

**User-Oriented Evaluation Methods for Information  
Retrieval: A Case Study Based on Conceptual Models for  
Query Expansion**

Jaana Kekäläinen & Kalervo Järvelin  
University of Tampere  
Department of Information Studies  
FIN-33014 University of Tampere  
FINLAND

Email: {[jaana.kekalainen](mailto:jaana.kekalainen@uta.fi), [kalervo.jarvelin](mailto:kalervo.jarvelin@uta.fi)}@uta.fi

Published in: Gerhard Lakemeyer and Bernhard Nebel (Eds.) *Exploring AI in the new millennium*. San Francisco: Morgan Kaufmann Publishers, 2002, 355-379.

# User-Oriented Evaluation Methods for Information Retrieval: A Case Study Based on Conceptual Models for Query Expansion

Jaana Kekäläinen & Kalervo Järvelin

University of Tampere

Department of Information Studies

FIN-33014 University of Tampere

FINLAND

Email: [jaana.kekalainen](mailto:jaana.kekalainen@uta.fi), [kalervo.jarvelin](mailto:kalervo.jarvelin@uta.fi)@uta.fi

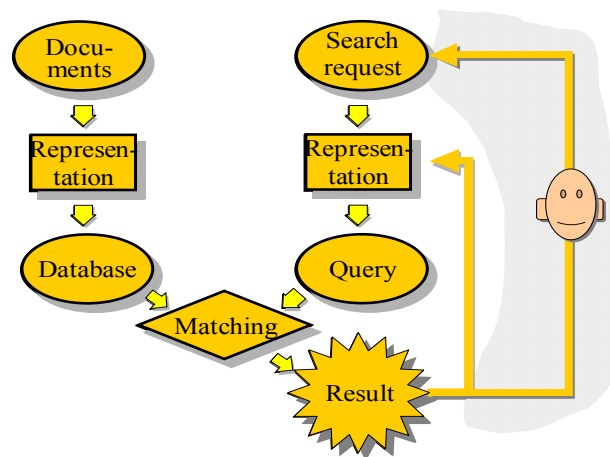
## Abstract

This paper discusses evaluation methods based on the use of non-dichotomous relevance judgements in information retrieval (IR) experiments. It is argued that evaluation methods should credit IR methods for their ability to retrieve highly relevant documents. This is desirable from the user's point of view in modern large IR environments. The proposed methods are (1) a novel application of P-R curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and (2) two novel measures computing the cumulated gain the user obtains by examining the retrieval result up to a given ranked position. We then demonstrate the use of these evaluation methods in a case study on the effectiveness of query types, based on combinations of query structures and expansion, in retrieving documents of various degrees of relevance. Query expansion is based on concepts, which are selected from a conceptual model, and then expanded by semantic relationships given in the model. The test is run with a best match retrieval system (InQuery) in a text database consisting of newspaper articles. The case study indicates the usability of domain dependent conceptual models in query expansion for IR. The results show that expanded queries with a strong query structure are most effective in retrieving highly relevant documents. The differences between the query types are practically essential and statistically significant. More generally, the novel evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable in IR experiments and allow harder testing of IR methods. Proposed methods are user-oriented because users' benefits and efforts – highly relevant documents and number of documents to be examined – are taken into account.

## 1 Introduction

Information retrieval (IR for short) research develops concepts, methods, and systems through which all information – in different forms and places – is easily accessible in forms as convenient as possible for those who need it. IR is concerned with storage and retrieval of – mainly digital – documents. While natural language is the most common means in communication of information, the problems of textual retrieval have been a major issue in research. In this article we shall discuss text-based IR, that is retrieval of text documents or text-based retrieval of multimedia documents.

The reasons why people seek information are called information needs. These needs are varied because persons looking for information have different background knowledge, and situations, tasks or activities from which information needs arise are varied. Information needs may be categorised into (1) verificative, (2) conscious topical, and (3) muddled or ill-defined needs. The first category refers to situation where documents with known properties are sought, e.g. by author name, titles of known authors, etc. The second type implies that the topic is known and definable, but less exact than in the first category. A person looking for information has some level of understanding of it. In the third category are the cases in which a person wishes to find new knowledge and concepts he is not able to describe in detail. (Ingwersen & Willett 1995.)



**Figure 1.** Information storage and retrieval process.

Figure 1 illustrates the information storage and retrieval process. An IR system encompasses the belief that information can be organised and represented for retrieval, the needed information can be described as a set of words. Because this study is about text retrieval, documents are assumed to have textual representations. In a database, documents are represented by their full text, some part of the text or a description in a documentation language, or by some combination of these. The representations – whatever they are – are saved as character strings, say, in an inverted file, which is a typical file structure used in text-based retrieval. A string representing a document is a *key*.

Information needs should be expressed in natural language to be communicable in text retrieval. This formulation is known as a *request*. If a decision is made to use an IR system, the request must be adapted to the conditions of the system; i.e. it must be translated to correspond to the representation of documents. A request translated into a form acceptable for the retrieval mechanism is known as a *query*. Queries consist of *search keys*, which are usually words or phrases represented by character strings, and operators, which express the requirements for the occurrences of the keys. The query language of an IR system defines the operators and a syntax for queries. Queries are matched with the representations of documents.

These representations consist of text keys, which are stored into the index of a database as character strings. Matching refers to matching of character strings, hence, no meanings are involved and talking about matching of search words or terms is somewhat misleading.

Query formulation is not a trivial task, because the information need behind it may be ill-defined. In natural language one idea or subject may be expressed in countless ways. To retrieve all documents that contain potentially relevant<sup>1</sup> information for the information need, one should find all the expressions that have been used to represent that information. Thus, sticking to words obscures the many-to-many relationship between subjects and expressions, or concepts and words.

Non-professional searchers tend to formulate queries with only few search keys. This causes problems because of the variability of expressions in the natural language of documents. *Query expansion* (QE) is a method of adding new search keys to a query to obtain a better correspondence between the query and documents carrying potential information. Expansion keys are usually elicited from the search results of an original (unexpanded) query or some external source, such as vocabularies. (Efthimiadis 1996; Xu & Croft 1996.)

Text retrieval methods may be divided into exact and partial (or best) match methods (Belkin & Croft 1987; Ingwersen & Willett 1995). The former is, in practice, Boolean retrieval, the latter consist of several methods, of which the most prevalent are perhaps probabilistic methods and methods based on the vector space model. In Boolean retrieval a database is divided into two parts: into documents that exactly match the query (the result set, presumed relevant documents), and documents that do not match (presumed non-relevant documents). The query expresses the retrieval conditions in Boolean logic. All documents in the result set are assumed to have equal relevance. In best match retrieval, all documents of the database or documents containing at least one search key are ranked according to their presumed relevance. The scores of documents are calculated from weights given to text keys, and possibly also to search keys. The weighting of the keys is usually based on the frequency of the key in a document and the inverse frequency of the key in the whole database (known as *tf.idf* weighting<sup>2</sup>). (Hersh 1996; Ingwersen & Willett 1995; Salton 1989; Sparck Jones 1972.)

The formulation of queries is different within the different retrieval methods. With *query structure* we refer to the use of operators to express relations between search keys. In Boolean retrieval, operators

---

<sup>1</sup> Relevance is a central concept in IR research, and it is much debated. By relevant information we mean information the searcher wants to retrieve for his information need. In the evaluation of IR methods relevant documents are those which match the topic of the information need. (See Saracevic, 1996; Cosijn & Ingwersen, 2000.)

<sup>2</sup> The abbreviation *tf* means term frequency, i.e. frequency of the key *t* in a document, and *idf* means inverse document frequency, which is often given as  $\log(N/n)$ , where *N* is the number of documents in the collection, and *n* is the number of documents con-

based on Boolean logic are available to indicate conjunctions, disjunctions and negations of search keys, as well as parentheses to mark the order of operations. In best match methods search keys may appear without explicitly marked relations, or the relations may be expressed with operators, which guide the calculation of scores from the weights of the keys. The structure of queries may be described as weak (queries without differentiated relations between search keys) or strong (queries with several operators, different relationships between search keys).

