University of Tampere

Department of Information Studies

Research Notes

RN • 2002 • 2

KALERVO JÄRVELIN & JAANA KEKÄLÄINEN

CUMULATED GAIN-BASED INDICATORS OF IR PERFORMANCE

Tampereen yliopisto • Informaatiotutkimuksen laitos • Tiedotteita 2002 • 2

CUMULATED GAIN-BASED INDICATORS OF IR PERFORMANCE

Kalervo Järvelin & Jaana Kekäläinen
University of Tampere
Department of Information Studies
FIN-33014 University of Tampere
FINLAND
Email: {kalervo.jarvelin, jaana.kekalainen}@uta.fi

Abstract

Modern large retrieval environments tend to overwhelm their users by their large output. Since all documents are not of equal relevance to their users, highly relevant documents should be identified and ranked first for presentation to the users. In order to develop IR techniques to this direction, it is necessary to develop evaluation approaches and methods that credit IR methods for their ability to retrieve highly relevant documents. This can be done by extending traditional evaluation methods, i.e., recall and precision based on binary relevance assessments, to graded relevance assessments. Alternatively, novel measures based on graded relevance assessments may be developed. This paper proposes three novel measures that compute the cumulative gain the user obtains by examining the retrieval result up to a given ranked position. The first one accumulates the relevance scores of retrieved documents along the ranked result list. The second one is similar but applies a discount factor on the relevance scores in order to devaluate late-retrieved documents. The third one computes the relative-to-the-ideal performance of IR techniques, based on the cumulative gain they are able to yield. The novel measures are defined and discussed and then their use is demonstrated 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. The test was run with a best match retrieval system (InQuery[1]) in a text database consisting of newspaper articles. The results indicate that the proposed measures credit IR methods for their ability to retrieve highly relevant documents and allow testing of statistical significance of effectiveness differences. The graphs based on the measures also provide insight into the performance IR techniques and allow interpretation, e.g., from the user point of view.

1. Introduction

Modern large retrieval environments tend to overwhelm their users by their large output. Since all documents are not of equal relevance to their users, highly relevant documents, or document components, should be identified and ranked first for presentation to the users. This

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

is desirable from the user point of view. In order to develop IR techniques to this direction, it is necessary to develop evaluation approaches and methods that credit IR methods for their ability to retrieve highly relevant documents.

The current practice of liberal binary assessment of topical relevance gives equal credit for a retrieval method for retrieving highly and fairly relevant documents. For example, TREC is based on binary relevance assessments with a very low threshold for accepting a document as relevant – the document needs to have at least one sentence pertaining to the request to count as relevant [TREC 2001]. Therefore differences between sloppy and excellent retrieval techniques, regarding highly relevant documents, may not become apparent in evaluation. To bring such differences into daylight, both graded relevance judgements and a method for using them are required.

In most laboratory tests in IR documents are judged relevant or irrelevant with regard to the request. 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; Järvelin & Kekäläinen 2000]. 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. 1988]. However, graded relevance assessments may be collected in field studies [Vakkari & Hakala 2000; Spink & al., 1998] and also produced for laboratory test collections [Sormunen 2001; Voorhees 2001], so they are available.

Graded relevance judgements may be used for IR evaluation, firstly, by extending traditional evaluation measures, such as recall and precision and P-R curves, to use them. Järvelin and Kekäläinen [2000; 2001] propose the use of each relevance level separately in recall and precision calculation. Thus different P-R curves are drawn for each level. They demonstrate that differing performance of IR techniques at different levels of relevance may thus be observed and analysed. In the latter study Järvelin and Kekäläinen generalise recall and precision calculation to directly utilise graded document relevance scores. They consider precision as a function of recall, but the approach extends to DCV (Document Cut-off Value) based recall and precision as well. They demonstrate that the relative effectiveness of IR techniques, and the statistical significance of their performance differences, may vary according to the relevance scales used.

In the present paper we develop three new evaluation measures, which seek to estimate the cumulative relevance gain the user receives by examining the retrieval result up to a given rank. The first one accumulates the relevance scores of retrieved documents along the ranked result list. The second one is similar but applies a discount factor on the relevance scores in order to devaluate late-retrieved documents. The third one computes the relative-to-the-ideal performance of IR techniques, based on the cumulated gain they are able to yield. The first two were originally presented in [Järvelin & Kekäläinen 2000] and were also applied in the TREC Web Track 2001 [Voorhees 2001] and in a text summarisation experiment by Sakai and Sparck Jones [2001]. These novel measures are akin to the average search length [briefly ASL; Losee 1998], sliding ratio [Korfhage 1997], and normalised recall [Pollack 1968; Salton & McGill 1983; Korfhage 1997] measures. They also have some resemblance to the ranked half life and relative relevance measures proposed by Borlund and Ingwersen [1998] for interactive IR. However, they offer several advantages by taking both the degree of relevance and the rank position (determined by the probability of relevance) of a document into account. [For a discussion of the degree of relevance and the probability of relevance, see Robertson & Belkin 1978.]

