

Document text characteristics affect the ranking of the most relevant documents by expanded structured queries

Eero Sormunen, Jaana Kekäläinen, Jussi Koivisto and Kalervo Järvelin
University of Tampere
Department of Information Studies
Finland
{lieeso, lijakr, likaja}@uta.fi

Abstract

The increasing flood of documentary information through the Internet and other information sources challenges the developers of information retrieval systems. It is not enough that an IR system is able to make a distinction between relevant and non-relevant documents. The reduction of information overload requires that IR systems provide the capability of screening the most valuable documents out of the mass of potentially or marginally relevant documents. This paper introduces a new concept-based method to analyze the text characteristics of documents at varying relevance levels. The results of the document analysis were applied in an experiment on query expansion (QE) in a probabilistic IR system.

Statistical differences in textual characteristics of highly relevant and less relevant documents were investigated by applying a facet analysis technique. In highly relevant documents a larger number of aspects of the request were discussed, searchable expressions for the aspects were distributed over a larger set of text paragraphs, and a larger set of unique expressions were used per aspect than in marginally relevant documents. A query expansion experiment verified that the findings of the text analysis can be exploited in formulating more effective queries for best match retrieval in the hunt for highly relevant documents. The results revealed that expanded queries with concept-based structures performed better than unexpanded queries or 'natural language' queries. Further it was shown that highly relevant documents benefit essentially more from the concept-based QE in ranking than marginally relevant documents.

1. Introduction

Fundamental problems of IR experiments are linked to the complex notion of relevance [1, 2, 3, 4, 5, 6]. One of the problems is that in most laboratory experiments documents are judged either relevant or irrelevant with regard to the request. Binary relevance cannot reflect the possibility

that documents may be relevant to a different degree; some documents contribute more information to the request, some less without being totally irrelevant. Relevance has been assessed at multiple levels in some studies of operational Boolean systems but even then the levels have been conflated into two categories at the analysis phase for the calculation of precision and recall [e.g. 7, 8, 9]. We therefore do not know how different best match IR methods are able to rank documents of varying relevance levels.

The need for IR methods that are more selective in retrieving highly relevant documents is quite obvious in large databases like those provided by the Internet search services [10, 11]. As more documents become available, also the number of potentially relevant items increases. From the user viewpoint, the major challenge for the IR systems is not how to make difference between relevant and non-relevant documents but rather to separate the highly relevant and potentially relevant documents. In the evaluation of IR systems, this challenge causes a pressure to raise the threshold for what is accepted as relevant, i.e. what is relevant enough.

One interpretation for the degree of relevance is that highly relevant documents tend to convey more information about the topic of interest than marginally relevant ones. From this viewpoint, one may hypothesize that highly relevant documents tend to have following characteristics:

- 1 (a) the topic is discussed in them at length
- (b) they deal with several aspects of the topic
- (c) they have many words pertaining to the topic of the request
- (d) authors use multiple unique expressions to refer to the concepts they discuss in order to avoid tautology.

In contrast, marginal documents mention the topic briefly; just present one aspect or contain just a few words referring to the topic; discuss the topic from a viewpoint not included in the request; no problem of tautology occurs in them. In this paper, we test these hypotheses by analyzing document text characteristics (expressions used and concepts referred to) through the facet analysis technique developed by Sormunen [12, 13].

In best match retrieval, documents are ranked according to scores calculated from the weights of search keys occurring in documents. These weights are typically based on the frequency of a key in a document and on the inverse collection frequency of the documents containing the key (tf.idf weighting) [14]. The development of tf.idf weighting schemes has been based on similar statistical hypotheses of document characteristics as were presented above (hypotheses 1c).

However, we will emphasize in this paper that the analysis of document texts can be elaborated, and further that the effectiveness of best match queries can be improved, especially in retrieving the most valuable documents.

Query structure refers to the syntactic structure of a query expression, marked with query operators and parentheses. Best match queries may either have a structure similar to Boolean queries, or queries may be 'natural language queries' without differentiated relations between search keys. In the former case, concepts are identified (henceforth concept-based or strong structures); in the latter, concepts are not identified, queries are mere sets of search keys, 'natural language queries' (henceforth weak structures). In the mainstream of experimental IR research weak query structures are nearly exclusively employed. However, recent findings have shown the positive influence of concept-based query structuring. For instance, strong query structures improve retrieval performance when queries are expanded [15, 16]. The positive effect of strong query structures seems to hold also for translation dictionary based CLIR [17, 18, 19]. However, in these studies the relevance assessments were dichotomous. In the present study, we investigate the effects of query structures and expansion on retrieving documents at different relevance levels.

Earlier statistical text analyses in the context of experimental IR have focused on the occurrences or co-occurrences of expressions (character strings). The narrow outlook has neglected the conceptual level: how concepts occur and co-occur in document texts in the form of expressions. Concepts are a key issue in finding the link between document characteristics and the effective use of query structures; a key to more effective concept-based retrieval methods.

The conceptual approach to the analysis of document texts helps in predicting the consequences of text characteristics (as stated in hypotheses 1a-d) on the effective querying of highly relevant documents. We already know that concept-based query structuring boosts the positive effects of QE on retrieval performance [15, 16, 20]. Expanded, concept-based queries contain multiple keys to refer to various aspects of a request, and should thus give credit to the highly relevant documents. We therefore may derive hypotheses concerning retrieval performance that

- 2 (a) concept-based expanded queries based on strong structures rank highly relevant documents better in the query results than unexpanded queries or queries based on other structures, whether expanded or not, and

(b) the performance differences between query types among marginally relevant documents should be marginal and among highly relevant documents essential.