IR systems typically deliver documents containing the searched information rather than direct answers to questions. Yet, the retrieved documents do not contain relevant information to the same degree. In modern large database environments, the number of topically relevant documents to a request may easily exceed the number of documents a user is willing to examine. It would therefore be desirable from the user's viewpoint to rank highly relevant documents highest in the retrieval results and to develop and evaluate IR methods accordingly. In the best match methods the scores given to documents do not reflect the degree of relevance but the probability of relevance (i.e. the probability that a document is either relevant or not, see Robertson & Belkin 1978).

For IR evaluation, documents are typically assessed for relevance and non-relevance. IR methods or systems are then compared according to their ability to retrieve relevant documents and rank them high among retrieved documents. Documents can also be assessed according to the relevance levels. The effects of using multiple relevance levels may be evaluated through traditional IR evaluation methods such as precision-recall (P-R) curves (see Chapter 3). In this paper we apply P-R curves in a new way, focusing on retrieval at each relevance level separately. Moreover, to emphasize the user's viewpoint, we develop new evaluation measures, which seek to estimate the cumulated relevance gain the user receives by examining the retrieval result up to a given rank. These measures facilitate evaluation where IR methods are credited more / only for highly relevant documents. (Järvelin & Kekäläinen, 2000.)

The case demonstrating the effects of multiple degree relevance assessments, and the application of traditional / novel evaluation measures explores query expansion and query structures in probabilistic IR. Kekäläinen and Järvelin (1998; Kekäläinen 1999) have earlier observed that the structure of queries influences retrieval performance when the number of search keys in queries is high, i.e. when queries are expanded. They reported significant retrieval improvements with expanded strongly structured queries. However, in their study the relevance assessments were dichotomous. We therefore do not know how different best match query types (based on expansion and structure) are able to rank documents of varying relevance levels. In the case study we investigate their ability to do this.

---

taining the key  $t$ . NB. We use 'key' rather than 'term' because we reserve 'term' to refer to certain units of documentation languages. A search key, by contrast, may be a natural language word, an abbreviation, a term, etc.

The article is organised as follows: In Chapter 2 we shall discuss conceptual models and concept-based IR, which are needed in the case study; in Chapter 3 we shall present the evaluation methods and compare them to other, related measures in IR, in Chapter 4 the case study. Discussion is in Chapter 5 and Conclusions in Chapter 6.

## 2 Conceptual models

Information science and related fields have employed conceptual models for data and information organisation, representation, classification and retrieval. A conceptual model describes explicitly a domain of interest by giving its concepts, relations between the concepts, and possibly the definitions of the concepts. Conceptual models are referred to by several names, like thesauri, classifications, and ontologies (Soergel, 1999). They differ in their formality, structure and use. Less formal models use natural language or a restricted vocabulary for representing concepts and semantic relations, which may or may be not differentiated. Formal models are expressed in artificial languages (e.g. description logics) which allow meticulous definitions for concepts. (Uschold & Gruninger, 1996.) As a consequence semantic relations between concepts are differentiated and clear because the fuzziness of natural language is eliminated.

Conceptual models are used as unifying frameworks (or shared understanding) for communication between people with different viewpoints and needs, and for inter-operability among systems with different paradigms, languages and software tools. (Guarino, Masolo & Vetere 1998; Uschold & Gruninger 1996.) Next we shall give some examples of the uses of different types of conceptual models.

*Classification and categorisation*<sup>3</sup> are the traditional methods for information organisation. Information items are attached into the best matching category in a predefined set of categories. Classification schemes are kind of conceptual models. They may use numeric notations like the ACM Computing Classification System<sup>4</sup> or terms like Unified Medical Language System<sup>5</sup>. Thesauri are kind of classifications used for document content description. Classification may be intellectual or automatic; the latter does not necessarily require predefined classes, is less expensive but also less reliable and less understandable from human point of view. Chen and Dumais (2000) report an example of automatic categorisation of WWW search results. The process involves a training phase and an operational phase. During the training phase, WWW pages with known – and predefined – category labels are used to train the categoriser. During the operational phase, the learned system is used to categorise new WWW pages on-the-fly.

---

<sup>3</sup> Classification and categorisation are not synonyms though often used as such (see Jacob, 1991). A rough distinction is that classes are mutually exclusive, but categories are necessarily not.

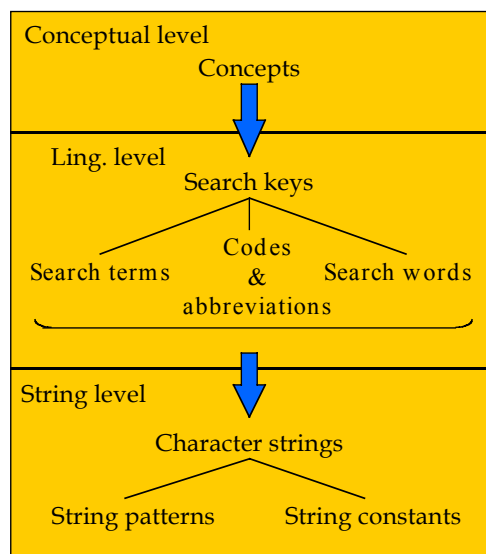
<sup>4</sup> CCS, see <http://www.acm.org/class/1998/>.

<sup>5</sup> UMLS, see <http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML>.

*Inference.* Conceptual models including conceptual relations have clear and explicit semantics that can be reasoned over. An obvious example is the use of hierarchical relations – if a document is about low-active waste, a conceptual model could state that then it also is about nuclear waste and radioactive waste. Ontologies typically have three major components that can be used in inference: a taxonomy (i.e. a generic classification with mutually exclusive classes), relations between concepts and axioms for the relations (Bechhofer & al. 2001).

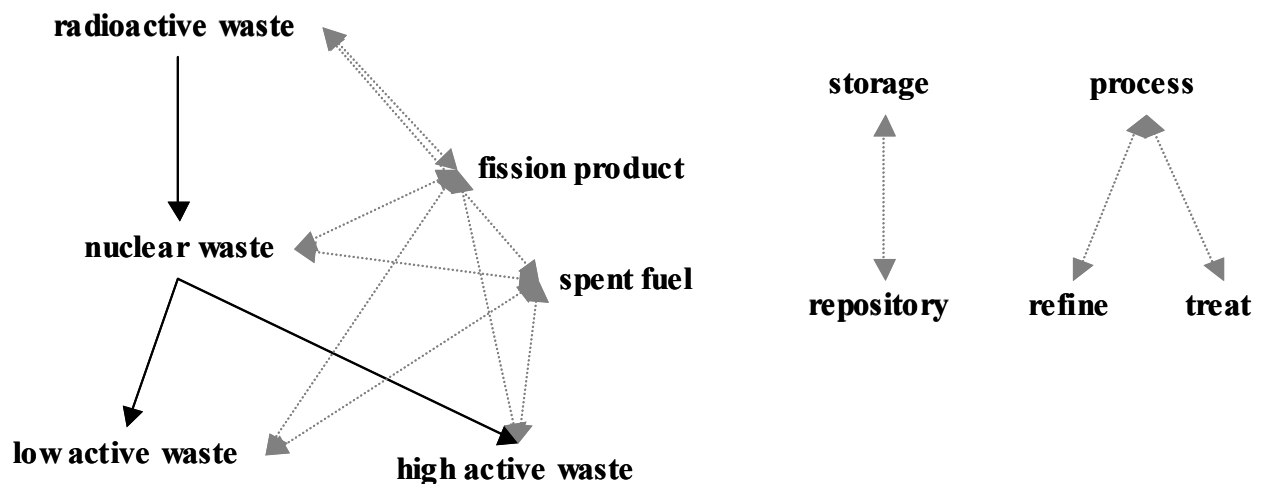
*Personalization.* Because information needs and information seeking situations vary from user to user, IR systems should support users in organising (retrieved) information according to their personal needs (worldview). Chaffee and Gauch (2000) report of an IR system that allows users to build their own hierarchical concept trees, and maps a reference ontology to these personal models. The system classifies WWW sites with concepts from the reference ontology and presents users the sites through their own model.