The novel measures are first defined and discussed and then their use is demonstrated 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. Kekäläinen [1999], and Kekäläinen and Järvelin [2000] have earlier observed that the structure of queries influences retrieval performance when queries are expanded. Query structure refers to the syntactic structure of a query expression, marked with query operators and parentheses. They reported significant differences in retrieval effectiveness of their query structures when expansion and relevance levels were varied [Järvelin & Kekäläinen 2000; Kekäläinen & Järvelin 2001]. This paper uses a similar case to demonstrate and analyse performance differences between IR techniques in the light of the proposed cumulated gain based measures. The tests were run with a best match retrieval system (InQuery) in a text database consisting of newspaper articles. The results indicate that the proposed measures credit IR methods for their ability to retrieve highly relevant documents and allow testing of statistical significance of effectiveness differences. The graphs based on the measures also provide insight into the performance IR techniques and allow interpretation, e.g., from the user point of view.

Section 2 explains our evaluation measures: the cumulated gain-based evaluation measures. Section 3 presents the case study. The test environment, relevance assessments, query structures and expansion, and the retrieval results are reported. Section 4 contains discussion and conclusions.

2 . Cumulated gain -based measurements

2.1 Direct cumulated gain

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

- highly relevant documents are more valuable than marginally relevant documents, and
- the greater the ranked position of a relevant document, the less valuable it is for the user, because the less likely it is that the user will ever examine the document.

The first point leads to comparison of IR techniques through test queries by their cumulated gain by document rank. In this evaluation, the relevance score 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 ranked 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 scores. Assume that the relevance scores 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:

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

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' = <3, 5, 8, 8, 8, 9, 11, 13, 16, 16, …>. The cumulated gain at any rank may be read directly, e.g., at rank 7 it is 11.

2.2 Discounted cumulated gain

The second point above stated that the greater the ranked position of a relevant document, the less valuable it is for the user, because the less likely it is that the user will ever examine the document due to time, effort, and cumulated information from documents already seen. This leads to comparison of IR techniques 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 score is added to the cumulated gain.

A discounting function is needed which progressively reduces the document score 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 discounting with this requirement is to divide the document score 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 it 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' given in the preceding section we obtain DCG' = <3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, …>.

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 (Section 3).

To compute the averaged vectors, we need vector sum operation and vector multiplication by a constant. Let $V = \langle v_1, v_2, \ldots, v_k \rangle$ and $W = \langle w_1, w_2, \ldots, w_k \rangle$ be two vectors. Their sum is the vector $V + W = \langle v_1 + w_1, v_2 + w_2, \ldots, v_k + w_k \rangle$. For a set of vectors $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$, each of *k* components, the sum vector is generalised as $\Sigma_{V \in \mathbf{V}} V = V_1 + V_2 + \ldots + V_n$. The multiplication of a vector $V = \langle v_1, v_2, \ldots, v_k \rangle$ by a constant *r* is the vector $r*V = \langle r*v_1, r*v_2, \ldots, r*v_k \rangle$. The average vector **AV** based on vectors $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$, is given by the function *avg-vect*(**V**):

$$avg\text{-}vect(\mathbf{V}) = |\mathbf{V}|^{-1} * \Sigma_{V \in \mathbf{V}} V \quad (3)$$

Now the average CG and DCG vectors for vector sets **CG** and **DCG**, over a set of test queries, are computed by *avg-vect*(**CG**) and *avg-vect*(**DCG**).

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 … *m* by the values 3, then the positions *m*+1 … *m*+*l* by the values 2, then the positions *m*+*l*+1 … *m*+*l* +*k* by the values 1, and finally the remaining positions by the values 0. More formally, the theoretically best possible score vector BV for a request of *k, l,* and *m* relevant documents at the relevance levels 1, 2 and 3 is constructed as follows:

$$BV[i] = \begin{cases} 3, \text{ if } i \leq m, \\ 2, \text{ if } m < i \leq m + l, \\ 1, \text{ if } m + l < i \leq m + l + k, \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

A sample ideal gain vector could be:

$$I' = \langle 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, \ldots \rangle$$

The CG and DCG vectors, as well as the average CG and DCG vectors and curves, are computed as above. Note that the curves turn horizontal when no more relevant documents (of any level) can be found (Section 3 gives examples). 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 (average) curve shows the effort wasted on less-than-perfect documents due to a particular IR method. Based on the sample ideal gain vector I', we obtain the ideal CG and DCG (b = 2) vectors:

CG_I' = <3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, …>

DCG_I' = <3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 11.21, 11.53, 11.83, 11.83, 11.83, …>.

Note that the ideal vector is based on the recall base of the search topic rather than on the result of some IR technique. This is an important difference with respect to some related measures, e.g. the sliding ratio and satisfaction measure [Korfhage 1997].

