Cumulated Gain-based Evaluation of IR Techniques

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

Submitted to ACM Transactions on Information Systems 25.1.2002, revised 9.7.2002.

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 using TREC data – sample system run results for 20 queries in TREC-7. As relevance base we used novel graded relevance assessments on a four-point scale. The test 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 often 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 technique for retrieving highly and marginally 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; Kekäläinen & Järvelin, 2002a] 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 Kekäläinen and Järvelin 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[1] and the rank position (determined by the probability of relevance) of a document into account.

The novel measures are first defined and discussed and then their use is demonstrated in a case study on the effectiveness of TREC-7 runs in retrieving documents of various degrees of relevance. 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, and the retrieval results are reported. Section 4 contains discussion and Section 5 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

[1] For a discussion of the degree of relevance and the probability of relevance, see Robertson & Belkin 1978.

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:

$$G' = <3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots >$$

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' = $<3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots>$. 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:

$$\mathrm{DCG}[i] = \begin{cases} \mathrm{CG}[i], \text{ if } i < b \\ \mathrm{CG}[i-1] + \mathrm{G}[i]/{}^{b}\log i, \text{if } i \geq b \end{cases} \tag{2}$$

Note that we must not apply the logarithm-based discount at rank 1 because ${}^{b}\log 1 = 0$. Moreover, we do not apply the discount case for ranks less than the logarithm base (it would give them a boost). This is also realistic, since the higher the base, the lower the discount and the more likely the searcher is to examine the results at least up to the base rank (say 10).

For example, let b = 2. From G' given in the preceding section we obtain DCG' = $<3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \ldots>$.

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 technique 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 $V = <v_1, v_2, \ldots, v_k>$ and $W = < w_1, w_2, \ldots, w_k>$ be two vectors. Their sum is the vector $V + W = <v_1 + w_1, v_2 + w_2, \ldots, v_k + w_k>$. 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 $V = <v_1, v_2, \ldots, v_k>$ by a constant r is the vector $r*V = <r*v_1, r*v_2, \ldots, r*v_k>$. The average vector **AV** based on vectors $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$, is given by the function *avg-vect*(**V**):

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

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:

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

A sample ideal gain vector could be:

I' = <3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, …>

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

$$\textit{norm-vect}(V, I) = \langle v_1/i_1, v_2/i_2, \ldots, v_k/i_k \rangle \qquad (5)$$

For example, based on CG’ and $CG_{\mathbf{I}}'$ from above, we obtain the normalised CG vector nCG’ = *norm-vect*(CG’, $CG_{\mathbf{I}}'$) =

$$\langle 1, 0.83, 0.89, 0.73, 0.62, 0.6, 0.69, 0.76, 0.89, 0.84, \ldots \rangle .$$

The normalised DCG vector nDCG’ is obtained in a similar way from DCG’ and $DCG_{\mathbf{I}}'$. Note that, as a special case, the normalised ideal (D)CG vector is always *norm-vect*(I, I) = $\langle 1, 1, \ldots, 1 \rangle$, 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:

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

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 by explicitly giving the number of documents for which each n(D)CG value holds. P-R curves do not make

the number of documents explicit for given performance and may therefore mask bad performance [Losee 1998].

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 2002a]. 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 limitations of the measures are considered in Chapter 4.

3. Case study: comparison of some TREC-7 results at different relevance levels

We demonstrate the use of the proposed measures in a case study testing runs from TREC-7 ad hoc track with binary and non-binary relevance judgements. 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 TREC-7 data

The seventh Text Retrieval Conference (TREC-7) had an ad hoc track in which the participants produced queries from topic statements – altogether 50 – and run those queries against the TREC text document collection. The collection includes about 528,000 documents, or 1.9 GB data. Participants returned lists of the best 1000 documents retrieved for each topic. These lists were evaluated against binary relevance assessments provided by the TREC organisers (National Institute of Standards and Technology, NIST). Participants were allowed to submit up to three different runs, which could be based on different queries or different retrieval methods. [Voorhees & Harman 1999.]

Ad hoc task had two subtracks, automatic and manual, with different query construction techniques. An automatic technique means deriving a query from a topic statement without manual intervention; manual technique is anything else. [Voorhees & Harman 1999.]