This paper seeks to test these hypotheses 1(a-d) and 2(a-b). Section 2 explains the methods and data: the test environment, relevance assessments, document analysis as well as query structures and expansion. Section 3 presents the results and Section 4 the discussion and conclusions.

2. Methods and data

2.1. Test collection

The test environment is a text database containing Finnish newspaper articles operated under the InQuery¹ retrieval system (version 3.1). The database contains 53,893 articles published in three Finnish newspapers. The average article length is 233 words, and typical paragraphs are two or three sentences in length. The database index contains all keys in their morphological basic forms, and all compound words are split into their component words in their morphological basic forms. For the database there is a collection of 35 requests, which are 1-2 sentences long, in the form of written information need statements.

For the requests of the collection there is a recall base of 17,337 articles, which fall into four relevance categories. The base was collected by pooling the result sets of thousands of different queries formulated from the requests in different studies, using both exact match, and partial match retrieval. [12, 20, 21.] In these efforts the consistency of relevance judgments was assessed as inter-assessor consistency as well as intra-assessor consistency by performing partially overlapping and repeated assessments. The average consistencies were 83-89 %. [12, 20].

The recall base of the test collection was constructed by assessing retrieved documents for relevance in the context of an imaginary journalist intending to write an article on the topic of a search request. A four level relevance scale was used to characterize the relevance degree of documents in the simulated work situation [12, 21]:

(0) non-relevant, totally off target,

¹ The InQuery software provided by the Center for Intelligent Information Retrieval, University of Massachusetts Computer Science Department, Amherst, MA, USA, was used in test queries.

- (1) marginally relevant, refers to the topic but does not convey more information than the topic description,
- (2) relevant, contains some new facts about the topic,
- (3) highly relevant, contains valuable information, the article's main focus is on the topic.

The consequence of these efforts is that we have a highly reliable recall base. Moreover, the assessments are given on a four level scale and are highly consistent. Therefore it is possible to study, using the recall base, the characteristics of documents of different degrees of relevance, and their ranking when the query structures and/or content are modified. Two sub-sets of the original test collection were used in this study. The characteristics of document texts were analyzed in a set of 18 requests originally created for another study by Sormunen [12] - called the *text analysis set*. A set of 30 requests amenable for concept-based query expansion - called the *query test set* - were chosen for the retrieval experiment. The latter set was originally created for a study by Kekäläinen [20].

2.2. Facet analysis of document texts

In text retrieval an information need is expressed as a query containing a set of search keys. In exact match – or Boolean – retrieval relations between search keys are marked with Boolean and proximity operators. The query has a structure based on conjunctions and disjunctions of search keys. Similar query structures are supported by some best match IR systems like InQuery [22]. Disjunction and conjunction vaguely mimic syntagmatic and paradigmatic relations of the syntax of natural language [23, 24]. The connection between the query structures and the syntax of natural language provide us a framework to analyze document texts.

The notion of facet is very useful in representing the relationship between query structures and requests as expressed information needs. A facet is a concept (or a family of concepts) identified from, and defining one exclusive aspect of a request or a search topic. A query constructed with the Boolean block search strategy (a query in the conjunctive normal form), is an example of a facet structure. Within a facet, search keys representing one aspect of a request are connected by disjunctions, and facets themselves are connected by conjunctions.

The facets of a request are difficult to identify plausibly in a general operational context by automatic means. On the other hand, the identification and selection of query facets is a routine query planning task for a professional searcher, see e.g. [25, 26, 27]. Query facets are quite consistently identified by experienced searchers assuming that a textual information need

description is available [12, 28]. The test collection used here provides inclusive query plans designed by an experienced search analyst for an earlier study by Sormunen [21]. Thus the list of query facets was available (with information need descriptions as a background) for the analysis of texts.

In the study by Sormunen [12], all relevant (Rel = 2), and highly relevant (Rel = 3) news articles in a sample of 18 requests were analyzed to identify all searchable expressions for all facets of the inclusive query plans. The set of 18 requests was selected so that the complexity and facet broadness² of topics were in line with the whole test collection: six requests were selected from each of the three complexity categories (C = 2-3, C = 4 and C = 5 facets). Within each complexity category, the sample contained three requests for both facet broadness categories (Br≤14 and Br>14 keys per facet). The average number of relevant documents was about the same per request in the sample and in the collection (34 vs. 36 articles).

The document sample contained 236 highly relevant (Rel = 3), and 407 relevant (Rel = 2) news articles. For the present study, additional samples of 371 marginally relevant (Rel = 1) and 261 non-relevant (Rel = 0) documents were analyzed. The samples were created by taking randomly up to 25 (15) marginally relevant (non-relevant, respectively) documents from the pools of relevance judged documents of each of the 18 requests (if available). One should note that the non-relevant documents *were not* randomly selected from the database. They were all retrieved by extensive queries used in recall base estimation and thus contained at least some expressions (or character strings) typical of the relevant documents. The total number of documents analyzed was 1275.

The articles were read by a research assistant and all searchable expressions (nouns, verbs or adjectives) related to some query facet were marked using a different color for each facet. Each occurrence of a searchable expression (single word, compound word or phrase) was recorded. Inflectional word forms were merged as occurrences of a single expression. For instance, if a document contained single words *crisis* and *crises*, and a phrase *economic crisis* for a facet [problem], two occurrences were counted for expression *crisis* and one for *economic crisis*.

² Complexity is the number of query facets identified from a request in inclusive query planning. Respectively, facet broadness is the average number of search keys identified for each search facet of a topic. A detailed description of inclusive query planning method is presented in [12].

