, `home`, `kill`, `more`, `people`, `province` and `the storm`, good for an average precision of 0.2429.

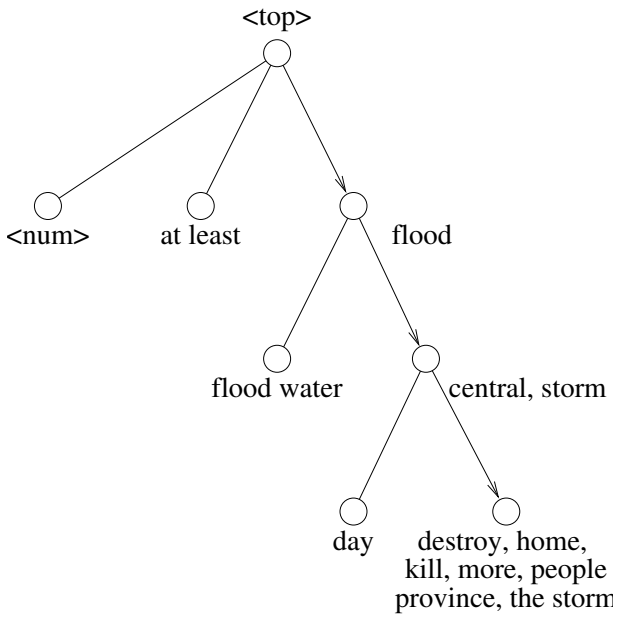

Figure 7: Example navigation

In the next section we will we try to make it plausible that a searcher is able to efficiently find a good query expansion concept by navigating the concept lattice in only a few steps.

5.3 Simulation

In order to simulate a searcher, we wrote a simple program program called `autonav` that is capable of navigating the concept lattice. The program starts in the top concept of the lattice (the least specific concept) and iteratively chooses the best sub or super concept (in terms of the 11-points average precision) and changes its focus to that new concept. The navigation process ends if there is no neighboring concept with a better score than the concept in focus.

5.4 Navigation results

The results of the simulated navigation for the selected top ranked document and top relevant document are presented in table 3. For top ranked document selection, the average number of navigation steps is 1, the average approximation of the best concept is 92%. Obviously, for top

relevant document selection the average number of navigation steps is somewhat higher (≈ 3). The average approximation is coincidentally the same.

Figure 8 shows the results together with standard relevance feedback. From the figure it is clear that:

- Conceptual relevance feedback delivers a significant gain in retrieval performance, even if the top document is not relevant.
- Since it is not guaranteed that the best concept is found during navigation, the performance is somewhat lower, but still significantly better than without feedback or standard relevance feedback.