2.3. Relative to the ideal measure – the normalised (D)CG-measure

Are two IR techniques significantly different in effectiveness from each other when evaluated through (D)CG curves? In the case of P-R performance, we may use the average of interpolated precision figures at standard points of operation, e.g., eleven recall levels or DCV points, and then perform a statistical significance test. The practical significance may be judged by the Sparck Jones [1974] criteria, e.g., differences less than 5% are marginal and differences over 10% are essential. P-R performance is also relative to the ideal performance: 100% precision over all recall levels. The (D)CG curves are not relative to an ideal. Therefore it is difficult to assess the magnitude of the difference of two (D)CG curves and there is no obvious significance test for the difference of two (or more) IR techniques either. One needs to be constructed.

The (D)CG vectors for each IR technique can be normalised by dividing them by the corresponding ideal (D)CG vectors, component by component. In this way, for any vector position, the normalised value 1 represents ideal performance, and values in the range [0, 1) the share of ideal performance cumulated by each technique. Given an (average) (D)CG vector V =

$<v_1, v_2, \ldots, v_k>$ of an IR technique, and the (average) (D)CG vector $I = <i_1, i_2, \ldots, i_k>$ of ideal performance, the normalised performance vector n(D)CG is obtained by the function:

$$norm\text{-}vect(V, I) = <v_1/i_1, v_2/i_2, \ldots, v_k/i_k> \qquad (5)$$

For example, based on CG' and CG_I' from above, we obtain the normalised CG vector nCG' = *norm-vect*(CG', CG_I') =

<1, 0.83, 0.89, 0.73, 0.62, 0.6, 0.69, 0.76, 0.89, 0.84, …> .

The normalised DCG vector nDCG' is obtained in a similar way from DCG' and DCG_I'. Note that, as a special case, the normalised ideal (D)CG vector is always *norm-vect*(I, I) = <1, 1, …, 1>, when I is the ideal vector.

The area between the normalised ideal (D)CG vector and the normalised (D)CG vector represents the quality of the IR technique. Normalised (D)CG vectors for two or more IR techniques also have a normalised difference. These can be compared in the same way as P-R curves for IR techniques. The average of a (D)CG vector (or its normalised variation), up to a given ranked position, summarises the vector (or performance) and is analogous to the non-interpolated average precision of a DCV curve up to the same given ranked position. The average of a (n)(D)CG vector V up to the position *k* is given by:

$$avg\text{-}pos(V, k) = k^{-1} * \Sigma_{i=1\ldots k} V[i] \qquad (6)$$

These vector averages can be used in statistical significance tests in the same way as average precision over standard points of operation, e.g., eleven recall levels or DCV points.

2.4. Comparison to earlier 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], the sliding ratio measure [SR for short; Pollack 1968; Korfhage 1997], 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 *SR measure* takes the degree of document relevance into account and actually computes the cumulated gain and normalises this by the ideal cumulated gain for *the same retrieval result.* The result thus is quite similar to our nCG vectors. However, SR is heavily dependent on the retrieved list size: with a longer list the ideal cumulated gain may change essentially and this affects all normalised SR ratios from rank one onwards. Because our nCG is based on the recall base of the search topic, the first ranks of the ideal vector are not affected at all by extension of the evaluation to further ranks. Improving on normalised recall, SR is not dependent on outliers, but it is too sensitive to the actual retrieved set size. SR does not have the discount feature of our (n)DCG measure.

The *SFT measure* consists of three components similar to the SR measure. 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. Like SR, also 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 $n << N$ (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 nDCG 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, DCG, nCG and nDCG 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.

Further, the normalised nCG and nDCG measures support evaluation:

- They represent performance as relative to the ideal based on a known (possibly large) recall base of graded relevance assessments.
- The performance differences between IR techniques are also normalised in relation to the ideal thereby supporting the analysis of performance differences.

Järvelin and Kekäläinen have earlier proposed recall and precision based evaluation measures to work with graded relevance assessments [Järvelin & Kekäläinen 2000; Kekäläinen & Järvelin 2001]. They first propose the use of each relevance level separately in recall and precision calculation. Thus different P-R curves are drawn for each level. Performance differences at different relevance levels between IR techniques may thus be analysed. Further, they generalise recall and precision calculation to directly utilise graded document relevance scores. They consider precision as a function of recall and demonstrate that the relative effectiveness of IR techniques, and the statistical significance of their performance differences, may vary according to the relevance scales used. The proposed measures are similar to standard IR measures while taking document relevance scores into account. They do not have the discount feature of our (n)DCG measure. The measures proposed in this paper are directly user-oriented in calculating the gain cumulated by consulting an explicit number of documents. P-R curves tend to hide this information. The generalised P-R approach extends to DCV (Document Cut-off Value) based recall and precision as well, however.