From the collected data set it is quite straightforward to calculate figures that characterize the statistical properties of documents needed in testing hypotheses 1b, 1c, and 1d. However, above data do not give a solid basis for measuring the length or the share of text discussing the topic of interest (hypothesis 1a). This characteristic was investigated by analyzing a sample of 308 documents (77 documents for each relevance level). The sample contained up to 5 randomly selected news articles (if available) per relevance level for each of the 18 requests (max 360 documents).

It turned out that average number of text paragraphs per document varied from one set to another: Rel = 0: 11.9; Rel = 1: 14.7; Rel = 2: 11.8, and Rel = 3: 15.5 (maximum 36, 74, 37, 52 / median 11, 12, 11, 14, respectively). To control potential biases in results, a balanced sample was compiled by removing 8 longest documents from sets Rel = 3 and Rel = 1, and an equal number of short documents from the other sets. The number of documents was maintained equal within a topic in this process. The average number of paragraphs per document was balanced between 12.6-13.0, maximum between 36-38 and median between 11-12 paragraphs. The original and balanced samples were used side by side to ensure the validity of results.

Searchable expressions for each query facet were identified similarly as in the original facet analysis but now the focus was on the text paragraphs of news articles. First, the total number of text paragraphs (including titles and subtitles) was counted. Next, the number of text paragraphs containing searchable expressions for at least one, two, ..., five query facets was calculated. On the basis of this data it was possible to reveal the proportion of text paragraphs that were associated with the request.

The statistical significance of differences in document characteristics was tested with the Friedman two-way analysis of variance by ranks [29, 30].

2.3. Query structures and expansion

The InQuery system was chosen for the test, because it has a wide range of operators, including probabilistic interpretations of the Boolean operators [22]. InQuery is based on Bayesian inference networks [31]. All keys are attached with a belief value, which is computed by a modification of the tf.idf weighting. The belief value calculation formula for the SUM operator used in this study is given below:

$$P_{\text{sum}}(Q_1, Q_2, \dots, Q_n) = (p_1 + p_2 + \dots + p_n) / n \quad (1)$$

where P denotes probability, $Q_i, i=1...n$ is either a search key or an InQuery expression, $p_i, i = 1...n$, is the belief value of Q_i [31, 32]. The probability for operands connected by SYN operator is calculated by modifying the tf.idf function so that all instances of synonymous keys are treated as instances of one key [15].

We have earlier tested the co-effects of query structures and query expansion on retrieval performance, and ascertained that the structure of the queries became important when queries were expanded [15, 20]. The best performance overall was achieved with expanded, facet structured queries. For the present study we selected the best weak structure (SUM) and one of the best strong structures, based on concepts (SSYN).

Query expansion was based on conceptual query plans and semantic query expansion, see [20]. In query formulation, search concepts were identified from requests and corresponding search keys were elicited from a test thesaurus. In QE, search keys representing concepts that were semantically related to the original search concepts (by synonymy, hierarchy, or association) in the test thesaurus were added to queries. This procedure gave unexpanded (denoted by u) and expanded (e) query versions, which were both formulated into different query structures.

The structures used to combine the search keys are exemplified in the following. Examples are based on a sample request *The processing and storage of radioactive waste*, and both unexpanded and expanded versions are given.

SUM (average of the weights of keys) queries represent weak structures. An unexpanded SUM query is constructed of the original concepts of a request, and each concept is represented by a single key or a set of keys corresponding to the term but without phrases. In the expansions all expressions are added as single words, i.e., no phrases are included. An example of this structure is given below:

SUM/ u #sum(radioactive waste process storage)

SUM/ e #sum(radioactive waste nuclear waste high active waste low active waste
spent fuel fission product storage store stock repository process refine)

In a SUM-of-synonym-groups-query (SSYN) each search concept forms a clause with the SYN operator. SYN clauses are combined with the SUM operator. Phrases are used, marked with #3,

i.e. the proximity operator allowing a three word distance for the phrase components. All keys within the SYN operator are treated as instances of one key.

SSYN/u #sum(#3(radioactive waste) #syn(process) #syn(storage))

SSYN/e #sum(#syn(#3(radioactive waste) #3(nuclear waste) #3(high active waste)
#3(low active waste) #3(spent fuel) #3(fission product)) #syn(storage store
stock repository) #syn(process refine))

2.4. Test queries and performance measures

The query test set of 30 requests was selected on the basis of their expandability, i.e., they provided possibilities for studying the interaction of query structure and expansion. The recall base includes 366 highly relevant documents (Rel = 3), 700 relevant documents (Rel = 2), 857 marginally relevant documents (Rel = 1). The rest of the database, 51,970 documents, are considered irrelevant (Rel = 0). In order to study the ranking effects separately at each relevance level, separate recall bases were compiled for each relevance level. In other words, relevant documents at the level 1 are listed in a separate recall base from those at the level 2, etc. for each request.

The complexity (number of facets) of the queries ranged from 3 to 5, the average being 3.7. The average coverage (the number of concepts) of the unexpanded queries was 4.9. The broadness (the number of search keys³) of unexpanded queries when no phrases were marked (i.e. SUM structure) was 6.1 on average, and for expanded queries without phrases, on average, 62.3. The broadness with phrases (i.e. SSYN structures) was 5.4 for unexpanded queries, and 52.4 for expanded queries, on average.