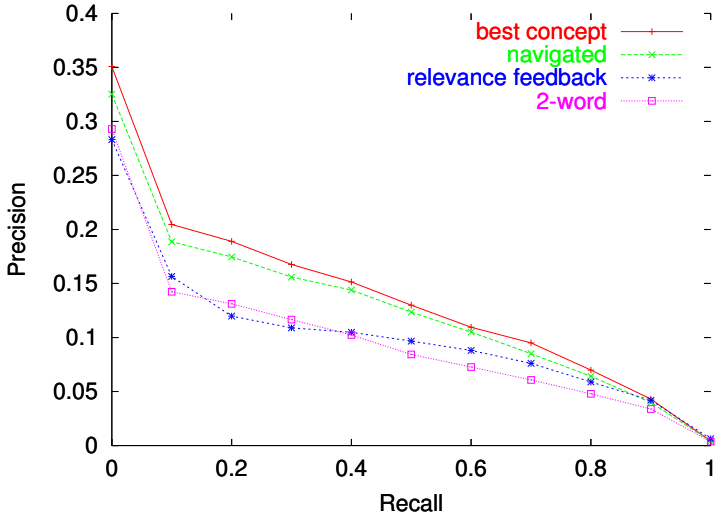

Figure 8: Navigation performance for top ranked document selection

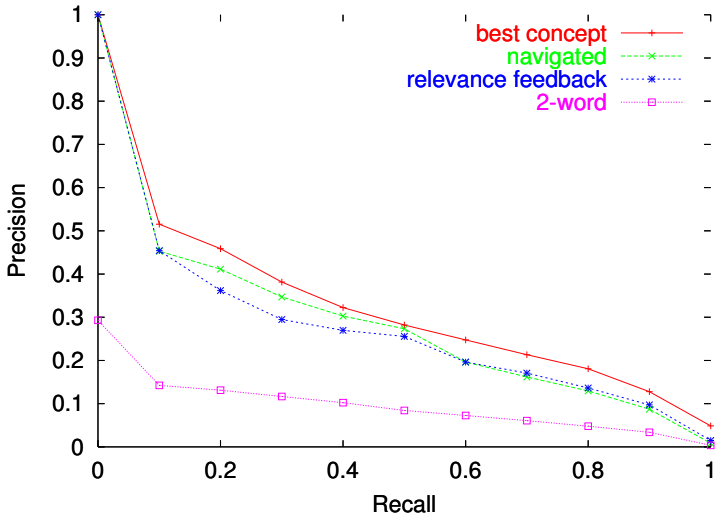

Figure 9: Navigation performance for top relevant document selection

top ranked document				top relevant document			
query	steps	score	approx	query	steps	score	approx
q51	1	0.1865	(89%)	q51	2	0.3081	100%
q52	1	0.2982	(100%)	q52	1	0.3436	100%
q53	1	0.2988	(100%)	q53	1	0.3053	90%
q54	2	0.2126	(100%)	q54	1	0.3116	85%
q55	1	0.2275	(84%)	q55	1	0.2729	95%
q56	0	0.1671	(100%)	q56	1	0.3993	100%
q57	1	0.1782	(98%)	q57	1	0.2415	86%
q58	1	0.2444	(93%)	q58	1	0.2807	100%
q59	0	0.0270	(100%)	q59	3	0.2429	100%
q60	1	0.0010	(100%)	q60	4	0.2642	100%
q61	1	0.0132	(49%)	q61	1	0.1589	42%
q62	1	0.1228	(87%)	q62	3	0.1995	100%
q63	1	0.0852	(100%)	q63	2	0.3207	91%
q64	1	0.1096	(58%)	q64	5	0.2267	100%
q66	1	0.0005	(100%)	q66	2	0.5482	55%
q67	1	0.0209	(45%)	q67	3	0.2586	100%
q68	2	0.2171	(100%)	q68	2	0.2701	100%
q69	1	0.0047	(100%)	q69	5	0.8960	100%
q70	1	0.3138	(100%)	q70	7	0.7396	100%
q71	1	0.2944	(100%)	q71	1	0.2944	98%
q72	1	0.0007	(100%)	q72	3	0.1539	100%
q73	1	0.0015	(100%)	q73	1	0.0984	80%
q74	1	0.0116	(100%)	q74	1	0.0936	100%
q75	1	0.0390	(98%)	q75	1	0.1254	92%

Table 3: Navigation results

6 Conclusions

In this paper we discussed a way to overcome the inherent shortcomings of short queries, and discussed its potential effectiveness. We showed that it is even possible to benefit from non-relevant documents. We proposed a mechanism to help searchers finding their way in the semantical richness of the meaning of a short query by exploiting latent knowledge stored in the collection..

A possible direction for further research is to incorporate previous behavior of the searcher. This may lead to heuristics for query reformulation without searcher interaction.

Another direction can be to investigate 'reference lattices' or cross lingual lattices for general usage.

References

[1] Smeaton A.F. and Rijsbergen C.J. van. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[2] A. Arampatzis, Th. P. van der Weide, C.H.A. Koster, and P. van Bommel. An evaluation of linguistically-motivated indexing schemes. In *Proceedings of BCS-IRSG 2000 Colloquium on IR Research*, Sidney Sussex College, Cambridge, England, 5–7 April 2000.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[4] P.D. Bruza and Th. P. van der Weide. Stratified hypermedia structures for information disclosure. *The Computer Journal*, 35(3):208–220, 1992.