In the case study, we used result lists for 20 topics by five participants from TREC-7 ad hoc manual track. These topics were selected because of the availability of non-binary relevance judgements for them [see Sormunen 2002].[2]

3.2 Relevance judgements

The non-binary relevance judgements were obtained by re-judging documents judged relevant by NIST assessors and about 5 % of irrelevant documents for each topic. The new assessments were made by six Master's students of Information Studies, all of them fluent in English though not native speakers. The relevant and irrelevant documents were pooled, and the

2 The numbers of topics are: 351, 353, 355, 358, 360, 362, 364, 365, 372, 373, 377, 378, 384, 385, 387, 392, 393, 396, 399, 400. For details see http://trec.nist.gov/data/topics_eng/topics.351-400.gz.

judges did not know the number of documents previously judged relevant or irrelevant in the pool. [Sormunen 2002.]

The assumption about relevance in the re-judgement process was topicality. This agrees with the TREC assessments for the ad hoc track: documents are judged one by one; general information with limitations given in the topic's narrative is searched, not details in sense of question answering. New assessments were done on a four-point scale:

- *Irrelevant document*. The document does not contain any information about the topic.
- *Marginally relevant document*. The document only points to the topic. It does not contain more or other information than the topic statement.
- *Fairly relevant document*. The document contains more information than the topic statement but the presentation is not exhaustive. In case of multi-faceted topic, only some of the sub-themes are covered.
- *Highly relevant document*. The document discusses the themes of the topic exhaustively. In case of multi-faceted topics, all or most sub-themes are covered.

Altogether 20 topics from TREC-7 and 18 topics from TREC-8 were re-assessed. In Table 1 the results of re-assessment are shown with respect to the original TREC assessments. It is obvious that almost all originally irrelevant documents were also assessed irrelevant in re-judgement (93.8 %). Of the TREC relevant documents 75 % were judged relevant at some level, and 25 % irrelevant. This seems to indicate that the re-assessors have been somewhat stricter than the original judges. The great overlap in irrelevant documents proves the new assessments reliable. However, in the case study we are not interested to compare the results based on different judgements but to show the effects of utilising non-binary relevance judgements in evaluation. Thus we do not use the original TREC assessments in any phase of the case study.

Levels of relevance	TREC relevant		TREC irrelevant		Total	
	# of docs	%	# of docs	%	# of docs	%
Rel = 0	691	25.0%	2780	93.8%	3471	60.5%
Rel = 1	1004	36.2%	134	4.5%	1138	19.8%
Rel = 2	724	26.1%	40	1.3%	764	13.3%
Rel = 3	353	12.7%	11	0.4%	364	6.4%
Total	2772	100%	2965	100%	5737	100%

Table 1. Distribution of new relevance assessment with relation to original TREC assessments.

In the subset of 20 topics, among all relevant documents (N = 1182), the share of highly relevant documents was 20.1%, the share of fairly relevant documents was 30.5%, and that of marginal documents was 49.4%.

3.3 The application of the evaluation measures

We run the TREC-7 result lists of five participating groups against the new, graded relevance assessments. For the cumulated gain evaluations we tested logarithm bases and handling of relevance levels varied as parameters as follows:

1. We tested different relevance weights at different relevance levels. First, we replaced document relevance levels 0, 1, 2, 3 with binary weights, i.e. we gave documents at level 0 weight 0, and documents at levels 1-3 weight 1 (weighting scheme 0-1-1-1 for the four point scale). Then, we replaced the relevance levels with weights 0, 0, 0, 1, to test the other extreme where only the highly relevant documents are valued. The last weighting scheme, 0, 1, 10, 100, is between the extremes; the highly relevant documents are valued hundred times more than marginally relevant documents, and ten times more than fairly relevant ones. Different weighting on highly relevant documents may affect the relative effectiveness of IR techniques as also pointed out by Voorhees [2001]. The first and last weighting schemes only are shown in graphs because the 0-0-0-1 scheme is very similar to the last one in appearance.
2. The logarithm bases 2 and 10 were tested for the DCG vectors. The base 2 models impatient users, base 10 persistent ones. While the differences in results do not vary markedly with the logarithm base, we show only the results for the logarithm base 2. We also prefer the stricter test condition the smaller logarithm base provides.
3. The average actual CG and DCG vectors were compared to the ideal average vectors.
4. The average actual CG and DCG vectors were normalised by dividing them with the ideal average vectors.