Performance was evaluated by using two measures: precision at standard recall levels $R_{0,1} \dots R_{1,0}$, and average non-interpolated precision (AvP) calculated after each retrieved relevant document and averaged over requests. We chose AvP rather than precision based on document cut-off values, because the sizes of recall bases vary at different relevance levels, and thus one cut-off value will not treat queries equally with relation to precision. Effectiveness measures were calculated for each query type at three relevance levels: first, documents with relevance value 1

³ Note the difference between broadness used here (summing the number of search keys across all query concepts and facets) and facet broadness defined in footnote 2 (average number of search keys per facet).

were considered relevant; second, documents with relevance value 2 were relevant; third, documents with relevance value 3 were relevant. The statistical significance of differences in the effectiveness of query types was established with the Friedman two-way analysis by ranks [29].

3 Results

3.1. Document facet analysis results

Our first hypothesis, 1(a), concerning the characteristics of highly relevant documents was that they tend to discuss the topic at length. The facet analysis of documents support this view. Table 1 represents the average numbers of document text paragraphs containing expressions for a given number of query facets. The figures indicate that the higher the relevance level of a document is, the larger is the number of text paragraphs containing searchable expressions for query facets. The differences between different relevance levels maintained although the numbers decrease when co-occurrences of expressions for two or more facets is required. The results from the balanced sample (the average number of paragraphs per documents was fixed) do not differ much from the original sample. The raising trend in the number of paragraphs is even steadier across the recall levels in the balanced sample.

(Table 1)

The Friedman test verified that the observed differences were statistically significant up to the level “ ≥ 3 facets expressed”. Table 1 represents significance results for the original sample, but they are in line with those of the balanced sample. Documents containing expressions for 4 or more query facets within a paragraph were quite rare, and it was not reasonable to make statistical tests for these data sets.

In the balanced sample, highly relevant document contained 1.6 times more paragraphs searchable by at least one facet than a non-relevant document (1.4 and 1.2 times more than marginally relevant or relevant documents, respectively). What is remarkable is that this ratio dramatically increases when the co-occurrence of expressions for multiple different query facets is required. For instance, if searchable expressions are required for two facets, highly relevant documents contain 4.1 times more paragraphs of this type than non-relevant documents (2.3 and

1.5 times more than in marginally relevant and relevant documents, respectively). The difference increases further if expressions are required for three query facets (13.0, 2.6 and 1.7 times more "searchable" paragraphs).

The findings suggest that highly relevant documents contain more paragraphs that discuss different aspects of a topic. The more the co-occurrence of expressions referring to different aspects of the topic is required within a restricted text context (here paragraphs), the more clear the differences between relevance levels become. The difference was most drastic between the non-relevant and highly relevant documents but the same phenomenon could be found between other relevance levels. Standard best match retrieval does not take advantage of passage level evidence, and this was not tested in the our experiment (see below). However, the phenomena of passage level retrieval have been studied earlier, for instance by Salton et al. [33], Callan [34], and Kaszkiel and Zobel [35].

The second hypothesis, 1(b), for the text analysis stated that highly relevant documents deal with several aspects of the request. We may verify this hypothesis by measuring how large a share of the query facets has been referred to in the documents by a searchable expression. Figure 1 illustrates the relative number of aspects covered in different document categories by using two measures. The percentage of explicitly expressed facets (EEF) measures how large a share of query facets per topic has been expressed by a searchable expression in *all* documents.

The EEF measure is interpreted in the following way: If an exact match Boolean query were expanded to cover all alternative search keys for each query facet, roughly two thirds of the available query facets could be used in focusing the query and still retrieve all highly relevant documents. At relevance level 2 only one half of the facets could be applied (and one third at recall level 1, respectively) if full recall is required at these relevance levels. In best match queries, full exhaustivity could be used in strongly structured and expanded queries. Highly relevant documents would benefit in ranking since a larger number of #syn -groups in the expanded queries would contribute to document scores than in less relevant documents.

Unfortunately, the differences were statistically significant only between highly relevant/relevant/marginally relevant and non-relevant documents but not between the former document sets. An obvious reason for the difficulty of achieving statistically significant results is that EEF is measured for each request, and the sample was quite small (18 requests).

(Figure 1)

The second series of columns in Figure 1 measures the coverage of request aspects from the facet angle. The likelihood that a document does contain a searchable expression for a query facet is about 94% and 90% for the set of highly relevant and relevant documents, respectively. Even for the set of marginally relevant documents the likelihood of a searchable expression for a facet is still nearly 85%. The differences between relevance levels are not so clear as above. However, assuming that the occurrences searchable expressions for different facets are independent events, the effect of facet conjunctions can be estimated by taking a product of individual probabilities across the number of facets for which searchable expressions are required to occur in documents. For example, at Rel = 3 the likelihood of documents containing expressions for a conjunction of four facets is $0.937^4 = 0.771$ while remaining at the level of $0.844^4 = 0.507$ for marginally relevant documents. Decreasing likelihood for explicit expressions rapidly begins to punish less relevant documents when the number of facets in the query increases.

Although the differences between relevance levels were here smaller than in the EEF figures, most of them were statistically significant. The Friedman test verified that a highly relevant document more likely contains an expression for a given facet than relevant ($p < 0.001$), marginally relevant ($p < 0.0001$), and non-relevant documents ($p < 0.0001$).

The third and fourth hypotheses were that (1c) highly relevant documents contain more words pertaining to the topic of the request, and that (1d) the author(s) use multiple expressions to refer to the concepts they discuss in order to avoid tautology. These document characteristics were investigated by counting the number of unique expressions and the number of their occurrences represented for each facet in an average document. Figure 2 summarizes the results. The average number of unique expressions per facet increased steadily from 1.42 in non-relevant documents to 3.15 in the highly relevant ones. A similar trend can be observed in the occurrence data. The total number of expression occurrences per facet increases as a function of relevance level from 2.91 to 6.66.