*Concept-based IR* seeks to rise from the level of search keys to the level of concepts; i.e. searchers should be able to express their information need in concepts rather than search keys. The IR system should support searchers in search key selection, query formulation and expansion. Our aim is to equip the searcher with a conceptual model representing semantic relationships among concepts, and giving for each concept a set of search strings that may represent concepts in different search environments. The thesaural structure controlling hierarchies, associative relations and synonymy suits well for this kind of conceptual model. The model is managed by a tool that supports (1) searchers to automatically construct and expand effective queries without prior understanding about query structures and their interaction with expansion in various retrieval environments, and (2) QE experimentation with query structures, expansion and other query construction parameters.



**Figure 2.** The abstraction levels of query formulation.

The conceptual model is based on deductive data model and three abstraction levels (see Figure 2): the conceptual level, the linguistic level and the string level (Järvelin & al. 1996; 2001). The conceptual level represents concepts and conceptual relationships (e.g. hierarchical generic and partitive relationship, association relationship). The linguistic level represents natural language expressions for the concepts and equivalence relations between them. Each concept may have several synonymous expressions with varied reliability. This is comparable to family resemblance or fuzzy membership: some expressions are more typical names for a concept, but others may be used as well. Each concept is denoted by a term (the principal name of the concept) and possibly a number of synonyms<sup>6</sup>. Concepts are identified through an identification code and the relations they have. One or more strings represent each expression – also of varied reliability – at the string level. Each string is a matching pattern representing how the expression may be matched in database indices built in various ways (e.g. with or without compound words split into component words, and with or without stemming<sup>7</sup>). An example of the conceptual relations is given in Figure 3.



**Figure 3.** A sample of conceptual relations

(hierarchical relations shown with black arrows, associative relations with grey dashed arrows).

### 3 Evaluation methods employing multiple degree relevance assessments

#### 3.1 Precision as a function of recall

Fundamental problems of IR experiments are linked to the assessment of relevance. In most laboratory tests documents are judged relevant or irrelevant with regard to the request. However, binary relevance cannot reflect the possibility that documents may be relevant to a different degree; some documents con-

---

<sup>6</sup> Synonymy is understood rather loosely here (cf. Miller 1995). Possible candidates are synonyms, quasisynonyms, corresponding verbs / nouns - a term is not necessarily a noun - and common names.

tribute more information to the request, some less without being totally irrelevant. In some studies relevance judgements are allowed to fall into more than two categories, but only a few tests actually take advantage of different relevance levels (e.g. Hersh & Hickam, 1995). More often relevance is conflated into two categories at the analysis phase because of the calculation of precision and recall (e.g. Blair & Maron, 1985; Saracevic & al., 1985).

*Recall* is defined as

$$recall = \frac{\text{number of relevant documents retrieved}}{\text{total number of relevant documents in collection}}$$

*Precision* is defined as

$$precision = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}}$$

(Lancaster 1986.)

Retrieval evaluation results achieved by precision and recall are often presented as an average precision-recall curve (*P-R curve*), which measures the precision at the same fixed levels of recall for all requests and averages the results. In the average P-R curve precision and recall are used as a bivariate measure of retrieval effectiveness. Recall is defined as an independent variable and precision as a dependent variable, since precision is averaged at the fixed levels of recall. Because precision cannot be exactly defined at all fixed recall levels, interpolation is needed. Further, 100 per cent recall is not reached by every algorithm or query, thus, P-R curve is not faithful to actual data points. The assumptions attached to the P-R curve are, that a particular level of recall must be attained by every request, and the best method is the one that reaches this level with fewest number of non-relevant documents. The P-R curve provides no information about the number of documents that have to be retrieved in order to reach a given recall level. Because requests have different number of relevant documents, a recall level of, say, 50% may mean a result set of 20 or 200 documents for different requests. (Keen, 1992; Tague-Sutcliffe, 1992; Hull, 1993.)

The current practice of liberal binary assessment of topical relevance gives equal credit for a retrieval method for retrieving highly and fairly relevant documents. Therefore differences between sloppy and excellent retrieval methods may not become apparent in evaluation. In order to see the difference in performance between retrieval methods, their performance should be evaluated separately at each relevance level. For example, in case of a four point assessment (say, 0 to 3 points), separate recall bases<sup>8</sup> are needed for highly relevant documents (relevance level 3), fairly relevant documents (relevance level 2),

---

<sup>7</sup> Stemming means cutting off affixes in order to bring out the word stem (see Alkula 2001).

<sup>8</sup> In ideal IR experiments whole test collection is judged for relevance, i.e. the relevance of each document in the collection in relation to every request is known. If the number of documents in the collection is large, it may be impossible to judge the relevance of all documents. Then, recall should be estimated. For this, the number of relevant documents in the collection for each test request should be estimated. We refer to the estimation of the total number of relevant documents for all test requests as a *recall base*.

and marginally relevant documents (relevance level 1). In this study, we compiled the recall bases for P-R curve computation in this way.

### 3.2 Cumulated gain -based measurements

When examining the ranked result list of a query, it is obvious that:

1. highly relevant documents are more valuable than marginally relevant documents, and
2. the greater the ranked position of a relevant document (of any relevance level) the less valuable it is for the user, because the less likely it is that the user will examine the document.

Point one leads to comparison of IR methods through test queries by their cumulated gain by document rank. In this evaluation, the relevance level of each document is somehow used as a gained value measure for its ranked position in the result and the gain is summed progressively from position 1 to  $n$ . Thus the ranked document lists (of some determined length) are turned to gained value lists by replacing document IDs by their relevance values. Assume that the relevance values 0 - 3 are used (3 denoting high value, 0 no value). Turning document lists up to rank 200 to corresponding value lists gives vectors of 200 components each having the value 0, 1, 2 or 3. For example:

$$G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$$

The cumulated gain at ranked position  $i$  is computed by summing from position 1 to  $i$  when  $i$  ranges from 1 to 200. Formally, let us denote position  $i$  in the gain vector  $G$  by  $G[i]$ . Now the cumulated gain vector  $CG$  is defined recursively as the vector  $CG$  where:

$$CG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i - 1] + G[i], & \text{otherwise} \end{cases} \quad (1)$$

For example, from  $G'$  we obtain  $CG' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots \rangle$ . The cumulated gain at any rank may be read directly, e.g. at rank 7 it is 11.

Point two leads to comparison of IR methods through test queries by their cumulated gain based on document rank with a rank-based discount factor: the greater the rank, the smaller share of the document value is added to the cumulated gain. The greater the ranked position of a relevant document – of any relevance level – the less valuable it is for the user, because the less likely it is that the user will examine the document due to time, effort, and cumulated information from documents already seen. A discounting function is needed which progressively reduces the document value as its rank increases but not too steeply (e.g. as division by rank) to allow for user persistence in examining further documents. A simple way of discount-

ing with this requirement is to divide the document value by the log of its rank. For example  $^2\log 2 = 1$  and  $^2\log 1024 = 10$ , thus a document at the position 1024 would still get one tenth of its face value. By selecting the base of the logarithm, sharper or smoother discounts can be computed to model varying user behaviour. Formally, if  $b$  denotes the base of the logarithm, the cumulated gain vector with discount DCG is defined recursively as the vector DCG where:

$$DCG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ DCG[i-1] + G[i] / b^{\log i}, & \text{otherwise} \end{cases} \quad (2)$$

Note that we must not apply the logarithm-based discount at rank 1 because  $b^{\log 1} = 0$ .

For example, let  $b = 2$ . From  $G'$  we obtain  $DCG' = \langle 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \dots \rangle$ .