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 n(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.

3. 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 query structures in a database with non-binary relevance judgements. Kekäläinen and Järvelin [1998; 2000] have shown that queries with same search keys but different structures have significant differences in performance. In the present study we shall test three differently structured queries with relation to the degree of relevance. We give the results as CG and DCG curves, which exploit the degrees of relevance. Further, we show the results as normalised nCG and nDCG curves, and present the results of a statistical test based on the averages of n(D)CG vectors.

3.1 Test environment

The test environment consists of a text database, the InQuery retrieval system (version 3.1) and a request collection with relevance judgements. The database contains 53,893 newspaper articles. The requests 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. The base was collected by pooling the result sets of hundreds of different queries, using both exact and partial match retrieval. We thus believe that our recall estimates are valid. For a more detailed description of the test environment, see Kekäläinen [1999], Kekäläinen and Järvelin [2000], and Sormunen [2000]. The tests of this study were run with 30 requests.

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. Differences in relevance judgements were handled as follows: 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 base 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).

3.2 Query structures and expansion

Query structure refers to the syntactic structure of a query expression, marked with query operators and parentheses [Kekäläinen & Järvelin 1998]. In exact match – or Boolean – retrieval a query has a structure based on conjunctions and disjunctions of search keys. [Green 1995; Keen 1991]. In best match retrieval, matching is ranking documents according to scores calculated from the weights of search keys occurring in documents. 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. Kekäläinen and Järvelin [2000] divide the structures of best match queries into strong and weak. In the former, concepts are marked through operators; in the latter, concepts cannot be recognised through syntax.

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, strongly structured queries. However, all strongly structured queries were not effective: the algorithmic interpretation of the operators is important. In the present study we adopted three of their query structures. A weak SUM structure presents a typical 'bag of words query'. A strong best match Boolean structure (BOOL) shows the problems of the traditional soft interpretation of the OR operator. A strong concept-based structure sum-of-synonym-groups (SSYN-C) is a modification of the Boolean structure – the AND and the OR operator are replaced with the SUM (an average of the weights of keys) and the SYN (all keys are treated as instances of one key) operators respectively.

Kekäläinen and Järvelin [1998] used a thesaurus for query expansion and had several expansion types. In the present study, all queries were run with the largest expansion including ex-

pressions for synonyms, narrower concepts, and related concepts. Thus, we compare three different query structures with one expansion type, i.e. all queries have the same search keys. For samples of the query structures and expansion types, see Kekäläinen and Järvelin [2000], or Kekäläinen [1999].

3.3 The application of the evaluation measures

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:

We tested different relevance weights at different relevance levels. First, we used document relevance levels 0, 1, 2, 3 directly as gained value measures. This can be criticised by asking whether a highly relevant document is (only) three times as valuable as a marginally relevant document. Thus, we also replaced the relevance levels with weights 0, 1, 4, 10, and 0, 1, 10, 100 in turn to give more emphasis to highly relevant documents.

The logarithm bases 2 and 10 were tested for the DCG vectors. The base 2 models impatient users, base 10 persistent ones.

The average actual CG and DCG vectors were compared to the ideal average vectors.

The average actual CG and DCG vectors were normalised by dividing them with the ideal average vectors.

In the following chapters the results of the weights 0, 1, 10, 100 are only shown because the sample case is only meant to illustrate the methods. Different weighting on highly relevant documents may however affect the relative effectiveness of IR techniques as also pointed out by Voorhees [2001]. Varying weighting affects scale in (D)CG graphs, and the normalised curves reach higher n(D)CG levels but effects in differences between curves and their shapes are minor with the present sample data. Also, we only give the results for the logarithm base 2. We prefer the stricter test condition the smaller logarithm base provides.

3.4 Cumulated gain

Figure 1 presents the CG vector curves for the three structure types at ranks 1 - 200, and the ideal curves. In the ranked result list, highly relevant documents add 100 points to the cumulated gain; fairly relevant documents add 10 points; marginally relevant documents add 1 point; and irrelevant documents add 0 points to the gain.

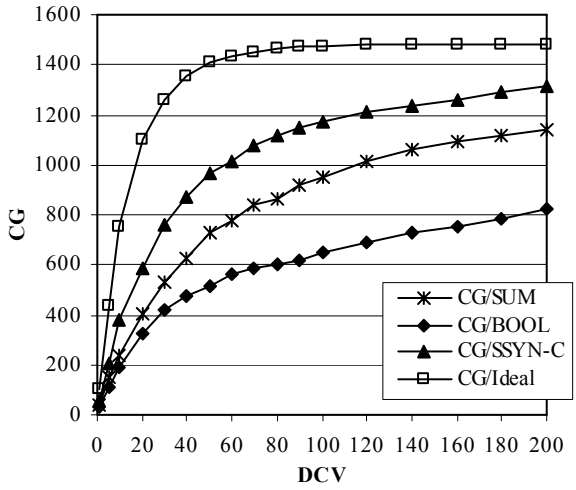

Figure 1. Cumulated gain (CG) curves for three query structure types and the ideal curve at ranks 1-200.