3.4 Cumulated gain

Figure 1 presents the CG vector curves for the five runs at ranks 1 - 100, and the ideal curves. Figure 1a shows the weighting scheme 0-1-1-1, and 1b the scheme 0-1-10-100. In the ranked result list, highly relevant documents add either 1 or 100 points to the cumulated gain; fairly relevant documents add either 1 or 10 points; marginally relevant documents add 1 point; and irrelevant documents add 0 points to the gain.

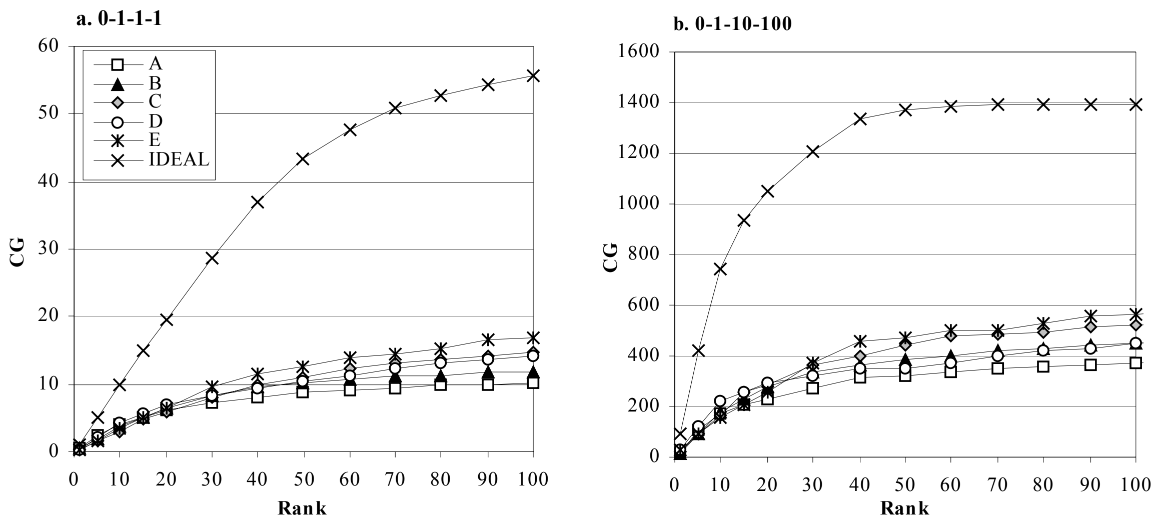

Figure 1. Cumulated gain (CG) curves for five TREC-7 runs and the ideal curve at ranks 1-100.

The different weighting schemes change the position of the curves compared to each other. For example, in Figure 1a – the liberal weighting scheme – the performance of (run) D is close to that of C; when highly relevant documents are given more weight, D is more similar to B, and C and E are close in performance. Note, that the graphs have different scales because of the weighting schemes.

In Figure 1a the best possible curve starts to level off at the rank 100 reflecting the fact that at the rank 100 practically all relevant documents have been found. In Figure 1b it can be observed, that the ideal curve has already found the most fairly and highly relevant documents by the rank 50. This, of course, reflects the sizes of the recall bases – average number of documents at relevance levels 2 and 3 per topic is 29.9. The best system (E) hangs below the ideal by 0 – 39 points with binary weights (1a), and 70 - 894 points with non-binary weights (1b). Note that the differences are not greatest at rank 100 but often earlier. The other runs remain further below by 0 – 6 points with binary weights (1a), and 0 – 197 points with non-binary weights (1b). The differences between the ideal and all actual curves are all bound to diminish when the ideal curve levels off.

The curves can be interpreted also in another way: In Figure 1a one has to retrieve 30 documents by the best run, and 90 by the worst run in order to gain the benefit that could theoretically be gained by retrieving only 10 documents (the ideal curve). In this respect the best run is three times as effective as worst run. In Figure 1b one has to retrieve 35 documents by the best run to get the benefit theoretically obtainable at rank 5; the worst run does not provide the same benefit even at rank 100.