(Figure 2)

The ratio of occurrences per unique expression was steadily around 2 (1.9...2.1) at all recall levels. This was a bit surprising since intuitively one would guess that a word tends to have more occurrences in a highly relevant than in a marginally relevant document. On average, authors have used a word of interest two times per document. This suggests that if we were analyzing

texts only at the level of expressions, we probably would have failed to identify the difference between highly and less relevant documents. The essential difference was observable at the level of concepts, only. Authors use different expressions to refer to the concepts of interest.

Both hypotheses 1(c) and 1(d) got support. For the occurrences of individual expressions the hypothesis does not seem to hold. All differences between relevance levels were statistically significant (Friedman: Rel = 3 >>> Rel = 1/Rel = 0; Rel = 3 > Rel = 2; Rel = 2 >>> Rel = 0; Rel = 2 > Rel = 1, where > = $p < 0.01$; >> = $p < 0.001$; >>> = $p < 0.0001$).

The findings underline the advantages of concept-based query expansion. Highly relevant documents benefit from query expansion since they contain more unique expressions, and get higher scores in weighting since, on average, they match a larger number of search keys per #syn-group than less relevant documents.

3.2. Performance of expanded queries at different relevance levels

The experimental results support our fifth hypothesis, 2(a), that expanded facet- and concept-based structures are most effective in retrieving highly relevant documents. The test also confirms the sixth hypothesis, 2(b), that the performance differences between query types among marginally relevant documents are minor whereas among highly relevant documents they are essential. Figure 3 presents the P-R curves of the four query types at different relevance levels. At the relevance level 1, the curves are almost inseparable. At the relevance level 2, expanded SSYN queries are more effective than the other combinations. At the relevance level 3, the difference is even more accentuated. The higher the relevance level is, the greater are the differences between the best and the worst combinations.

(Figure 3)

In Table 2 the average precision (AvP) figures are given. It can be seen that QE never enhances the average precision of SUM queries. In contrast, QE always improves the average precision of strongly structured queries. When queries are unexpanded the differences in precision are negligible within each relevance level. The best effectiveness over all relevance levels is obtained with expanded SSYN queries. At the best, the difference in average precision between unexpanded SUM and expanded SSYN queries is at the relevance level 3 (AvP: difference of 13.2 percentage units or a change of 51 %). In other words, expanded queries with strong structure are most effective in retrieving the most relevant documents.

(Table 2)

The Friedman test corroborates that the differences in precision figures are more significant at relevance level 3 than at the other relevance levels. Expanded strong queries outperform most often expanded weak queries, but also unexpanded weak and unexpanded strong queries. The number of significant differences is greatest, and the level of significance is highest, at the relevance level 3.

4 Discussion and Conclusions

The goal of this study was to emphasize the importance of the degree of relevance in the evaluation of IR systems. We were able to verify empirically our hypotheses concerning differences in the characteristics of documents at different levels of relevance. We also showed that the statistical differences in document characteristics could be an explanation for good performance in concept-based query expansion. We tested the performance of different query types in regard to different relevance levels. Our hypothesis that the performance differences between query types among marginally relevant documents should be smaller than among highly relevant documents was shown true. Therefore query expansion through strongly structured queries seems more effective than through the usual best match alternatives, weakly structured queries.

It seems that highly relevant documents have the following characteristics: the topic is discussed in them at length; they have more words pertaining to the topic of the request; they deal with several aspects of the topic; the authors use multiple expressions to refer to the concepts they discuss in order to avoid tautology. In contrast, marginal documents may mention the topic briefly; contain just a few words pertaining to the topic; discuss the topic from a viewpoint not included in the request. The number of occurrences per document is about the same for an individual key word (expression) at all relevance levels. As consequence, strongly structured expanded queries, which use multiple keys to refer to various aspects of the request, give credit to the highly relevant documents and push them up in the ranked result list. These findings indicate that there are clear textual phenomena that provide a foundation for concept based QE.

Even at the highest relevance level, only two thirds of the facets were present in all relevant documents. This is a clear support for best match retrieval (albeit facet-based). A Boolean query on all facets would automatically fail in retrieving many highly relevant documents. The strategy

in Boolean retrieval used to fight this problem, reducing the exhaustivity (number of facets) of queries, fails to use the existing evidence *for* relevance in documents for the dropped facets. One has to use several trials in Boolean retrieval when one does not know *which* facets are present in all/most documents. On the contrary, best match methods are able to utilize any evidence for any facet present in a particular document even when the explicitly present facets vary between documents. This is among the ideas behind the best match models. However, our findings show that it works also at the level of concepts / facets.

We also investigated the effects of QE on retrieval performance in regard to different relevance levels. The results reveal that concept-based query expansion with appropriate strong query structures enhances performance in retrieval of highly relevant documents. We verified the earlier findings [15, 16] that query expansion through strongly structured queries is more effective than through the usual best match alternatives, weakly structured queries. Our hypothesis that the performance differences between query types among marginally relevant documents should be smaller than among highly relevant documents was shown true. This finding suggests that concept-based query expansion with strong query structures has a solid basis in the characteristics of highly relevant documents.

The idea of concept-based weighting, i.e. merging the occurrences of all expressions of a concept together, is not a new one. The positive effects of synonym thesaurus were already showed by Salton and Lesk [36] in the vector space model. Hawking, Thistlewaite and Craswell [37] tested concept based scoring in a best match environment. The main idea was that in order to be relevant a document should contain evidence for the presence of all search concepts, not just one. The researchers point out that concept scoring improves the ranking of documents that contain representatives of all search concepts. The problem of focusing the query (or query drift) in QE based on relevance feedback has been discussed by Mitra, Singhal and Buckley [38]. The researchers state that all aspects of a request should be represented in documents assumed to be relevant and used as a source for expansion keys. These studies corroborate our results of the effectiveness of concept-based query structures combined with QE. However, we went further by showing the advantage of query expansion in favoring highly relevant documents and the link between document characteristics and concept-based query expansion.