The (lack of) ability of a query to rank highly relevant documents toward the top of the result list should show on both the cumulated gain by document rank (CG) and the cumulated gain with discount by document rank (DCG) vectors. By averaging over a set of test queries, the average performance of a particular IR method can be analysed. Averaged vectors have the same length as the individual ones and each component  $i$  gives the average of the  $i$ th component in the individual vectors. The averaged vectors can directly be visualised as gain-by-rank –graphs (see Section 4.6).

The actual CG and DCG vectors by a particular IR method may also be compared to the theoretically best possible. The latter vectors are constructed as follows. Let there be  $k$ ,  $l$ , and  $m$  relevant documents at the relevance levels 1, 2 and 3 (respectively) for a given request. First fill the vector positions  $1 \dots m$  by the values 3, then the positions  $m+1 \dots m+l$  by the values 2, then the positions  $m+l+1 \dots m+l+k$  by the values 1, and finally the remaining positions by the values 0. Then compute CG and DCG as well as the average CG and DCG vectors and curves as above. Note that the curves turn horizontal when no more relevant documents (of any level) can be found. They do not unrealistically assume as a baseline that all retrieved documents could be maximally relevant. The vertical distance between an actual (average) (D)CG curve and the theoretically best possible curve shows the effort wasted on less-than-perfect documents due to a particular IR method.

### 3.3. Comparison of (D)CG measures to related measures

The novel measures have several advantages when compared with several previous and related measures. The *average search length* (ASL) measure (Losee 1998) estimates the average position of a relevant document in the retrieved list. The *expected search length* (ESL) measure (Korfhage 1997; Cooper 1968)

is the average number of documents that must be examined to retrieve a given number of relevant documents. Both are dichotomical, they do not take the degree of document relevance into account. The former also is heavily dependent on outliers (relevant documents found late in the ranked order).

The normalised recall measure (NR for short; Rocchio 1966; Salton & McGill 1983), and the satisfaction – frustration – total measure (SFT for short; Myaeng & Korfhage 1990; Korfhage 1997) all seek to take into account the order in which documents are presented to the user. *The NR measure* compares the actual performance of an IR technique to the ideal one (when all relevant documents are retrieved first). Basically it measures the area between the ideal and the actual curves. NR does not take the degree of document relevance into account and is highly sensitive to the last relevant document found late in the ranked order.

The *SFT measure* consists of three components. The satisfaction measure only considers the retrieved relevant documents, the frustration measure only the irrelevant documents, and the total measure is a weighted combination of the two. SFT assumes the same retrieved list of documents, which are obtained in different orders by the IR techniques to be compared. This is an unrealistic assumption for comparison since for any retrieved list size  $n$ , when  $n \ll N$  (the database size), different IR techniques may retrieve quite different documents – that is the whole idea (!). A strong feature of SFT comes from its capability of punishing an IR technique for retrieving irrelevant documents while rewarding for the relevant ones. SFT does not have the discount feature of our DCG measure.

The relative relevance and ranked half life measures (Borlund & Ingwersen 1998; Borlund 2000) were developed for interactive IR evaluation. The *relative relevance* (RR for short) measure is based on comparing the match between the system-dependent probability of relevance and the user-assessed degree of relevance, the latter by the real person-in-need or a panel of assessors. The match is computed by the cosine coefficient (Borlund 2000) when *the same* ranked IR technique output is considered as vectors of relevance weights as estimated by the technique, by the user, or by the panel. RR is (intended as) an association measure between types of relevance assessments, and is not directly a performance measure. Of course, if the cosine between the IR technique scores and the user relevance assessments is low, the technique cannot perform well from the user point of view. The ranked order of documents is not taken into account.

The *ranked half life* (RHL for short) measure gives the median point of accumulated relevance for a given query result. It thus improves on ASL by taking the degree of document relevance into account. Like ASL, RHL is dependent on outliers. The RHL may also be the same for quite differently performing queries. RHL does not have the discount feature of DCG.

The strengths of the proposed CG and DCG measures can now be summarized as follows:

- They combine the degree of relevance of documents and their rank (affected by their probability of relevance) in a coherent way.
- At any number of retrieved documents examined (rank), CG and DCG give an estimate of the cumulated gain as a single measure no matter what is the recall base size.
- They are not heavily dependent on outliers (relevant documents found late in the ranked order) since they focus on the gain cumulated from the beginning of the result up to any point of interest.
- They are obvious to interpret, they are more direct than P-R curves, and do not mask bad performance.

In addition, the DCG measure has the following further advantages:

- It realistically weights down the gain received through documents found later in the ranked results.
- It allows modelling user persistence in examining long ranked result lists by adjusting the discounting factor.

The measures considered above, both the old and the new ones, have weaknesses in two areas. Firstly, none of them take into account order effects on relevance judgements, or document overlap (or redundancy). In the TREC interactive track (Over 1999), *instance recall* is employed to handle this. The user-system pairs are rewarded for retrieving distinct instances of answers rather than multiple overlapping documents. In principle, the (D)CG measures may be used for such evaluation. Secondly, the measures considered above all deal with relevance as a single dimension while it really is multidimensional (Vakkari & Hakala 2000). In principle, such multidimensionality may be accounted for in the construction of recall bases for search topics but leads to complexity in the recall bases and in the evaluation measures. Nevertheless, such added complexity may be worth pursuing because so much effort is invested in IR evaluation.

#### **4 Case study: the effectiveness of QE and query structures at different relevance levels**

We demonstrate the use of the proposed measures in a case study testing the co-effects of query expansion and structured queries in a database with non-binary relevance judgements. Based on the results by Kekäläinen and Järvelin (1998) we already know that weak query structures are not able to benefit from query expansion whereas the strong ones are. In the present study we shall test whether the performance of differently structured queries varies with relation to the degree of relevance. We give the results as traditional P-R curves for each relevance level, and as CG and DCG curves which exploit the degrees of relevance. We hypothesize that expanded queries based on strong structures are better able to rank highly relevant documents high in the query results than unexpanded queries or queries based on other structures, whether expanded or not. Consequently, the performance differences between query types among margin-

ally relevant documents should be marginal and among highly relevant documents essential. Expanded queries based on strong structures should cumulate higher CG and DCG values than unexpanded queries or queries based on other structures, whether expanded or not.

#### 4.1 Test environment

The test environment was a text database containing newspaper articles operated under the InQuery retrieval system (version 3.1). The database contains 53,893 articles published in three different newspapers. 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 requests, which are 1 - 2 sentences long, in the form of written information need statements. For these requests there is a recall base of 16,540 articles which fall into four relevance categories (see below *Relevance assessments*). The base was collected by pooling the result sets of hundreds of different queries formulated from the requests in different studies, using both exact and partial match retrieval. We thus believe that our recall estimates are valid. For a set of tests concerning query structures, 30 requests were selected on the basis of their expandability, i.e. they provided possibilities for studying the interaction of query structure and expansion. (Kekäläinen, 1999; Kekäläinen & Järvelin, 2000; Sormunen, 2000.)

The InQuery system was chosen for the test, because it has a wide range of operators, including probabilistic interpretations of the Boolean operators, and it allows search key weighting. Moreover, InQuery has shown good performance in several tests (e.g. Allan et al. 1997; Harman 1995; Xu & Croft 1996). InQuery is based on Bayesian inference networks (see, Allan et al. 1997; Turtle 1990). All keys are attached with a *belief value*, which is approximated by the following tf.idf modification:

$$0.4 + 0.6 * \left( \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 * \left( \frac{dl_j}{adl} \right)} \right) * \left( \frac{\log \left( \frac{N + 0.5}{df_i} \right)}{\log(N + 1.0)} \right) \quad (3)$$

where  $tf_{ij}$  = the frequency of the key  $i$  in the document  $j$

$dl_j$  = the length of document  $j$  (as a number of keys)

$adl$  = average document length in the collection

$N$  = collection size (as a number of documents)

$df_i$  = number of documents containing key  $i$ .