3.5 Discounted cumulated gain

Figure 2 shows the DCG vector curves for the five runs at ranks 1 - 100, and the ideal curve. The $\log_2$ of the document rank is used as the discounting factor. Discounting alone seems to narrow the differences between the systems (1a compared to 2a, and 1b to 2b). Discounting and non-binary weighting changes the performance order of the systems: in Figure 2b, run A seems to lose and run C to benefit.

In Figure 2a, the ideal curve levels off upon the rank 100. The best run hangs below by 0.5 - 10 points. The other runs remain further below by 0.25 - 1 points. Thus, with discounting factor and binary weighting, the runs seem to perform equally. In Figure 2b, the ideal curve levels off upon the rank 50. The best run hangs below by 71 – 408 points. The other runs remain further below by 13 - 40 points. All the actual curves still grow at the rank 100, but beyond that the differences between the best possible and the other curves gradually become stable.

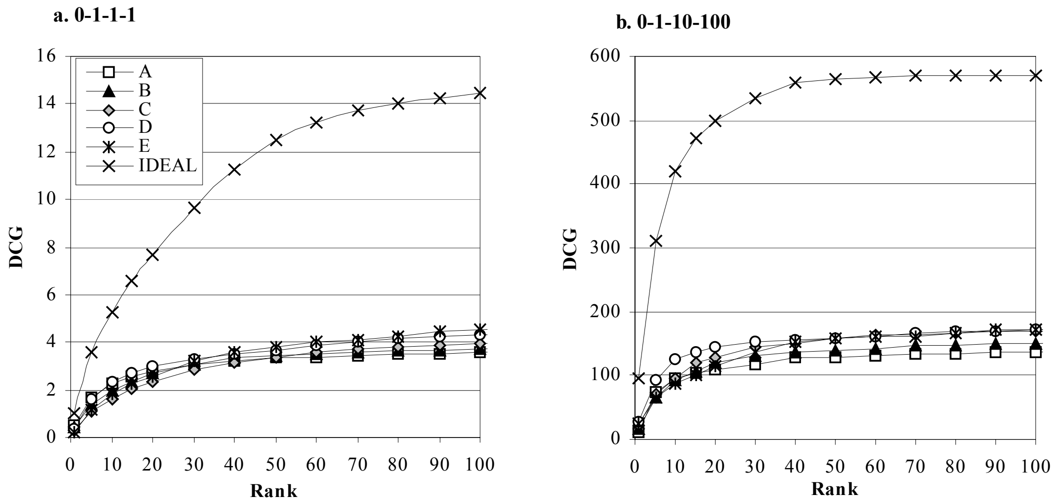

Figure 2. Discounted cumulated gain (DCG) curves for five TREC-7 runs and the ideal curve at ranks 1-100

Also these graphs can be interpreted in another way: In Figure 2a one has to expect the user to examine 40 documents by the best run in order to gain the (discounted) benefit that could theoretically be gained by retrieving only 5 documents. The worst run reaches that gain round rank 95. In Figure 2b, none of the runs gives the gain that would theoretically be obtainable at rank 5. Given the worst run the user has to examine 50 documents in order to get the (discounted) benefit that is obtained with the best run at rank 10. In that respect the difference in the effectiveness of runs is essential.

One might argue that if the user goes down to, say, 50 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. The nCG curves readily show the differences between methods to be compared because of the same scale but they lack the straightforward interpretation of the gain at each rank given by CG curves. In Figure 3b the curves start lower than in Figure 2a – it is obvious that highly relevant documents are more difficult to retrieve.

