IR evaluation methods for retrieving highly relevant documents

Kalervo Järvelin & Jaana Kekäläinen
University of Tampere
Department of Information Studies
Finland

Published in: Belkin, N.J., Ingwersen, P. and Leong, M.-K. (eds.) *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, pp. 41–48.

IR evaluation methods for retrieving highly relevant documents

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

This paper proposes evaluation methods based on the use of non-dichotomous relevance judgements in IR experiments. It is argued that evaluation methods should credit IR methods for their ability to retrieve highly relevant documents. This is desirable from the user point of view in modern large IR environments. The proposed methods are (1) a novel application of P-R curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and (2) two novel measures computing the cumulative gain the user obtains by examining the retrieval result up to a given ranked position. We then demonstrate the use of these evaluation methods in a case study on the effectiveness of query types, based on combinations of query structures and expansion, in retrieving documents of various degrees of relevance. 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 tested strong query structures are most effective in retrieving highly relevant documents. The differences between the query types are practically essential and statistically significant. More generally, the novel evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable in IR experiments, may reveal interesting phenomena, and allow harder testing of IR methods.

1. Introduction

Fundamental problems of IR experiments are linked to the assessment of relevance. In most laboratory tests documents are judged relevant or irrelevant with regard to the request. However, binary relevance cannot reflect the possibility that documents may be relevant to a different degree; some documents contribute more information to the request, some less without being totally irrelevant. In some studies relevance judgements are allowed to fall into more than two categories, but only a few tests actually take advantage of different relevance levels (e.g., [6]). More often relevance is conflated into two categories at the analysis phase because of the calculation of precision and recall (e.g., [2, 15]).

In modern large database environments, the number of topically relevant documents to a request may easily exceed the number of documents a user is willing to examine. It would therefore be desirable from the user viewpoint to rank highly relevant documents highest in the retrieval results and to develop and evaluate IR methods accordingly. However, the current practice of liberal binary assessment of topical relevance gives equal credit for a retrieval method for retrieving highly and fairly relevant documents. Therefore differences between sloppy and excellent retrieval methods may not become apparent in evaluation. In this paper, we want to examine the effects of using multiple degree relevance assessments in retrieval method evaluation and to demonstrate, by virtue of a case, that such assessments indeed may reveal important differences between retrieval methods.

The effects of using multiple degree relevance assessments may be evaluated through traditional IR evaluation methods such as P-R curves. In this paper we apply P-R curves in a new way, focusing on retrieval at each relevance level separately. Moreover, to emphasize the user viewpoint, we develop new evaluation measures, which seek to estimate the cumulative relevance gain the user receives by examining the retrieval result up to a given rank. These measures facilitate evaluation where IR methods are credited more / only for highly relevant documents. These novel measures are akin to the average search length (briefly ASL; [12]), ranked half life and relative relevance (briefly RHL and RR; [3]) measures but 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 [14].)

The case demonstrating the effects of multiple degree relevance assessments, and the application of traditional / novel evaluation measures explores query expansion and query structures in probabilistic IR. Kekäläinen [9], and Kekäläinen and Järvelin [11] have earlier observed that the structure of queries influences retrieval performance when the number of search

SIGIR 2000 7/00 Athens, Greece

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

keys in queries is high, i.e., when queries are expanded. Query structure refers to the syntactic structure of a query expression, marked with query operators and parentheses. Kekäläinen and Järvelin classify the structures of best match queries into strong and weak. In the former, search keys are grouped according to concepts they represent; in the latter, queries are mere sets of search keys. They reported significant retrieval improvements with expanded strongly structured queries. However, in their study the relevance assessments were dichotomous.We therefore do not know how different best match query types (based on expansion and structure) are able to rank documents of varying relevance levels. In the case study we investigate their ability to do this.

Section 2 explains our evaluation methodology: the novel application of the P-R curves and 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 Evaluation methods employing multiple degree relevance assessments

2.1 Precision as a function of recall

Average precision over recall levels and P-R curves are the typical ways of evaluating IR method performance. They are normally computed by using dichotomical relevance assessments. Even if the original assessments may have had multiple degrees, these are generally collapsed into two for evaluation. In order to see the difference in performance between retrieval methods, their performance should be evaluated separately at each relevance level. For example, in case of a four point assessment (say, 0 to 3 points), separate recall bases are needed for highly relevant documents (relevance level 3), fairly relevant documents (relevance level 2), and marginally relevant documents (relevance level 1). The rest of the database is considered irrelevant (relevance level 0). In this study, we compiled the recall bases for P-R curve computation in this way.