The InQuery query language provides a set of operators to specify relations between search keys. As with Boolean operators it is possible to formulate structured queries, and mark relationships between concepts. The probabilistic interpretations for the operators used in this study are given below:

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

$$P_{wsum}(w_s, w_1Q_1, w_2Q_2, \dots, w_nQ_n) = w_s(w_1p_1 + w_2p_2 + \dots + w_np_n) / (w_1 + w_2 + \dots + w_n)$$

where  $P$  denotes probability,  $Q_i$  is either a key or an InQuery expression,  $p_i, i = 1 \dots n$ , is the belief value of  $Q_i$ ,  $w_i, i = 1 \dots n$ , is the weight of  $Q_i$ , and  $w_s$  is a weight given for a clause (Rajashekar & Croft 1995; Turtle 1990).

The probability for operands connected by SYN operator is calculated by modifying the tf.idf function as follows:

$$0.4 + 0.6 * \left( \frac{\sum_{i \in S} tf_{ij}}{\sum_{i \in S} tf_{ij} + 0.5 + 1.5 * \left( \frac{dl_j}{adl} \right)} \right) * \left( \frac{\log \left( \frac{N + 0.5}{df_s} \right)}{\log(N + 1.0)} \right) \quad (4)$$

where  $tf_{ij}$  = the frequency of the key  $i$  in the document  $j$

$S$  = a set of search keys within the SYN operator

$dl_j$  = the length of document  $j$  (as a number of keys)

$adl$  = average document length in the collection

$N$  = collection size (as a number of documents)

$df_s$  = number of documents containing at least on key of the set  $S$ .

## 4.2 Relevance assessments

For the test requests and test collection of the present experiment, relevance was assessed by four persons, two experienced journalists and two information specialists. They were given written information need statements (requests), and were asked to judge the relevance on a four level scale: (0) irrelevant, the document is not about the subject of the request, (1) marginally relevant, the topic of the request is mentioned, but only in passing, (2) fairly relevant, the topic of request is discussed briefly, (3) highly relevant, the topic is the main theme of the article. The relevance of 20 requests (of 30) was assessed by two (one by three) persons, the rest by one person. The assessors agreed in 73% of the parallel assessments, in 21% of the cases the difference was one point, and in 6% two or three points. If the difference was one point, the assessment was chosen from each judge in turn. If the difference was two or three points, the article was checked by the researcher to find out if there was a logical reason for disagreement, and a more plausible alternative was selected. (Kekäläinen, 1999; Sormunen, 2000.)

The recall bases for the 30 requests of the present study includes 366 highly relevant documents (relevance level 3), 700 fairly relevant documents (relevance level 2), 857 marginally relevant documents (relevance level 1). The rest of the database, 51,970 documents, is considered irrelevant (relevance level 0).

### **4.3 Query structures and expansion**

In text retrieval an information need is typically expressed as a set of search keys. In exact match – or Boolean – retrieval relations between search keys in a query are marked with the AND operator, the OR operator, or proximity operators which, in fact, are stricter forms of the AND operator. Thus, the query has a structure based on conjunctions and disjunctions of search keys. (Green, 1995; Keen, 1991.) 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 with the OR operator, and facets are connected with the AND operator. A facet may consist of one or several concepts. In best match retrieval, queries may either have a structure similar to Boolean queries, or queries may be ‘natural language queries’ without differentiated relations between search keys.

Kekäläinen and Järvelin (1998) 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. The best performance overall was achieved with expanded, facet structured queries. For the present study, we selected their best weak structure (SUM) and two of their best strong structures, one based on concepts (SSYN-C) and another based on facets (WSYN). SUM queries may be seen as typical ‘best match’ queries and therefore suitable as a baseline.

We formulated and expanded queries using a conceptual model with a thesaural structure. Since the test database includes newspaper articles, we needed a conceptual model for this domain to test QE based on semantic relationships. No such model was available, thus, we chose to construct a test model. The collection of concepts was started by identifying all concepts from the test requests. Then, for each of these concepts all plausible hierarchically narrower and associatively related concepts were collected. Consideration was given to the completeness of hierarchies. In the organisation of concept relations, concepts were treated independently of the context of the requests. However, the newspaper domain guided the selection of concepts and relations. For each concept, all plausible expressions were gathered, and these expressions were turned into search strings. The conceptual model was constructed by three persons using dictionaries, handbooks, primary literature and their own knowledge. The relations between concepts and between expressions are valid for the whole domain, i.e. they are standard thesaurus or semantic relations. The test model was aimed at QE in a database of Finnish newspaper articles, thus, its language was Finnish. (Sormunen 1994; Kekäläinen 1999.)

The conceptual model includes 832 concepts, 1,345 expressions for the concepts, and 1,558 search strings for the expressions. Concepts have hierarchic (generic, partitive and instance) and association relationships. Expressions representing concepts are each other's synonyms or quasi-synonyms, i.e. equivalence exists between expressions at the linguistic level. The most typical or obvious of the synonyms, in a linguistic sense, was chosen as a principal expression, *term*, of the concept. Several strings, which are spelling variants, may represent each expression. If these strings are phrases, they are formed with different proximity operators. There is no variation caused by word truncation because the words are in their basic forms in the database index. The conceptual model is a database managed with the ExpansionTool, which is a tool for concept based query construction and expansion (see Järvelin et al. 1996; 2001).

In query formulation, researchers identified search concepts from requests and elicited corresponding search keys from the conceptual model. In a practical setting, users would select the concepts of interest from the model by themselves, or their search keys would be mapped to the model. In QE, search keys that were semantically related (synonyms, hierarchies, associations) to the original search concepts in the test model were added to queries. This procedure gave unexpanded (u) and expanded (e) query versions, which both were 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*. In the following samples queries are expanded, the expressions of the unexpanded queries are in italics.

*SUM* (average of the weights of keys) queries represent weak structures. In these queries search keys are single words, i.e. no phrases are included.

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-C*) each search concept forms a clause with the SYN operator. SYN clauses were combined with the SUM operator. Phrases were used (marked with #3). All keys within the SYN operator are treated as instances of one key.

SSYN-C/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))

WSYN queries were similar to SSYN, but based on facets instead of concepts. Facets were divided into major and minor facets according to their importance for the request. In WSYN queries, the weight of major facets was 10 and of minor facets 7.

```

WSYN/e
#wsum(1 10 #syn(#3(radioactive waste) #3(nuclear waste) #3(high active waste)
#3(low active waste) #3(spent fuel) #3(fission product))
7 #syn(storage store stock repository process refine))

```

#### 4.4 Test queries and the application of the evaluation measures

In the queries for the 30 test requests, the average number of facets was 3.7. The average number of concepts in unexpanded queries was 4.9, and in expanded queries 26.8. The number of search keys 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 number of search keys with phrases (i.e. SSYN-C, and WSYN structures) was 5.4 for unexpanded queries, and 52.4 for expanded queries, on average.

The length of relevant documents at all relevance levels exceeded the average length of documents in the database (233 words). However, the documents at relevance level 3 were, on average, shorter than documents at relevance levels 2 or 1. The average document lengths were 334 words at relevance level 1; 314 words at level 2; and 306 words at level 3. Because the differences in average document lengths are minor, highly relevant documents did not gain from higher document length.

We present the analysis of the search results in two forms: First, we apply the conventional measures in the form of P-R curves. We also calculated precision after each retrieved relevant document and took an average over requests (average non-interpolated precision, AvP for short). 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. The statistical significance of differences in the effectiveness of query types was established with the Friedman test (see Conover, 1980).