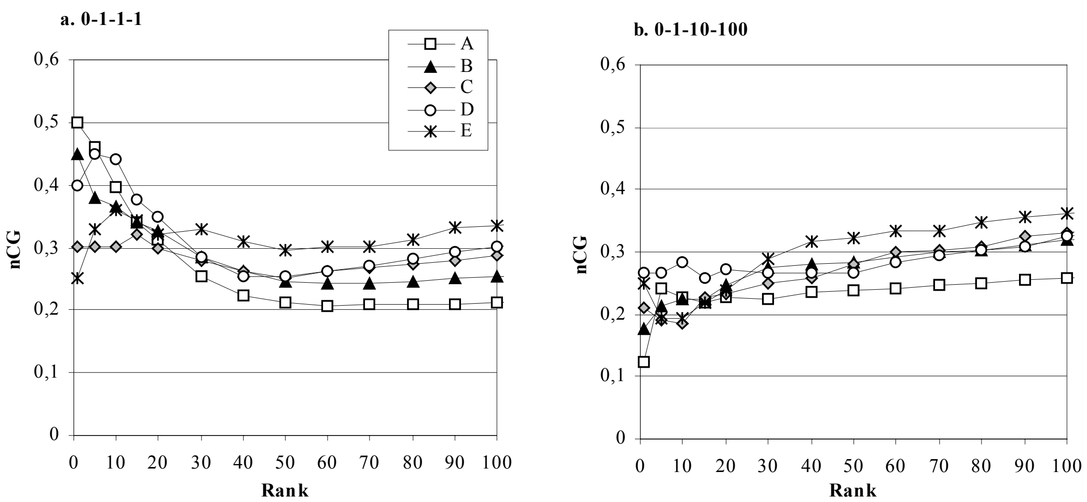

Figure 3. Normalised cumulated gain (nCG) curves for five TREC-7 runs.

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 100. The effect of discounting can be seen by comparing Figures 3 and 4, e.g. the order of the runs changes. The effect of normalisation can be detected by comparing Figure 2 and Figure 4: the differences between the IR techniques are easier to detect and comparable.

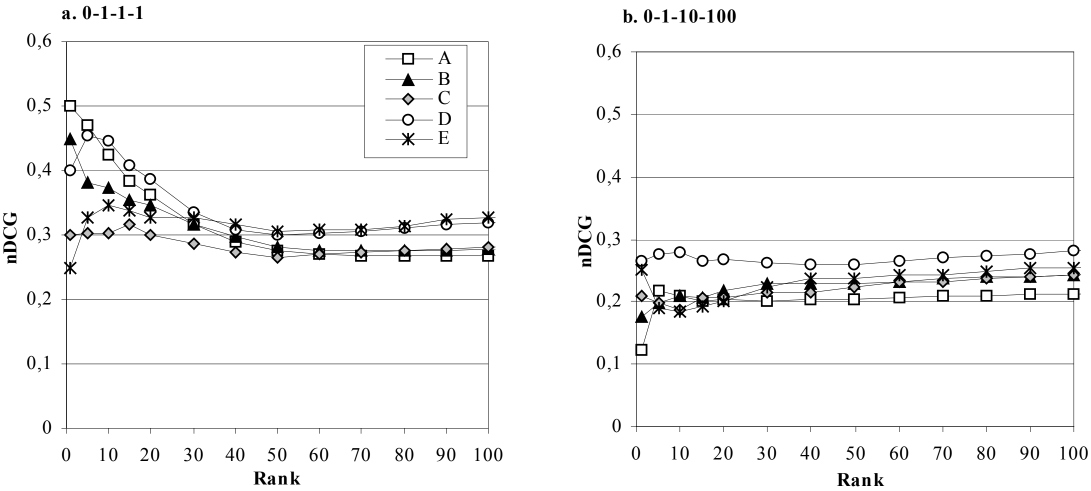

Figure 4. Normalised discounted cumulated gain (nDCG) curves for five TREC-7 runs.

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 document cut-off values. 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 grand averages of the vectors of length 200, and the results the Friedman test, ANOVA did not prove any differences significant.

In the table, the average is first calculated for each topic, then an average is taken over the topics. If the average would have been taken of vectors of different length, the results of the statistical tests might change. Also, the number of topics (20) is rather small to provide reliable results. However, even these data illuminate the behaviour of the (n)(D)CG measures.

	A	B	C	D	E	Statistical significance
nCG (0-1-1-1)	0.242	0.271	0.293	0.318	0.343	
nDCG (0-1-1-1)	0.292	0.294	0.287	0.335	0.331	
nCG (0-1-10-100)	0.254	0.305	0.313	0.316	0.342	D, E > A**
nDCG (0-1-10-100)	0.211	0.238	0.236	0.279	0.247	D > A *
nCG (0-0-0-1)	0.244	0.301	0.309	0.309	0.329	
nDCG (0-0-0-1)	0.192	0.220	0.223	0.259	0.224	

Table 2. n(D)CG averages over topics and statistical significance the results for five TREC-7 runs (legend: ** = p< 0.01; * = p< 0.05, Friedman test).