Finally, we want to make a methodological comment on the use of multi-level relevance judgments in system-oriented IR experiments. We do not really know how stringent the relevance criteria in other studies have been. For instance, judges for TREC might have accepted

as relevant only documents that we would have considered highly relevant. However, that is not probable since making the threshold of acceptance higher requires some extra effort in form of instructions to judges. The simplest way of reminding judges about the degrees of relevance is the use of multi-level relevance categories. On the other hand, the degree of relevance is quite an abstract notion in a general context, and it is difficult to control what aspects of document contents the judges prefer valuable. A more credible approach is to link all requests to a simulated work task, and use experts of that work task as judges.

In this study, we applied traditional formulas to calculate recall and precision for each relevance level. However, the availability of multiple relevance levels raises a question of developing measures that accumulate the total contribution of documents from different relevance levels. One method was proposed and discussed in a recent paper [39].

Acknowledgements. We are grateful for the FIRE research group for the helpful comments and collaboration. The study was funded in part by the Academy of Finland, Research project 44704.

References

1. Froehlich, T.J. Relevance reconsidered – towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3), 1994, 124–134.
2. Harter, S.P. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 1992, 602–615.
3. Saracevic, T. Relevance reconsidered '96. In: Ingwersen, P and Pors, P.O. eds. *Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Copenhagen: The Royal School of Librarianship, 1996, 201–218.
4. Saracevic, T. The stratified model of information retrieval interaction: Extension and applications. In: Schwartz, C. and Rorvik, M. eds. *ASIS '97: Proceedings of the 60th ASIS annual meeting*. Medford, NJ: Information Today, 1997, 313–327.
5. Swanson, D.R. Historical note: information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39(4), 1988, 92–98.
6. Cosijn, E. and Ingwersen, P. Dimensions of relevance. *Information Processing & Management*, 36 (4), 2000, 533–550.
7. Blair, D.C. and Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 1985, 289–299.

8. Lancaster, W., MEDLARS: Report on the Evaluation of Its Operating Efficiency. *American Documentation* 20(2), 1969, 641-664.
9. Saracevic, T., Kantor, P., Chamis, A. and Trivison, D. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3), 1988, 161–176.
10. Gordon, M. and Pathak, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2), 1999, 141–180.
11. Hawking, D., Craswell, P. and Thistlewaite, P. Overview of TREC-7 Very Large Collection Track. In: Voorhees, E.M. and Harman, D.K. eds. *Proceedings of TREC-7*. Available online from: <URL: http://trec.nist.gov/pubs/trec7/papers/vlc_overview.pdf.gz>. [Cited 16/06/00.]
12. Sormunen, E. *A method for measuring wide range performance of Boolean queries in full-text databases*. Doctoral Thesis. Acta Electronica Universitatis Tamperensis, URL: <http://acta.uta.fi/pdf/951-44-4732-8.pdf>. Tampere: University of Tampere, 2000.
13. Sormunen, E. A novel method for the evaluation of Boolean query effectiveness across a wide operational range. In: Belkin, N.J., Ingwersen, P. and Leong, M.-K: eds. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 2000, 25–32.
14. Ingwersen, P. and Willett, P. An introduction to algorithmic and cognitive approaches for information retrieval. *Libri*, 45, 1995, 160–177.
15. Kekäläinen, J. and Järvelin, K. The impact of query structure and query expansion on retrieval performance. In: Croft, W.B., Moffat, A, van Rijsbergen, C. J., Wilkinson, R. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1998, 130–137.
16. Kekäläinen, J. and Järvelin, K. The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval*, 1(4), 2000, 329–344.
17. Ballesteros, L. and Croft, W.B. Resolving ambiguity for cross-language retrieval. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1998, 64–71.
18. Pirkola, A. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Croft, W.B., Moffat, A, van Rijsbergen, C. J., Wilkinson, R. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1998, 55–63.

19. Pirkola, A. *Studies on linguistic problems and methods in text retrieval*. Ph.D. dissertation. Acta Universitatis Tampereensis 672. Tampere: TAJU, 1999.
20. Kekäläinen, J. *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Ph.D. dissertation. Acta Universitatis Tampereensis 678. Tampere: TAJU, 1999.
21. Sormunen, E. *Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa* [Free-text searching efficiency and factors affecting it in a newspaper article database]. VTT Publications 790. Espoo: Technical Research Centre of Finland, 1994. [In Finnish.]
22. Allan, J., Callan, J., Croft, B., Ballesteros, L., Broglio, J., Xu, J. and Shu, H. INQUERY at TREC 5. In: Voorhees, E. M. and Harman, D. K., eds. *Information technology: The Fifth Text Retrieval Conference (TREC-5)*. Gaithersburg, MD: National Institute of Standards and Technology, 1997, 119–132.
23. Green, R. The expression of conceptual syntagmatic relationships: A comparative survey. *Journal of Documentation*, 51(4), 1995, 315–338.
24. Keen, E. M. The use of term position devices in ranked output experiments. *Journal of Documentation*, 47(1), 1991, 1–22.
25. Fidel, R. Searcher's Selection of Search Keys: I. The Selection Routine. II. Controlled Vocabulary of Free-Text Searching. III. Searching Styles. *Journal of the American Society of Information Science*, 42(7), 1991, 490–500, 501–514, 515–527.
26. Harter, S.P. *Online Information retrieval*. Orlando: Academic Press, 1986.
27. Lancaster, F.W. and Warner, A.J. *Information Retrieval Today*. Arlington: Information Resources Press, 1993.
28. Iivonen, M. Searchers and Searchers: Differences Between the Most and Least Consistent Searchers. In: Fox, E.A., Ingwersen, P. and Fidel, R. eds. *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*. New York, NY: ACM, 1995, 149-157.
29. Conover, W.J. *Practical nonparametric statistics* (2nd ed.). New York: John Wiley and Sons, 1980.
30. Siegel, S. and Castellan, N.J. *Nonparametric statistics for the behavioral Sciences*. Singapore: McGraw-Hill, 1998.
31. Turtle, H.R. *Inference networks for document retrieval*. Ph.D. dissertation. Computer and information Science Department, University of Massachusetts. COINS Technical Report 90–92, 1992.