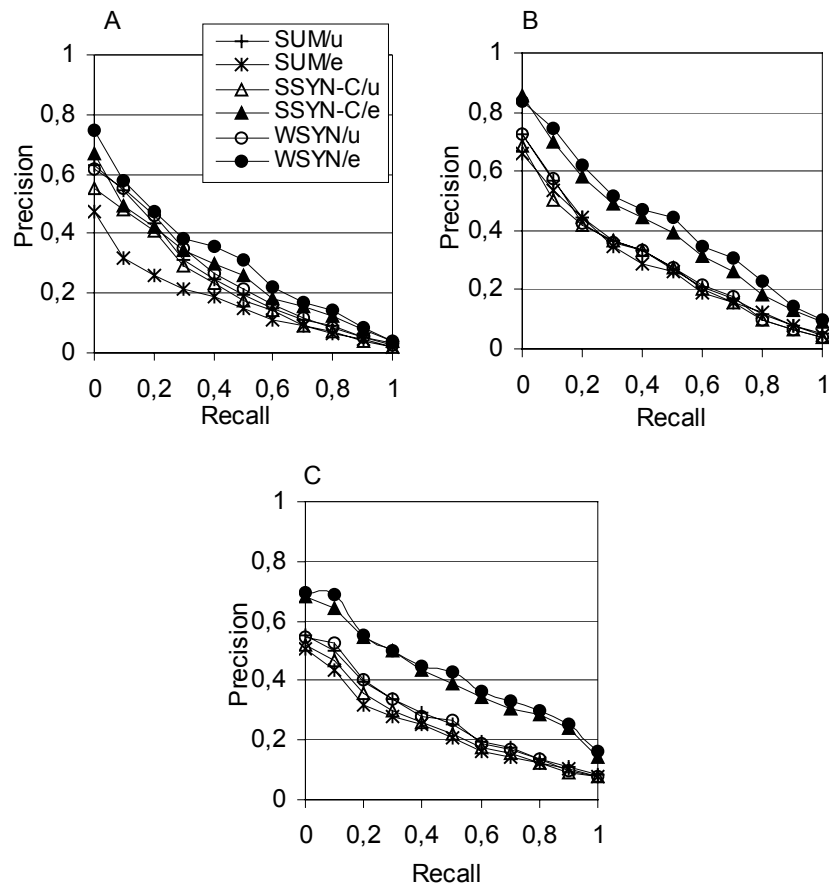
Second, we present the CG and DCG curves. For the cumulated gain evaluations we tested the same query types in separate runs with the logarithm bases and the handling of relevance levels varied as parameters as follows:

- The logarithm bases 2,  $e$ , and 10 were tested for the DCG vectors. The base 2 models impatient users, base 10 persistent ones. For brevity, we show only the results obtained with the base 2 in Section 4.7.

- We used document relevance levels 0 - 3 directly as gained value measures. This can be criticised, e.g. by asking whether a highly relevant document is (only) three times as valuable as a marginally relevant document. Nevertheless, even this gives a clear difference for document quality to look at.
- We first took all documents at relevance levels 1 - 3 into account, secondly nullified the values of documents at relevance level 1 (to reflect that they practically have no value), and finally nullified the values of documents at relevance levels 1 - 2 in order to focus on the highly relevant documents.
- The average actual CG and DCG vectors were compared to the theoretically best possible average vectors.

#### 4.5 P-R curves and average precision

Figure 4 presents the P-R curves of the six query types at different relevance levels. At the relevance level 1, the curves are almost inseparable. At the relevance level 2, expanded WSYN and SSYN-C queries are more effective than the other query types. 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 query types.



**Figure 4.** P-R curves of SUM, SSYN-C, and WSYN queries at relevance levels 1 (A), 2 (B), and 3 (C).

In Table 1 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 WSYN queries. At the best, the difference in average precision between unexpanded SUM and expanded WSYN queries is at the relevance level 3 (AvP: a change of 15.1 percentage units or an improvement of 58.3 %). In other words, expanded queries with strong structure are most effective in retrieving the most relevant documents.

Average non-interpolated precision				
Rel. level	Expansion type	Structure type		
		SUM	SSYN-C	WSYN
1	u	12.8	12.4	13.8
	e	10.1	13.3	14.3
2	u	22.4	21.5	22.9
	e	21.1	27.4	29.3
3	u	25.9	23.5	25.7
	e	22.2	39.1	41.0

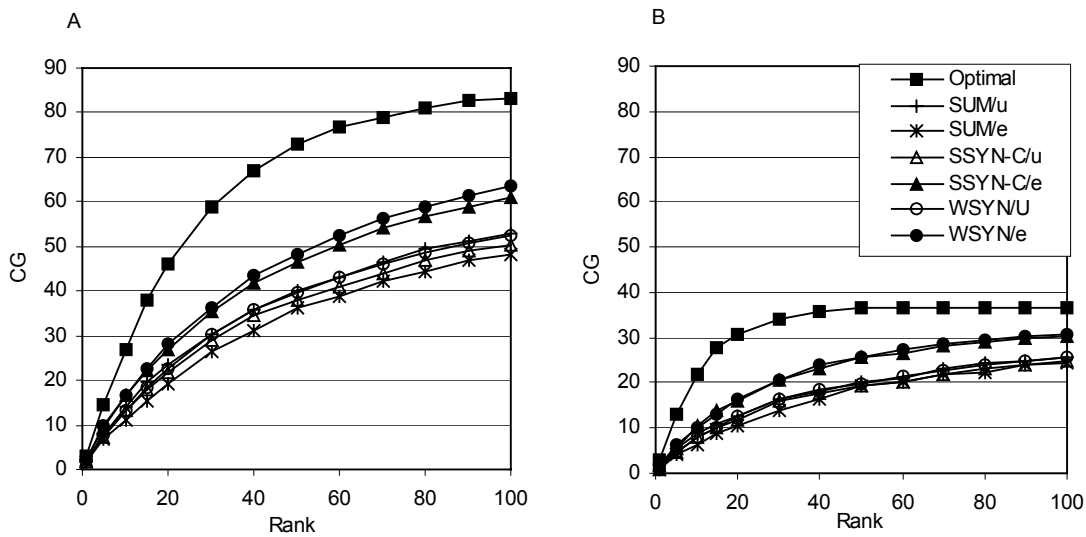
**Table 1.** Average non-interpolated precision figures for different query types.

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.

#### 4.6 Cumulated gain

Figure 5 presents the CG vector curves for ranks 1 - 100, the six query types studied above and the theoretically best possible (average) query. Figure 5A shows the curves when documents at both relevance levels 2 and 3 are taken into account (i.e. they earn 2 and 3 points, respectively). The best possible curve almost becomes a horizontal line at the rank 100 reflecting the fact that at rank 100 practically all relevant documents have been found. The two best (synonym structured) query types hang below by 18 - 27 points (35 - 39 %) from the rank 20 to 100. The difference is the greatest in the middle range. The other four query types remain further below by 5 - 15 points (about 16 - 24 %) from rank 20 to 100. The difference to the best possible curve is 23 - 38 points (50 %). Beyond the rank 100 the differences between the best possible and all actual curves are all bound to diminish. Figure 5B shows the curves when documents only

at the relevance level 3 considered. The precise figures are different and the absolute differences smaller. However, the proportional differences are larger.



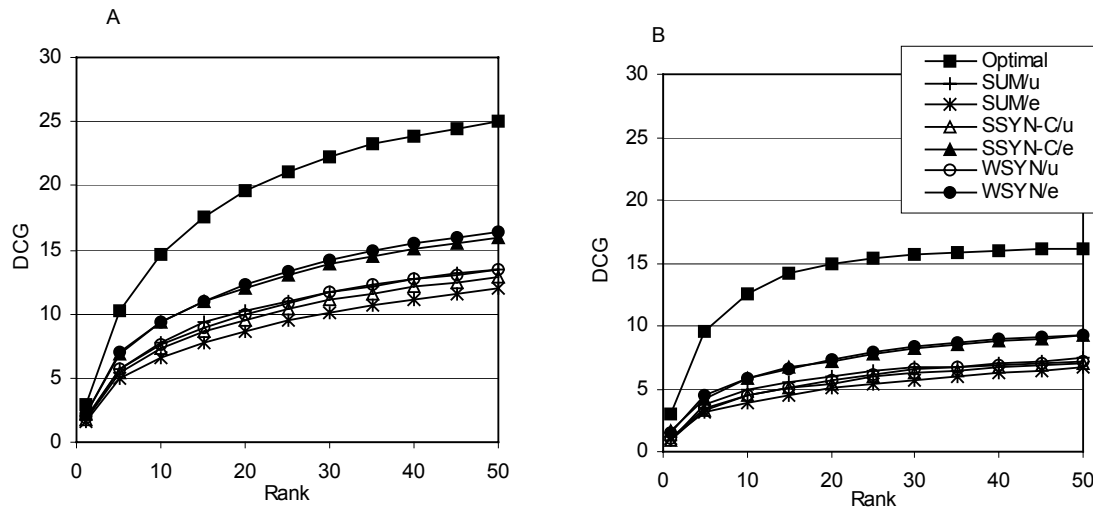
**Figure 5.** Cumulated gain curves at ranks 1-100, relevance levels 2&3 (A), and 3 (B).