4 Discussion

The proposed measures are based on several parameters – the last rank considered, the gain values to employ, and discounting factors to apply. An experimenter needs to know which parameter values and combinations to use. In practice, the evaluation context and scenario should suggest these values. Alternatively, several values and/or combinations may be used to obtain a richer picture on IR system effectiveness under different conditions. Below we consider the effects of the parameters. Thereafter we discuss statistical testing, relevance judgements and limitations of the measures.

Last Rank Considered

Gain vectors of various length from 1 to *n* may be used for computing the proposed measures and curves. If one analyzes the curves alone, the last rank does not matter. Eventual differences between the IR methods are observable for any rank region. The gain difference for any point (or region) of the curves may be measured directly.

If one is interested in differences in average gain up to a given last rank, then the last rank matters, particularly for nCG measurements. Suppose IR method A is somewhat better than the method B in early ranks (say, down to rank 10) but beyond them the methods B starts catching up so that they are *en par* at rank 50 with all relevant documents found. If one now evaluates the methods by nCG, they might be statistically significantly different for the ranks 1 – 10, but there probably would be no significant difference for the ranks 1 – 100 (or down to lower positions).

If one uses nDCG in the previous case the difference earned by the method A would be preserved due to discounting low ranked relevant documents. In this case the difference between the methods may be statistically significant also for the ranks 1 – 100 (or down to lower positions).

The measures themselves cannot tell how they should be applied – down to which rank? This depends on the evaluation scenario and the sizes of recall bases. It makes sense to produce the n(D)CG curves liberally, i.e., down to quite low ranks. The significance of differences between IR methods, when present, can be tested for selected regions (top n) when justified by the scenario. Also our test data demonstrate that one run may be significantly better than another, if just top ranks are considered, while being similarly effective as another, if also low ranks are included (say up to 100; see e.g. runs C and D in Figure 3).

Gain Values

Justifying different gain values for documents relevant to different degrees is inherently quite arbitrary. It is often easy to say that one document is more relevant than another, but the quantification of this difference still remains arbitrary. However, determining such documents as equally relevant is another arbitrary decision – and less justified in the light of the evidence from relevance studies [Tang, Shaw & Vevea 1999; Sormunen 2002].

Since graded relevance assessments can be provided reliably, the sensitivity of the evaluation results to different gain quantifications can easily be tested. Sensitivity testing is also typical in cost-benefit studies, so this is no new idea. Even if the evaluation scenario would not advice us on the gain quantifications, evaluation through several flat to steep quantifications informs us on the relative performance of IR methods better than a single one. Voorhees [2001] used this approach in the TREC Web Track evaluation, when she weighted highly relevant documents by factors 1–1000 in relation to marginal documents. Varying weighting affected relative effectiveness order of IR systems in her test. Our present illustrative findings based on TREC data also show that weighting affects the relative effectiveness order of IR systems. We can observe in Figures 4a-b (Section 3.6) that by changing from weighting 0–1–1–1, i.e., flat TREC-type weights, to weights 0–1–10–100 for the irrelevant to highly relevant documents, run D appears more effective than the others.

Tang, Shaw and Vevea [1999] proposed seven as the optimal number of relevance levels in relevance assessments. While our findings are for four levels the proposed measures are not tightly coupled with any particular number of levels.

Discounting Factor

The choice between (n)CG and (n)DCG measures in evaluation is essential: discounting the gain of documents retrieved late affects the order of effectiveness of runs as we saw in Sections 3.4. and 3.5 (Figures 1b and 2b). It is however again somewhat arbitrary to apply any specific form of discounting. Consider the discounting case of the DCG function:

$$DCG[i] = DCG[i - 1] + G[i] / df,$$

where df is the discounting factor and *i* the current ranked position. There are three cases of interest:

- If df = 1 then DCG = CG and no discounting is performed – all documents, at whatever rank retrieved, retain their relevance score.
- If df = *i* then we have a very sharp discount – only the first documents would really matter, which hardly is desirable and realistic for evaluation.
- If df = ${}^{b}\log i$ then we have a smooth discounting factor, the smoothness of which can be adjusted by the choice of the base b. A relatively small base (b = 2) models an impatient searcher for whom the value of late documents drops rapidly. A relatively high base (b > 10) models a patient searcher for whom even late documents are valuable. A very high base (b >100) yields a very marginal discount from the practical IR evaluation point of view.