The best possible curve becomes a horizontal line at the rank 100 reflecting the fact that at the rank 100 practically all relevant documents have been found. The best (SSYN) query type hangs below the ideal by 43 – 531 points[2] (11 – 54 %). The absolute difference is greatest in the range from rank 20 to 100; the relative difference is greatest in the early ranks. The other two query types remain further below by 20 - 535 points (about 13 - 57 %). The difference to the best possible curve is 63 - 900 points (23 -79 %). Beyond the rank 200 the differences between the best possible and all actual curves are all bound to diminish.

The curves can be interpreted also in another way: one has to retrieve 30 documents by the best query type, and 60 by the second best, and up to 170 documents by the worst query type query in order to gain the benefit that could theoretically be gained by retrieving only 10 documents. In this respect the best query type is twice as effective as any of the others.

3.5 Discounted cumulated gain

Figure 2 shows the DCG vector curves for the three structure types at ranks 1 - 200, and the ideal curve. The $\log_2$ of the document rank is used as the discounting factor. The ideal curve levels off at the rank 145. The best query type hangs below by 43 - 258 points (33 - 53 %). The other two query types remain further below by 20 - 179 points (20 - 57 %). The differ-

[2] NB. The difference of 530 points means 5,3 most relevant documents.

ence to the best possible curve is 63 - 408 points (47 - 78 %). All the actual curves still grow at the rank 200, but beyond that the differences between the best possible and the other curves gradually become stable.

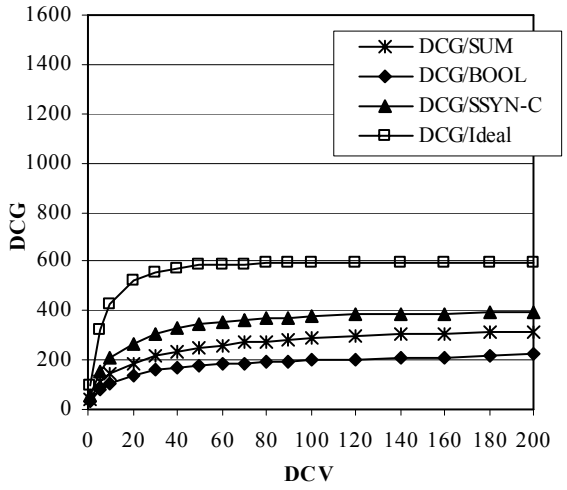

Figure 2. Discounted cumulated gain (DCG) curves for the three query structure types and the ideal curve at ranks 1-200

Also these graphs can be interpreted in another way: one has to expect the user to examine 40 documents by the best query type in order to gain the (discounted) benefit that could theoretically be gained by retrieving only 5 documents. The other two curves never reach that gain. The difference in query type effectiveness is essential. The discounted gains of any query type never reach the gain theoretically possible at the rank 10.

One might argue that if the user goes down to 40 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 and less likely to scan further and thus documents placed there do not have their real relevance value, a retrieval technique placing relevant documents later in the ranked results should not be credited as much as another technique ranking them earlier.

3.6 Normalised (D)CG vectors and statistical testing

Figure 3 shows the curves for CG vectors normalised by the ideal vectors. The curve for the normalised ideal CG vector has value 1 at all ranks. The actual normalised CG vectors reach it in due course when all relevant documents have been found. Differences at early ranks are easier to observe than in Figure 1. For example, the performance of the SUM queries is closer to the BOOL queries at ranks 1 – 20 but reaches somewhat the SSYN-C queries after that. The nCG curves readily show the differences between methods to be compared but they lack the straightforward interpretation of the gain at each rank given by CG curves.

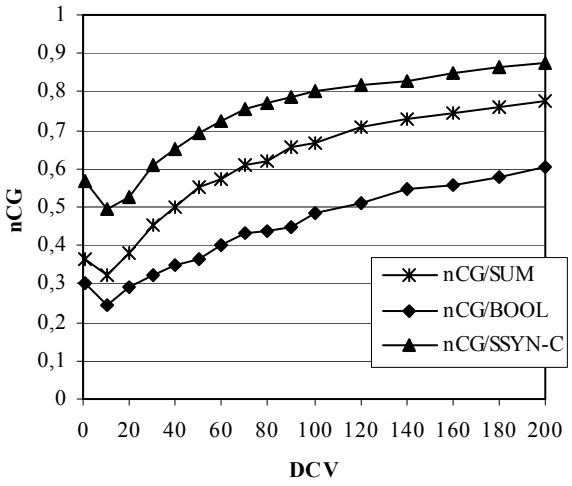

Figure 3. Normalised cumulated gain (nCG) curves for the three query structure types.