32. Rajashekar, T. B. and Croft, W. B. Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science*, 46(4), 1995, 272–283.
33. Salton, G., Allan, J. and Buckley, C. Approaches to passage retrieval in full text information systems. In: Korfhage, R., Rasmussen, E.M. & Willett, P. eds. *Proceedings of the 16th International ACM/SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1993, 49–58.
34. Callan, J.P. Passage-Level Evidence in Document Retrieval. In: Croft, W.B. and van Rijsbergen, C.J. eds. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. London: Springer, 1994, 302–310.
35. Kaszkiel, M. and Zobel, J. Passage retrieval revisited. In: Belkin, N. J., Narasimhalu, A. D. and Willett, P. eds. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1997, 178–185.
36. Salton, G. and Lesk, M.E. Computer evaluation of indexing and text processing. *Journal of the ACM* 15(1), 1968, 8-36. Reprinted in: Jones, K.P. & Willett, P. eds. *Readings in Information retrieval*. San Francisco: Morgan Kaufmann, 60-84.
37. Hawking, D., Thistlewaite, P. and Craswell, P. ANU/ACSys TREC-6 experiments. In: Voorhees, E.M. and Harman, D.K. eds. *Information technology: The Sixth Text Retrieval Conference (TREC-6)*. Gaithersburg, MD: National Institute of Standards and Technology, 1997, 275–290.
38. Mitra, M., Singhal, A. and Buckley, C. Improving automatic query expansion. In: Croft, W.B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1998, 206–214.
39. Järvelin, K. and Kekäläinen, J. IR evaluation methods for highly relevant documents. In: Belkin, N.J., Ingwersen, P. and Leong, M.-K: eds. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 2000, 41–48.

Number of paragraphs of a document										
Relevance degree	≥ 1 facets expressed *		≥ 2 facets expressed *		≥ 3 facets expressed **		≥ 4 facets expressed ***		5 facets expressed ****	
	Original	Balanced	Original	Balanced	Original	Balanced	Original	Balanced	Original	Balanced
Rel = 0	5.7	6.1	1.3	1.3	0.3	0.2	0.0	0.0	0.0	0.0
Rel = 1	7.7	7.0	2.4	2.3	1.0	1.0	0.3	0.3	0.1	0.1
Rel = 2	7.7	8.2	3.4	3.6	1.4	1.5	0.5	0.5	0.1	0.1
Rel = 3	11.0	9.5	6.0	5.3	2.7	2.6	1.2	1.3	0.6	0.7
Significance	Rel = 3 > Rel = 2 Rel = 3 >> Rel = 1 Rel = 3 >>> Rel = 0 Rel = 2 > Rel = 0		Rel = 3 >>> Rel = 2 Rel = 3 >>>> Rel = 1 Rel = 3 >>>>> Rel = 0 Rel = 2 > Rel = 1 Rel = 2 >>>> Rel = 0 Rel = 1 >>>>> Rel = 0		Rel = 3 > Rel = 2 Rel = 3 >>>> Rel = 1 Rel = 3 >>>>> Rel = 0 Rel = 2 >>>> Rel = 0 Rel = 1 >>>>> Rel = 0		N/A		N/A	

*18 requests, 77/69 docs per recall level; **17 topics, 72/65 docs; ***12 requests, 51/46 docs; ****6 requests, 28/25 docs

> = p < 0.01; >> = p < 0.001; >>> = p < 0.0001

Table 1. The average number and percentage of text paragraphs of a document containing searchable expressions for at least a given number of query facets at varying levels of document relevance.