[5] H. Cui, J. Wen, J. Nie, and W. Ma. Query expansion for short queries by mining user logs. `http://research.microsoft.com/asia/dload_files/group/mediasearching/2002p/QE-TKDE.pdf`.

[6] E.A. Fox. Lexical relations enhancing effectiveness of information retrieval systems. *SIGIR Forum*, 26(5):629–640, 1980.

[7] Gregory Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 89–97. ACM, 1992.

[8] F.A. Grootjen. Indexing using a grammerless parser. In *Proceedings of the 2001 IEEE International Conference on Systems, Man and Cybernetics, (NLPKE 2001)*, Tucson, Arizona, USA, October 2001.

[9] F.A. Grootjen and Th. P. van der Weide. Effectiveness of index expressions. Technical Report NIII-R0329, University of Nijmegen, 2003.

[10] F.A. Grootjen and Th. P. van der Weide. Information retrieval as a semantics transformation mechanism. a formal theory for latent semantics. Technical Report NIII-R0303, University of Nijmegen, 2003.

[11] G. M. Hardegree. An approach to the logic of natural kinds. In *Pacific Philosophical Quarterly*, volume 63, pages 122–132, 1982.

[12] D. Harman. Overview of the first TREC conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference*, 1993.

[13] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*, pages 337–354, 1971.

[14] Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.

[15] Stefan Klink. Query reformulation with collaborative concept-based expansion. In *First International Workshop on Web Document Analysis*, pages 19–22, 2001.

[16] Stefan Klink, Armin Hust, Markus Junker, and Andreas Dengel. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. In *Proceedings of DAS 2002, 5th International Workshop on Document Analysis Systems*, volume 2423 of *Lecture Notes in Computer Science*, pages 376–387, Princeton, NJ, USA, August 2002. Springer.

[17] C.H.A. Koster. Head-modifier frames for everyone. In *Proceedings of SIGIR-2003, International Conference on Research and Development in Information Retrieval*, page 466, Toronto, 2003.

[18] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. The use of WordNet in information retrieval. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 31–37. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[19] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 191–197. ACM, 1999.

[20] G. A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.

[21] Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.

[22] Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.

[23] C.J van. Rijsbergen, D.J. Harper, and M.F. Porter. The selection of good search terms. *Information Processing and Management*, 17:77–91, 1981.

[24] J.J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System – Experiments in Automatic Document Processing*, Englewood Cliffs, NJ, 1971.

[25] G. Salton, editor. *The SMART Retrieval System – Experiments in Automatic Document Processing*, Englewood Cliffs, NJ, 1971.

[26] J. Sarbo. Building sub-knowledge bases using concept lattices. *The Computer Journal*, 39(10):868–875, 1997.

[27] H. Schutze and J.O. Pederson. A coocurrence-based thesaurus and two applications to information retrieval. In *Proceedings of RIAO Conference*, pages 266–274, 1994.

[28] A. Sieg, B. Mobasher, S. Lytinen, and R. Burke. Using concept hierarchies to enhance user queries in web-based information retrieval. *Proceedings of the The IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria*, pages 226–234, febr 2004.

[29] A.F. Smeaton and C. Berrut. Tresholding posting lists, query expansion by word-word distances and the pos tagging of spanish text. In *Proceedings of the fourth Text Retrieval Conference*, 1996.

[30] A. Spink, B.J. Jansen, D. Wolfram, and Saracevic. T. Searching of the web: the public and their queries. *Journal of the American Society of Information Science and Technology*, 52(3):226–234, 2001.

[31] E.M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69. ACM, 1994.

[32] R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470. D. Reidel Publishing Company, Dordrecht–Boston, 1982.