Figure 4 displays the normalised curves for DCG vectors. The curve for the normalised ideal DCG vector has value 1 at all ranks. The actual normalised DCG vectors never reach it, they start to level off upon the rank 200. The effect of discounting can be seen by comparing Figures 3 and 4: The difference between the SSYN-C queries and the two other query types does not diminish in Figure 4 as in Figure 3. The effect of normalisation can be detected by comparing Figure 2 and Figure 4: the differences between the IR techniques are clear and comparable.

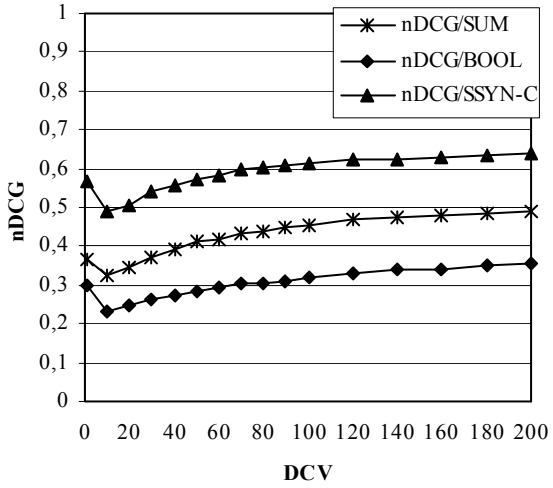

Figure 4. Normalised discounted cumulated gain (nDCG) curves for the three query structure types.

Statistical testing of differences between query types was based on normalised average n(D)CG vectors. These vector averages can be used in statistical significance tests in the same way as average precision over standard points of operation, or average non-interpolated precision. The classification we used to label the relevance levels through numbers 0 – 3 is on an ordinal scale. Holding to the ordinal scale suggests non-parametric statistical tests, such as the Friedman test [see Conover 1980]. However, we have based our calculations on class weights to represent their relative differences. The weights 0, 1, 10 and 100 denote differences on a ratio scale. This suggests the use of parametric tests such as ANOVA provided that its assumptions on sampling and measurement distributions are met. Next we give the results of both tests.

	Query structure type			Differences and statistical significance		
	BOOL	SUM	SSYN-C	Difference SUM - BOOL	Difference SSYN-C - BOOL	Difference SSYN-C - SUM
nCG	0.457	0.627	0.755	0.170 (##)	0.298** (##)	0.128**
nDCG	0.308	0.437	0.596	0.129 (##)	0.288** (##)	0.159** (##)

Table 1. n(D)CG averages over requests and statistical significance the results for three query types (legend: ** = p< 0.01 Friedman test; ## = p<0.01 ANOVA). Weighting scheme for documents at different relevance levels is 0-1-10-100.

Table 1 shows the average n(D)CG figures for different query types, their differences and the significance of the differences. In the table, the query type average is first calculated for each request, then an average is taken over the requests. It is worth noting that (n)(D)CG curves and Friedman test in Table 1 give somewhat conflicting information. In Figure 3 it is obvious that the curve of SUM queries is closer to the curve of SSYN-C queries than that of BOOL

queries. Nevertheless, the difference between SUM and BOOL is not significant although the difference between SSYN-C and SUM is. This example demonstrates how the average curves and statistical testing based on requests may lead to inconsistent conclusions. The curves averaged over requests at the ranks from 1 to 200 are affected by the magnitude of differences between requests, whereas the Friedman test is based on the order of differences. In other words, the average figures at given ranks may favour a query type A over type B, but the number of better performing requests may be smaller for the former.

The results of the parametric ANOVA test show significant differences between BOOL and the other query types in nCG, and between all query types in nDCG. The result corroborates that the parametric tests taking into account the magnitude of differences emphasise individual requests which have great variability in standard errors, whereas rank-based tests treat requests equally [see, Hull 1993].

4 . Discussion and conclusions

We have argued 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 viewpoint and presents a not too liberal test for IR techniques.

We then developed novel methods for IR technique evaluation, which aim at taking the document relevance degrees into account. These are the CG and the DCG measures, which give the (discounted) cumulated gain up to any given document rank in the retrieval results, and their normalised variants nCG and nDCG, based on the ideal retrieval performance. They are related to some traditional measures like *average search length* [ASL; Losee 1998], *expected search length* [ESL; Cooper 1968], normalised recall [NR; Rocchio 1966; Salton & McGill 1983], sliding ratio [SR; Pollack 1968; Korfhage 1997], and satisfaction – frustration – total measure [SFT; Myaeng & Korfhage 1990], and RHL [Borlund & Ingwersen 1998].