The curves can be interpreted also in another way: at the relevance level 3 one has to retrieve 34 documents by the best query types, and 62 by the other query types, in order to gain the benefit that could theoretically be gained by retrieving only 10 documents. In this respect the best query types are nearly twice as effective as the others. At the relevance levels 2&3 the corresponding figures are 20 and 26 documents. At the greatest, the difference between the best and the remaining query types is 6 - 8 points (or two documents, relevance level 3) at ranks 40 - 60. At relevance levels 2&3 the greatest differences are 5 - 15 points (or 2 - 7 documents) at ranks 40 - 100.

#### 4.7 Discounted cumulated gain

Figure 6 shows the DCG vector curves for ranks 1 - 50, the six query types studied above and the theoretically best possible (average) query. The  $\log_2$  of the document rank is used as the discounting factor. Figure 6A shows the curves when documents both at the relevance levels 2 and 3 are taken into account. The best possible curve still grows at the rank 50 (it levels off at the rank 90). The two best (synonym structured) query types hang below by 5 - 9 points (35 - 36 %) from the rank 10 to 50. The difference is growing. The other four query types remain further below by 2 - 4 points (15 - 27 %) from rank 10 to 50. The difference to the best possible curve is 7 - 13 points (47 - 50 %). Beyond the rank 50 the differences between the best possible and all actual curves gradually become stable. Figure 6B shows the curves when documents only at the relevance level 3 considered. The precise figures are different and the absolute differences smaller. However, the proportional differences are larger. At the greatest, the difference between

the best and the remaining query types is 3 points (or one level - 3 document) at the rank 40 and further. It is a consistent and statistically significant difference but are the users able to notice it?



**Figure 6.** Discounted ( $\log_2$ ) cumulated gain curves ranks 1-50, relevance levels 2&3 (A), and 3 (B).

Also these curves can be interpreted in another way: at the relevance level 2&3 one has to expect the user to examine 35 documents by the best query types, and 70 by the other query types, in order to gain the (discounted) benefit that could theoretically be gained by retrieving only 10 documents. User persistence up to 35 documents is not unrealistic whereas up to 70 it must be rare. The difference in query type effectiveness is essential. At the relevance level 3 the discounted gains of the best query types never reach the gain theoretically possible at the rank 10. The theoretically possible gain at the rank 5 is achieved at the rank 50 and only by the best query types.

One might argue that if the user goes down to 70 documents, she gets the real value, not the discounted one and therefore the DCG data should not be used for effectiveness comparison. While this may hold for the user situation, the DCG-based comparison is valuable for the system designer. The user is less likely to scan that far and thus documents placed there do not have their real relevance value; a retrieval system or method placing relevant documents later in the ranked results should not be credited as much as another system or method ranking them earlier.

The main findings are similar with the other logarithm bases we tested. However, the magnitude of the differences between the best and worst query types grows from 4 points for  $\log_2$  to 13 points for  $\log_{10}$  at the rank 50 (obviously). This means that for a persistent user the best methods are 13 points (or 27 %) better than the remaining ones. For an impatient one, they are only 4 points better.

## 5 Discussion

We have argued that users should be allowed to start information retrieval from concepts rather than words. In the empirical case study we used a conceptual model for query formulation and expansion. This model is developed for the news domain and contains 832 concepts, 1,345 expressions for the concepts, and 1,558 search patterns for the expressions. It represents generally valid hierarchical, associative and equivalence relationships for the news domain. The findings suggest that, when available, such conceptual models may be automatically used for query expansion and construction for varying retrieval environments. Thus, conceptual models or conceptual modelling can be used to support conceptual level interaction between the user and the collection, freeing the user from many details of search key selection and query formulation.

Our second argument is that in modern large database environments, the development and evaluation of IR methods should be based on their ability to retrieve highly relevant documents. This is desirable from the user's viewpoint and presents a not too liberal test for IR methods. We developed two methods for IR method evaluation, which aim at taking the document relevance degrees into account. One is based on a novel application of the traditional P-R curves and separate recall bases for each relevance level of documents. The other is based on two novel evaluation measures, the CG and the DCG measures, which give the (discounted) cumulated gain up to any given document rank in the retrieval results.

In the case study we demonstrated the use of these evaluation methods in the evaluation of the effectiveness of various query types which were varied in structure and expansion. Our hypotheses were that:

- the performance differences between query types among marginally relevant documents should be marginal and among highly relevant documents essential when measured by the P-R curves,
- strongly structured expanded queries present better effectiveness than unexpanded queries or queries based on other structures, whether expanded or not, and
- expanded queries based on strong structures cumulate higher CG and DCG values than unexpanded queries or queries based on other structures, whether expanded or not.

These hypotheses were confirmed. The differences between the performance figures of the best and worst query types are consistent and statistically very significant. We valued the documents at different relevance levels rather equably, however, the user might value documents at relevance level 3 much higher than documents at other relevance levels. Thus, our analysis perhaps led to rather conservative, although significant results.

Sormunen and others analysed the contents of documents at different relevance levels (Sormunen & al. 2001). Their results suggest that highly relevant documents have the following characteristics: the topic of the request is discussed in them at length; they have more words pertaining to the topic; 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. Queries that contain several alternative search keys for each search concept retrieve highly relevant documents in best match systems because documents containing several search keys have 'more evidence of being on the topic' than documents with few search keys. The structure is needed to balance the queries: strong structured queries with the SYN operator were the best alternatives, because treating all alternative search keys as instances of one search key is a proper way to interpret the idea of Boolean disjunction in best match retrieval. Facet structuring with other operators<sup>9</sup> was not successful (Kekäläinen 1999).

The P-R curves demonstrate that the good performance of the expanded structured query types is due to, in particular, their ability to rank the highly relevant documents toward the top of retrieval results. The cumulated gain curves illustrate the value the user actually gets, but discounted cumulated gain curves can be used to forecast the system performance with regard to a user's patience in examining the result list. With a small log base, the value of a relevant document decreases quickly along the ranked list and a DCG curve turns horizontal. This assumes an impatient user for whom late coming information is not useful because it will never be read. If the CG and DCG curves are analysed horizontally, we may conclude that a system designer would have to expect the users to examine by 50 to 100 % more documents by the worse query types to collect the same gain collected by the best query types. While it is possible that persistent users go way down the result list, e.g. from 30 to 60 documents, it often is unlikely to happen, and a system requiring such a behaviour is, in practice, much worse than a system yielding the gain within a 50 % of the documents.

The novel CG and DCG measures complement the modified P-R measure. Precision over fixed recall levels hides the user's effort up to a given recall level. The DCV-based precision - recall curves are better but still do not make the value gained by ranked position explicit. The CG and DCG curves provide this directly. The distance to the theoretically best possible curve shows the effort wasted on less-than-perfect or useless documents. The advantage of the P-R measure is that it treats requests with different number of relevant documents equally, and from the system's point of view the precision at each recall level is com-

---

<sup>9</sup> For example, the OR operator in InQuery, which gives the product of argument probabilities, was not a suitable facet operator.

parable. In contrast, CG and DCG curves show the user's point of view as the number of documents needed to achieve a certain gain. Together with the theoretically best possible curve they also provide a stopping rule, that is, when the best possible curve turns horizontal, there is nothing to be gained by retrieving or examining further documents.