We propose the use of the logarithmic discounting factor. However, the choice of the base is again somewhat arbitrary. Either the evaluation scenario should advice the evaluator on the base or a range of bases could be tried out. Note that in the DCG function case $DCG[i] = DCG[i - 1] + G[i] / {}^{b}\log i$, the choice of the base would not affect the order of effectiveness of IR methods because ${}^{b}\log i = {}^{b}\log a * {}^{a}\log i$ for any pair of bases a and b since ${}^{b}\log a$ is a constant. This is the reason for applying the discounting case for DCG only after the rank indicated by the logarithm base. This also is the point where discounting begins because ${}^{b}\log b = 1$. In the rank region 2 to b discounting would be replaced by boosting.

There are two borderline cases for the logarithm base. When the base b ($b \geq 1$) approaches 1, discounting becomes very aggressive and finally only the first document would matter – hardly realistic. On the other hand, if b approaches infinity, then DCG approaches CG – neither realistic. We believe that the base range 2 to 10 serves most evaluation scenarios well.

Practical Methodological Problems

The discussion above leaves open the proper parameter combinations to use in evaluation. This is unfortunate but also unavoidable. The mathematics work for whatever parameter combinations and cannot advice us on which to choose. Such advice must come from the evaluation context in the form of realistic evaluation scenarios. In research campaigns such as TREC, the scenario(s) should be selected.

If one is evaluating IR methods for very busy users who are only willing to examine a few best answers for their queries, it makes sense to evaluate down to shallow ranks only (say 30), use fairly sharp gain quantifications (say 0–1–10–100) and a low base for the discounting factor (say 2). On the other hand, if one is evaluating IR methods for patient users who are willing to dig down in the low ranked and marginal answers for their queries, it makes sense to evaluate down to deep ranks (say 200), use moderate gain quantifications (say 0–1–2–3) and a high base for the discounting factor (say 10). It makes sense to try out both scenarios in order to see whether some IR methods are superior in one scenario only.

When such scenarios are argued out, they can be critically assessed and defended for the choices involved. If this is not done, an arbitrary choice is committed, perhaps unconsciously. For example, precision averages over 11 recall points with binary relevance gains models well only very patient users willing to dig deep down the low ranked answers, no matter how relevant vs. marginal the answers are. Clearly this is not the only scenario one should look at.

The normalized measures nCG and nDCG we propose are normalized by the best possible behavior for each query on a rank-by-rank basis. Therefore also the averages of the normalized vectors are less prone to the problems of recall base size variation which plague the precision – recall measurements, whether they are based on DCVs or precision as function of recall.

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. 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 500 % 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.

Relevance Judgements

Kekäläinen & Järvelin [2002a] argue on the basis of several theoretical, laboratory and field studies that the degree of document relevance varies and document users can distinguish between them. Some documents are far more relevant than others are. Further, in many studies on information seeking and retrieval, multiple degree relevance scales have been found pertinent, while the number of degrees employed varies. It is difficult to determine, how many degrees there should be, in general. This depends on the study setting and the user scenarios. When multiple degree approaches are justified, evaluation methods should utilize / support them.

TREC has been based on binary relevance assessments with a very low threshold for accepting a document as relevant for a topical request – the document needs to have at least one sentence pertaining to the request to count as relevant [TREC 2001]. This is a very special evaluation scenario and there are obvious alternatives. In many scenarios, at that level of contribution one would count the document at most as marginal unless the request is factual – in which case a short factual response should be regarded highly relevant and another not giving the facts marginal if not irrelevant. This is completely compatible with the proposed measures. 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. In our data sample, about 50% of the relevant documents were marginally relevant. Possible differences between IR techniques in retrieving highly relevant documents might be evened up by their possible indifference in retrieving marginal documents. The net differences might seem practically marginal and statistically insignificant.