The benefits of the proposed novel measures are many. They systematically combine document rank and degree of relevance. 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. Performance is determined on the basis of recall bases for search topics and thus does not vary in an uncontrollable way which is true of measured based on the re-

trieved lists only. The novel measures are not heavily dependent on outliers since they focus on the gain cumulated from the beginning of the result up to any point of interest. They are obvious to interpret, and do not mask bad performance. They are directly user-oriented in calculating the gain cumulated by consulting an explicit number of documents. P-R curves tend to hide this information. In addition, the DCG measure realistically down weights the gain received through documents found later in the ranked results and allows modelling user persistence in examining long ranked result lists by adjusting the discounting factor. Further, the normalised nCG and nDCG measures support evaluation by representing performance as relative to the ideal based on a known (possibly large) recall base of graded relevance assessments. The performance differences between IR techniques are also normalised in relation to the ideal thereby supporting the analysis of performance differences.

An essential feature of the proposed measures is the weighting of documents at different levels of relevance. What is the value of a highly relevant document compared to the value of fairly and marginally relevant documents? There can be no absolute value because this is a subjective matter which also depends on the information seeking situation. Even classifying documents into different relevance levels does not correspond with the understanding of individual users but this holds for binary relevance classification as well. The original classification we used to label the relevance levels through numbers 0 – 3 is on an ordinal scale, thus the magnitude of the differences has no unequivocal meaning. However, we have based our calculations on the assumption that these classes can be given weights denoting differences on a ratio scale. It is difficult to justify any particular weighting scheme. Nevertheless, it is possible to illustrate different evaluation situations by giving varying weights to the documents of different relevance levels, i.e., 'what if the most relevant documents were three times, ten times or one hundred times more valuable than the others?'.

As to statistical testing, holding to the ordinal scale interpretation suggests non-parametric statistical tests, such as the Wilcoxon test or the Friedman test applied above. However, when weights are used, the scale of measurement becomes one of interval or ratio scale. This suggests the use of parametric tests such as ANOVA or t-test provided that their assumptions on sampling and measurement distributions are met. For example, Zobel [1998] used parametric tests when analysing the reliability of IR experiment results. Also Hull [1993] argues that with sufficient data parametric tests may be used. Our test case ANOVA gave a result different from Friedman – an effect of the magnitude of the differences between the IR techniques.

TREC has been based on binary relevance assessments with a very low threshold for accepting a document as relevant – the document needs to have at least one sentence pertaining to the request to count as relevant [TREC 2001]. At that level we would count the document at most as marginal. If the share of marginal documents were high in the test collection, then by utilising TREC-like liberal binary relevance assessments would lead to difficulties in identifying the better techniques as such. The differences between the techniques in retrieving highly relevant documents would be evened up by their indifference in retrieving marginal documents. The net differences might seem practically marginal and statistically insignificant.

The DCG measure has been applied in the TREC Web Track 2001 [Voorhees 2001] and in a text summarisation experiment by Sakai and Sparck Jones [2001]. Voorhees' findings are based on a three-point relevance scale. She examined the effect of incorporating highly relevant documents (HRDs) into IR system evaluation and weighting them in more or less sharply in a DCG-based evaluation. She found out that the relative effectiveness of IR systems is affected when evaluated HRDs. Voorhees pointed out that moderately sharp weighting of HRDs in DCG measurement supports evaluation for HRDs but avoids problems caused by instability due to small recall bases of HRDs in test collections. Sakai and Sparck Jones first assigned the weight 2 to each highly relevant document, and the weight 1 to each partially relevant document. They also experimented with other valuations, e.g., zero for the partially relevant documents. Sakai and Sparck Jones used log base 2 as the discounting factor to model user's (lack of) persistence. The DCG measure served to test the hypotheses in the summarisation study. These applications exemplify the usability of the cumulated gain –based approach to IR evaluation.

The cumulated gain curves illustrate the value the user actually gets, but discounted cumulative 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 100 to 600 % 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 CG and DCG measures complement P-R based measures [Järvelin & Kekäläinen 2000; Kekäläinen & Järvelin 2001]. Precision over fixed recall levels hides the user's effort up to a given recall level. The DCV-based precision - recall graphs are better but still do not make the value gained by ranked position explicit. The CG and DCG graphs provide this directly. The distance to the theoretically best possible curve shows the effort wasted on less-than-perfect or useless documents. The normalised CG and DCG graphs show explicitly the share of ideal performance given by an IR technique and make statistical comparisons possible. The advantage of the P-R based measures is that they treat requests with different number of relevant documents equally, and from the system's point of view the precision at each recall level is comparable. 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 proposed evaluation measures and the case further demonstrate that graded relevance assessments are applicable in IR experiments. The dichotomous and liberal 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 novel measures should be used whenever possible, because they provide richer information for evaluation.

Acknowledgements

This study was funded in part by Academy of Finland under the grant numbers 44703 and 49157. This research was done, in part, while Kal Jarvelin was on a leave at the Department of Information Studies, University of Sheffield, UK, winter 2001. We thank the FIRE group at University of Tampere for helpful comments, and the IR Lab for programming.