2.2 Cumulated gain -based measurements

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

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

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

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

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.

Point two leads to comparison of IR methods through test queries by their cumulated gain based on document rank with a rank-based discount factor: the greater the rank, the smaller share of the document value is added to the cumulated gain. The greater the ranked position of a relevant document – of any relevance level – the less valuable it is for the user, because the less likely it is that the user will examine the document due to time, effort, and cumulated information from documents already seen. A discounting function is needed which progressively reduces the document value as its rank increases but not too steeply (e.g., as division by rank) to allow for user persistence in examining further documents. A simple way of discounting with this requirement is to divide the document value by the log of its rank. For example $^2\log 2 = 1$ and $^2\log 1024 = 10$, thus a document at the position 1024 would still get one tenth of 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' 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 ith component in the individual vectors. The averaged vectors can directly be visualised as gain-by-rank -graphs.

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. Then compute CG and DCG as well as the average CG and DCG vectors and curves as above. Note that the curves turn horizontal when no more relevant documents (of any level) can be found. They do not unrealistically assume as a baseline that all retrieved documents could be maximally relevant. The vertical distance between an actual (average) (D)CG curve and the theoretically best possible curve shows the effort wasted on less-than-perfect documents due to a particular IR method.

The CG measure has several advantages when compared with the average search length (ASL) measure [12] or the RR and RHL measures [3]:

1. It combines the degree of relevance of documents and their rank (affected by their probability of relevance) in a coherent way. The RR is based on comparing the match between the system-dependent probability of relevance and the user-assessed degree of relevance. The ASL measure is dichotomical.
2. At any number of retrieved documents examined (rank), it gives an estimate of the cumulated gain as a single measure no matter what is the recall base size. The ASL measure only gives the average position of a relevant document for a given recall base. The RHL measure gives the median point of accumulated relevance for a given query result, which may be the same for quite differently performing queries.
3. It is not heavily dependent on outliers (relevant documents found late in the ranked order) since it focuses on the gain cumulated from the beginning of the result. The ASL and RHL are dependent on outliers although RHL is less so.
4. It is obvious to interpret, it is more direct than P-R curves, and it does not mask bad performance. The RHL alone is not sufficient as a performance measure.

In addition, the DCG measure has the following further advantages not provided by the ASL or RHL measures:

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

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 structured queries in a database with non-binary relevance judgements. Based on the results by Kekäläinen and Järvelin [11] we already know that weak query structures are not able to benefit from query expansion whereas the strong ones are. In the present study we shall test whether the performance of differently structured queries varies with relation to the degree of relevance. We give the results as traditional P-R curves for each relevance level, and as CG and DCG curves which exploit the degrees of relevance. We hypothesize that expanded queries based on strong structures are better able to rank highly relevant documents high in the query results than unexpanded queries or queries based on other structures, whether expanded or not. Consequently, the performance differences between query types among marginally relevant documents should be marginal and among highly relevant documents essential. Expanded queries based on strong structures should cumulate higher CG and DCG values than unexpanded queries or queries based on other structures, whether expanded or not.

3.1 Test environment

The test environment was a text database containing newspaper articles operated under the InQuery retrieval system (version 3.1). The database contains 53,893 articles published in three different newspapers. The database index contains all keys in their morphological basic forms, and all compound words are split into their component words in their morphological basic forms. For the database there is a collection of requests, which are 1 - 2 sentences long, in the form of written information need statements. For these requests there is a recall base of 16,540 articles which fall into four relevance categories (see below *Relevance assessments*). The base was collected by pooling the result sets of hundreds of different queries formulated from the requests in different studies, using both exact and partial match retrieval. We thus believe that our recall estimates are valid. For a set of tests concerning query structures, 30 requests were selected on the basis of their expandability, i.e., they provided possibilities for studying the interaction of query structure and expansion. [9, 10, 17.]

The InQuery system was chosen for the test, because it has a wide range of operators, including probabilistic interpretations of the Boolean operators, and it allows search key weighting. InQuery is based on Bayesian inference networks. For details of the InQuery system, see [1, 13, 18].