Generally, the evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable even in IR experiments, and may reveal interesting phenomena. The dichotomous relevance assessments generally applied may be too permissive, and, consequently, too easily give credit to IR system performance. We believe that, in modern large environments, the proposed modified P-R measure and the novel (D)CG measures should be used whenever possible, because they provide richer information for evaluation.

## 6 Conclusions

We tested the performance of different query types in regard to different relevance levels. Queries varied according to their structure and expansion. Queries had either a weak 'bag of words' structure or a strong facet structure. Query expansion was based on a conceptual model which gives several alternative search keys for concepts. Our hypothesis was that the performance differences between query types among marginally relevant documents should be minor than among highly relevant documents. This was shown true by clear differences between P-R curves at different relevance levels. We also introduced a new method for illustrating the performance of ranked, non-dichotomously assessed retrieval results from the user point-of view. This was based on scoring the documents according to their relevance level and drawing a cumulated curve of scores of the result set. Our hypothesis was that expanded queries based on strong structures cumulate higher value for the user than unexpanded queries or queries based on other structures, whether expanded or not. This should show clearly also when the ranked position of each document is used to discount its worth in the cumulated value. These hypotheses were confirmed. The results also indicate the usability of domain dependent conceptual models in query expansion for IR.

## References

- Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval* 4(3/4), 195-208.
- Allan, J., Callan, J., Croft, B., Ballesteros, L., Broglio, J., Xu, J. & Shu, H. (1997). INQUERY at TREC 5. In E. M. Voorhees & D. K. Harman (Eds.), *Information technology: The Fifth Text Retrieval Conference (TREC-5)*, pp. 119-132. National Institute of Standards and Technology.

- Bechhofer, S., Carr, L., Goble, C., Hall, W. (2001). Conceptual open hypermedia = the semantic Web? Position paper. To appear in *Proceedings of the second International Workshop on the Semantic Web – SemWeb '2001*.
- Belkin, N. J. & Croft, W. B. (1987). Retrieval techniques. In M. E. Williams (Ed.), *Annual review of information science and technology*, vol. 22. New York, NY: Elsevier, 109–145.
- Blair, D.C. & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM* 28(3), 289-299.
- Borlund, P. (2000). *Evaluation of interactive information retrieval systems*. Ph.D. dissertation. Åbo Akademi University Press.
- Borlund, P & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324-331, ACM.
- Chaffee, J. & Gauch, S. (2000). Personal ontologies for Web navigation. In *Proceedings of the ninth international conference on Information knowledge management – CIKM 2000*, pp. 227-234.
- Chen, H. & Dumais, S. (2000). Bringing order to the Web: Automatically categorizing search results. *CHI Letters* 2 (1), 145-152.
- Conover, W.J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley & Sons.
- Cooper, W.S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science* 19(1), 30 – 41.
- Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management* 36, 533–550.
- Efthimiadis, E.N. (1996). Query expansion. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology*, Vol. 31, pp. 121–187. Information Today.
- Green, R. (1995). The expression of conceptual syntagmatic relationships: A comparative survey. *Journal of Documentation* 51(4), 315-338,.
- Guarino, N., Masolo, C. & Vetere, G. (1998). *OntoSeek: Using large linguistic ontologies for gathering information resources from the Web*. LADSEB-CNR Technical Report 01/98.
- Harman, D. K. (1995). *Overview of the fourth text retrieval conference (TREC-4)* [online]. [Cited 2.1.2002] Available at: <URL: <http://trec.nist.gov/pubs/trec4/papers/overview.ps>>.
- Hersh, W. R. (1996). *Information retrieval: A health care perspective*. Springer Verlag.
- Hersh, W.R. & Hickam, D.H. (1995). An evaluation of interactive Boolean and natural language searching with an online medical textbook. *Journal of the American Society for Information Science* 46(7), 478-489.

- Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Korfhage, R., Rasmussen, E.M. & Willett, P. (Eds.), *Proceedings of the 16th International Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 349–338.
- Ingwersen, P. & Willett, P. (1995). An introduction to algorithmic and cognitive approaches for information retrieval. *Libri 45*, 160-177.
- Jacob, E.K. (1991). Classification and categorization: Drawing the line. In B. H. Kwasnik & R. Fidel (Eds.), *Advances in classification research*. Vol. 2. Proceedings of the 2nd ASIS SIG/CR classification research workshop. Medford, NJ: Learned Information, 67–83.
- Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for highly relevant documents. In: Belkin, N.J., Ingwersen, P. and Leong, M.-K. (Eds.) *Proceedings of the 23<sup>rd</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, pp. 41–48.
- Järvelin K, Kekäläinen J and Niemi T (2001) ExpansionTool: concept-based query expansion and construction. *Information Retrieval* 4(3/4), 231-255.
- Järvelin, K., Kristensen, J., Niemi, T., Sormunen, E. & Keskustalo, H. (1996). A deductive data model for query expansion. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, pp. 235–249.
- Keen, E.M. (1991). The use of term position devices in ranked output experiments. *Journal of Documentation* 47(1), 1-22.
- Keen, E. M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing & Management* , 28(4), 491–501.
- Kekäläinen, J. (1999). *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Ph.D. dissertation. Department of Information Studies, University of Tampere. (Available at: [http://www.info.uta.fi/research/postscript\\_docs/JK1\\_99.pdf](http://www.info.uta.fi/research/postscript_docs/JK1_99.pdf).)
- Kekäläinen, J. & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 130–137.
- Kekäläinen, J. & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval* 1(4), 329-344.
- Korfhage, R.R. (1997). *Information storage and retrieval*. Wiley & Sons, New York.
- Lancaster, F. W. (1986). *Vocabulary control for information retrieval* (2nd ed.). Information Resources Press.

- Losee, R.M. (1998). *Text retrieval and filtering: Analytic models of performance*. Kluwer Academic Publishers.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM* 38(11), 39-41.
- Myaeng, S.H. & Korfhage, R.R. (1990). Integration of user profiles: Models and experiments in information retrieval. *Information Processing & Management* 26(6), 719-738.
- Over, P. (1999). *TREC-7 interactive track report* [On-line]. Available at <http://trec.nist.gov/pubs/trec7/papers/t7irep.pdf.gz>.
- Rajashekar, T.B. & Croft, W.B. (1995). Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science* 46(4), 272-283.
- Robertson, S.E. & Belkin, N.J. (1978). Ranking in principle. *Journal of Documentation* 34(2), 93-100.
- Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Salton, G. & McGill, M.J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, London.
- Saracevic, T. (1996). Relevance reconsidered '96. In P. Ingwersen & N.O. Pors (Eds.), *Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*. The Royal School of Librarianship, pp. 201-218.
- Saracevic, T, Kantor, P., Chamis, A. & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science* 39(3), 161-176.
- Schamber, L. (1994). Relevance and Information Behavior. In: M.E. Williams (Ed.), *Annual Review of Information Science and Technology* 29. Medford, NJ: Information Today (pp. 3-48).
- Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Association for Information Science*, 50(12), 1119-1120.
- Sormunen, E. (2000). *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases*. Ph.D. dissertation. Department of Information Studies, University of Tampere. (Available at: <http://acta.uta.fi/teos.phtml?3786>.)
- Sormunen, E., Kekäläinen, J., Koivisto, J., Järvelin, K. (2001). Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. *Journal of Documentation* 57(3), 538-376.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 11-21.
- Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 28(4), 467-490.

- Turtle, H.R. (1990). *Inference networks for document retrieval*. Ph.D. dissertation. Computer and information Science Department, University of Massachusetts.
- Uschold, M. & Gruninger, M. (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review* 11(2), 93-136.
- Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation* 56, 540 – 562.
- Xu, J. & Croft, W. B. (1996). Query expansion using local and global document analysis. In H.-P. Frei, D. Harman, P. Schäuble & R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, pp. 4-11.