References

BLAIR, D.C. & MARON, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM 28, 3 (1985), 289–299.

BORLUND, P. Evaluation of Interactive Information Retrieval Systems. Ph.D. disseration. Åbo University Press (2000).

BORLUND, P. & INGWERSEN, P. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. & Zobel, J. (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York (1998), 324–331.

CONOVER, W.J. Practical Nonparametric Statistics (2nd ed.). John Wiley & Sons, New York (1980).

COOPER, W.S. 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 (1968), 30 – 41.

GREEN, R. The expression of conceptual syntagmatic relationships: A comparative survey. Journal of Documentation 51, 4 (1995), 315–338.

HERSH, W.R. & HICKAM, D.H. An evaluation of interactive Boolean and natural language searching with an online medical textbook. Journal of the American Society for Information Science, 46, 7 (1995), 478–489.

HULL, D. 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. ACM, New York (1993), 349–338.

INGWERSEN, P. & WILLETT, P. An introduction to algorithmic and cognitive approaches for information retrieval. Libri 45 (1995), 160–177.

JÄRVELIN, K. & KEKÄLÄINEN, J. IR evaluation methods for retrieving highly relevant documents. In Belkin, N., Ingwersen, P. & Leong, M-K. (Eds.), Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York (2000), 41-48.

KEEN, E.M. The use of term position devices in ranked output experiments. Journal of Documentation 47, 1 (1991), 1–22.

KEKÄLÄINEN, J. The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval [On-line]. Available: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf. Ph.D. dissertation. Department of Information Studies, University of Tampere (1999).

KEKÄLÄINEN, J. & JÄRVELIN, K. The impact of query structure and query expansion on retrieval performance. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkin-

son & J. Zobel (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York (1998), 130–137.

KEKÄLÄINEN, J. & JÄRVELIN, K. The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. Information Retrieval 1, 4 (2000), 329-344.

KEKÄLÄINEN, J. & JÄRVELIN, K. Using Graded Relevance Assessments in IR Evaluation. To appear in JASIST (2002).

KORFHAGE, R.R. Information storage and retrieval. Wiley & Sons, New York (1997).

LOSEE, R.M. Text retrieval and filtering: Analytic models of performance. Kluwer Academic Publishers, Boston (1998).

MYAENG, S.H. & KORFHAGE, R.R. Integration of user profiles: Models and experiments in information retrieval. Information Processing & Management 26, 6 (1990), 719-738.

POLLACK, S.M. Measures for the comparison of information retrieval systems. American Documentation, 19, 4 (1968), 387-397.

OVER, P. TREC-7 interactive track report [On-line]. Available: http://trec.nist.gov/pubs/trec7/papers/t7irep.pdf.gz (1999).

RAJASHEKAR, T.B. & 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.

ROBERTSON, S.E. & BELKIN, N.J. Ranking in principle. Journal of Documentation 34, 2 (1978), 93–100.

ROCCHIO, J.J. Jr. Document retrieval systems – Optimization and evaluation. Ph.D. dissertation. Harvard Computation Laboratory, Harvard University (1966).

SAKAI, T. & SPARCK JONES, K. Generic summaries for indexing in information retrieval. In Croft, W.B., Harper, D.J., Kraft, D.H. & Zobel, J. (Eds.) Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York (2001), 190–198.

SALTON, G. & MCGILL, M.J. Introduction to modern information retrieval. McGraw-Hill, London (1983).

SARACEVIC, T. KANTOR, P. CHAMIS, A. & 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.

SORMUNEN, E. A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases[On-line]. Available: http://acta.uta.fi/pdf/951-44-4732-8.pdf.. Ph.D. dissertation. Department of Information Studies, University of Tampere (2000).

SORMUNEN, E. Extensions to the STAIRS Study – Empirical Evidence for the Hypothesised Ineffectiveness of Boolean Queries in Large Full-Text Databases. Information Retrieval 4, 3/4 (2001), 257-273.

SPARCK JONES, K. (1974). Automatic indexing. Journal of Documentation 30 (1974), 393–432.

SPINK, A., GEISDORF, H. & BATEMAN, J. From highly relevant to non relevant: examining different regions of relevance. Information Processing & Management 34, 5 (1998), 599–622.

TREC homepage, Data – English Relevance Judgements. Available: http://trec.nist.gov/data/reljudge_eng.html (2001).

VAKKARI, P., & HAKALA, N. Changes in relevance criteria and problem stages in task performance. Journal of Documentation 56 (2000), 540 – 562.

VOORHEES, E. Evaluation by highly relevant documents. In Croft, W.B., Harper, D.J., Kraft, D.H. & Zobel, J. (Eds.) Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York (2001), 74 – 82.

ZOBEL, J. (1998). How reliable are the results of large-scale information retrieval experiments? 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. ACM, New York (1998), 307 – 314.