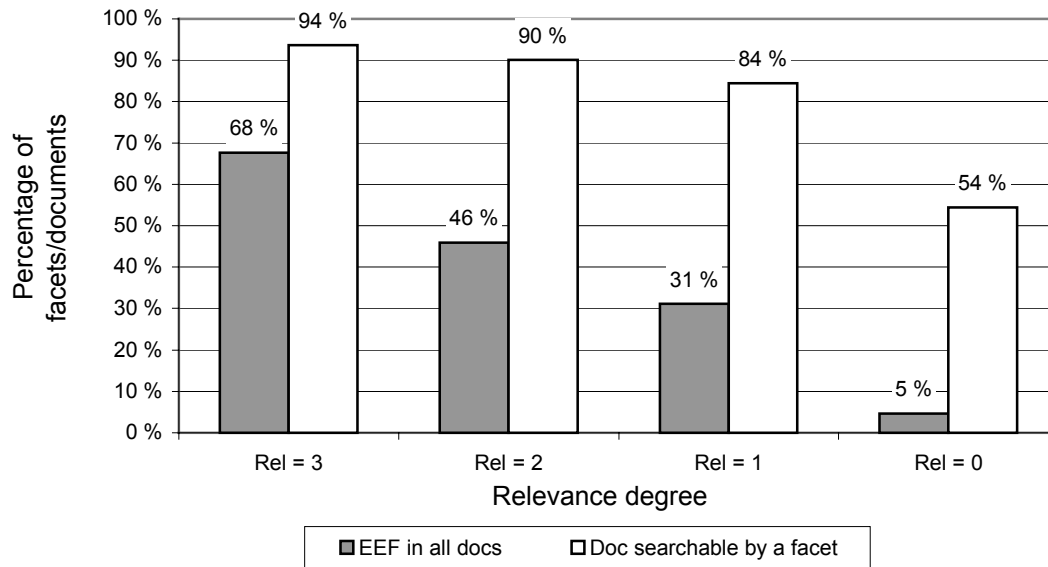


Figure 1. Average percentage of facets of a search topic explicitly expressed in all documents, and the share of documents containing a searchable expression for a given facet (18 topics/71 facets).

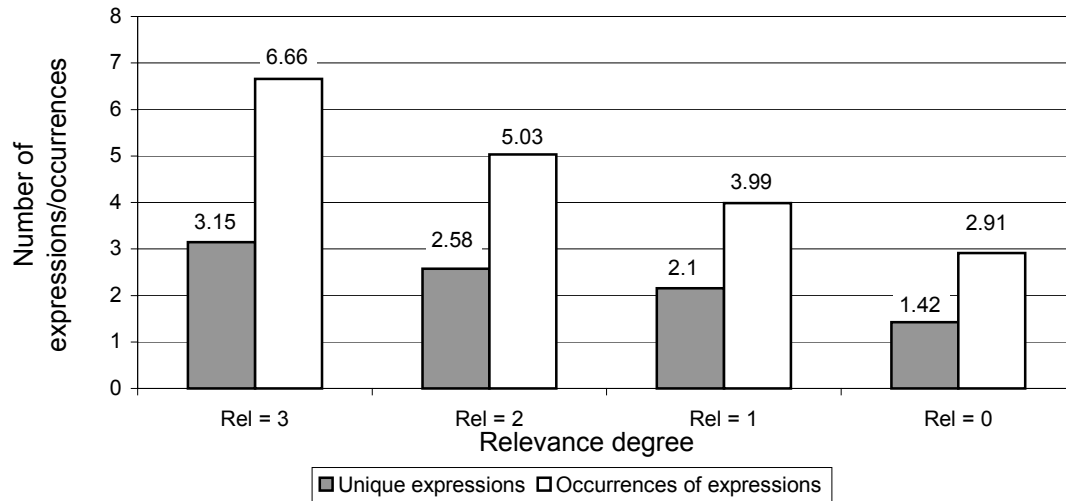


Figure 2. Average number of unique expressions and the sum of their occurrences per query facet in documents of varying relevance degree (18 topics, 1275 documents).

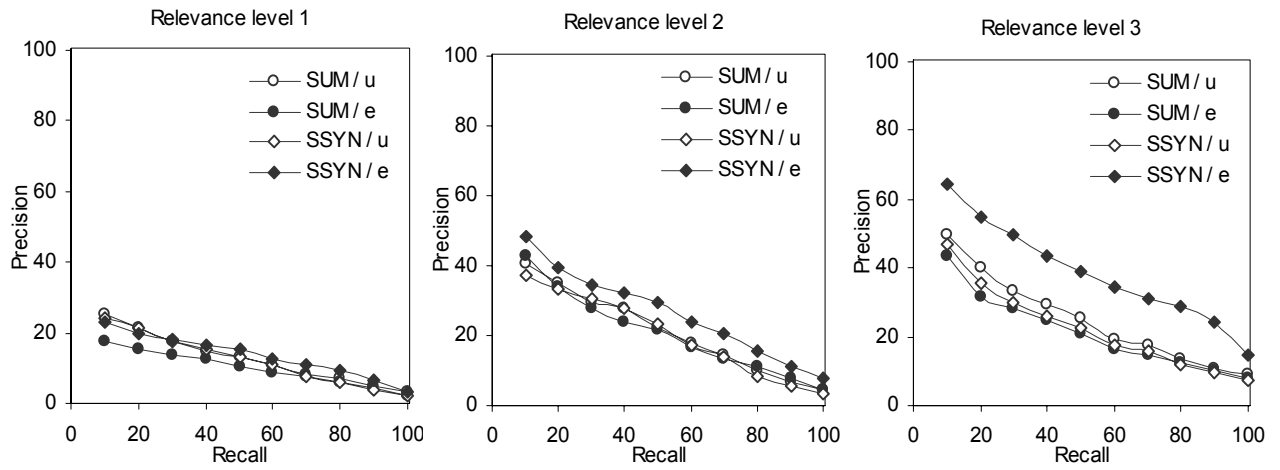


Figure 3. P-R curves of SUM and SSYN queries at relevance levels 1, 2, and 3.

Rel. level	Expansion type	Structure type		Significance
		SUM	SSYN	
1	u	12.8	12.4	SSYN/e > SUM/e
	e	10.1	13.3	
2	u	22.4	21.5	SSYN/e > SUM/u
	e	21.1	27.4	SSYN/e >> SUM/e, SSYN/u
3	u	25.9	23.5	SSYN/e > SUM/u
	e	22.2	39.1	SSYN/e >> SUM/e SSYN/e >>> SSYN/u
> = p < 0.01; >> = p < 0.001; >>> = p < 0.0001				

Table 2. Average non-interpolated precision figures for different query structure and expansion combinations.