Statistical Testing

Holding to the ordinal scale of relevance judgements suggests non-parametric statistical tests, such as the Wilcoxon test or the Friedman test. 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. In our test case ANOVA gave a result different from Friedman – an effect of the magnitude of the differences between the IR runs considered. However, the data set used in the demonstration was fairly small.

Empirical Findings Based on the Proposed Measures

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 by 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. Our present demonstrative findings based on TREC-7 data also show that weighting affects the relative effectiveness order of IR systems. These results exemplify the usability of the cumulated gain-based approach to IR evaluation.

Limitations

The measures considered in this paper, both the old and the new ones, have weaknesses in three 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 in Section 2.4 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.

Thirdly, any measure based on static relevance judgments is unable to handle dynamic changes in real relevance assessments. However, when changes in user's relevance criteria lead to a reformulated query, an IR system should retrieve the best documents for the reformulated query. Kekäläinen and Järvelin [2002b] argue that complex dynamic interaction is a sequence of simple topical interactions and thus good one-shot performance by a retrieval system should be rewarded in evaluation. Changes in the user's information need and relevance criteria affect consequent requests and queries. While this is likely happen, it has not been shown that this should affect the design of the retrieval techniques. Neither has it been shown that this would invalidate the proposed or the traditional evaluation measures.

It may be argued that IR systems should not rank just highly relevant documents to top ranks. Consequently, they should not be rewarded in evaluation for doing so. Spink, Greisdorf and Bateman [1998] have argued that partially relevant documents are important for users at the early stages of their information seeking process. Therefore one might require that IR systems be rewarded for retrieving partially relevant documents in the top ranks.

For about 40 years IR systems have been compared on the basis of their ability to provide relevant – or useful – documents for their users. To us it seems plausible, that highly relevant documents are those people find useful. The findings by Spink, Greisdorf and Bateman do not really disqualify this belief, they rather state that students in the early states of their information seeking tend to change their relevance criteria and problem definition and that the number of partially relevant documents correlate with these changes.

However, if it should turn out that for some purposes, IR systems should rank partially relevant documents higher than, say, highly relevant documents, our measures suit perfectly comparisons on such basis: the documents should just be weighted accordingly. We do not intend to say how or on what criteria the relevance judgments should be made, we only propose measures that take into account differences in relevance.

The limitations of the proposed measures being similar to those of traditional measures, the proposed measures offer benefits taking the degree of document relevance into account and modeling user persistence.

5. 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 often 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 measures based on the retrieved 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. It may difficult to justify any particular weighting scheme. If the evaluation scenario does not suggest otherwise, several weight values may be used to obtain a richer picture on IR system effectiveness under different conditions. Regarding all at least somewhat relevant documents as equally relevant is also an arbitrary (albeit traditional) decision – and also counter-intuitive.

It may be argued that IR systems should not rank just highly relevant documents to top ranks. One might require that IR systems be rewarded for retrieving partially relevant documents in the top ranks. However, our measures suit perfectly comparisons on such basis: the documents should just be weighted accordingly. The traditional measures do not allow this.

The CG and DCG measures complement P-R based measures [Järvelin & Kekäläinen 2000; Kekäläinen & Järvelin 2002a]. 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. dissertation. 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.

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.

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.

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. Wilkinson & 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. Journal of the American Society for Information Science and Technology (2002a), to appear.

KEKÄLÄINEN, J. & JÄRVELIN, K. Evaluating Information Retrieval Systems under the Challenges of Interaction and Multidimensional Dynamic Relevance. To appear in Proceedings of the CoLIS 4 Conference, Seattle, July 2002 (2002b).

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. In NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7), (1999).
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.
SORMUNEN, E. Liberal relevance criteria of TREC – Counting on negligible documents? To appear in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2002).
SPARCK JONES, K. 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 [On-line]. 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.
VOORHEES, E. & HARMAN, D. Overview of the Seventh Text REtrieval Conference (TREC-7) [On-line]. Available: http://trec.nist.gov/pubs/trec7/papers/overview_7.pdf.gz. In NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7), (1999).

ZOBEL, J. 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.