3.2 Relevance assessments

For the test requests and test collection of the present experiment, relevance was assessed by four persons, two experienced journalists and two information specialists. They were given written information need statements (requests), and were asked to judge the relevance on a four level scale: (0) irrelevant, the document is not about the subject of the request, (1) marginally relevant, the topic of the request is mentioned, but only in passing, (2) fairly relevant, the topic of request is discussed briefly, (3) highly relevant, the topic is the main theme of the article. The relevance of 20 requests (of 35) was assessed by two (one by three) persons, the rest by one person. The assessors agreed in 73% of the parallel assessments, in 21% of the cases the difference was one point, and in 6% two or three points. If the difference was one point, the assessment was chosen from each judge in turn. If the difference was two or three points, the article was checked by the researcher to find out if there was a logical reason for disagreement, and a more plausible alternative was selected. [9, 17.]

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

3.3 Query structures and expansion

In text retrieval an information need is typically expressed as a set of search keys. In exact match – or Boolean – retrieval relations between search keys in a query are marked with the AND operator, the OR operator, or proximity operators which, in

fact, are stricter forms of the AND operator. Thus, the query has a structure based on conjunctions and disjunctions of search keys. [5, 8.] A query constructed with the Boolean block search strategy (a query in the conjunctive normal form), is an example of a facet structure. Within a facet, search keys representing one aspect of a request are connected with the OR operator, and facets are connected with the AND operator. A facet may consist of one or several concepts.

In best match retrieval, matching is ranking documents according to scores calculated from the weights of search keys occurring in documents. These weights are typically based on the frequency of a key in a document and on the inverse collection frequency of the documents containing the key (tf*idf weighting). [7.] In best match retrieval, queries may either have a structure similar to Boolean queries, or queries may be 'natural language queries' without differentiated relations between search keys.

Kekäläinen and Järvelin [11] tested the co-effects of query structures and query expansion on retrieval performance, and ascertained that the structure of the queries became important when queries were expanded. The best performance overall was achieved with expanded, facet structured queries. For the present study, we selected their best weak structure (SUM) and two of their best strong structures, one based on concepts (SSYN-C) and another based on facets (WSYN). SUM queries may be seen as typical 'best match' queries and therefore suitable as a baseline.

In query formulation, researchers identified search concepts from requests and elicited corresponding search keys from a test thesaurus containing more than 1000 concepts and more than 1500 expressions for the domains of the test requests (see [9]). In QE, search keys that were semantically related (synonyms, hierarchies, associations) to the original search concepts in the test thesaurus were added to queries. This procedure gave unexpanded (u) and expanded (e) query versions, which both were formulated into different query structures.

The structures used to combine the search keys are exemplified in the following. Examples are based on a sample request *The processing and storage of radioactive waste.* In the following samples queries are expanded, the expressions of the unexpanded queries are in italics.

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

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

In a SUM-of-synonym-groups-query (*SSYN-C*) each search concept forms a clause with the SYN operator. SYN clauses were combined with the SUM operator. Phrases were used (marked with *#3*). All keys within the SYN operator are treated as instances of one key [13].

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

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

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

3.4 Test queries and the application of the evaluation measures

In the queries for the 30 test requests, the average number of facets was 3.7. The average number of concepts in unexpanded queries was 4.9, and in expanded queries 26.8. The the number of search keys of unexpanded queries when no phrases were marked (i.e., SUM structure) was 6.1 on average, and for expanded queries without phrases, on average, 62.3. The number of search keys with phrases (i.e., SSYN-C, and WSYN structures) was 5.4 for unexpanded queries, and 52.4 for expanded queries, on average.

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

We present the analysis of the search results in two forms: First, we apply the conventional measures in the form of P-R curves. We also calculated precision after each retrieved relevant document and took an average over requests (average noninterpolated precision, AvP for short). We chose AvP rather than precision based on document cut-off values, because the sizes of recall bases vary at different relevance levels, and thus one cut-off value will not treat queries equally with relation to precision. The statistical significance of differences in the effectiveness of query types was established with the Friedman test (see [4]).

Second, we present the CG and DCG curves. For the cumulative 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:

1. The logarithm bases 2, *e,* and 10 were tested for the DCG vectors. The base 2 models impatient users, base 10 persistent ones.
2. We used document relevance levels 0 - 3 directly as gained value measures. This can be criticised, e.g., by asking whether a highly relevant document is (only) three times as valuable as a marginally relevant document. Nevertheless, even this gives a clear difference for document quality to look at.
3. We first took all documents at relevance levels 1 - 3 into account, secondly nullified the values of documents at relevance level 1 (to reflect that they practically have no value), and finally nullified the values of documents at relevance levels 1 - 2 in order to focus on the highly relevant documents.

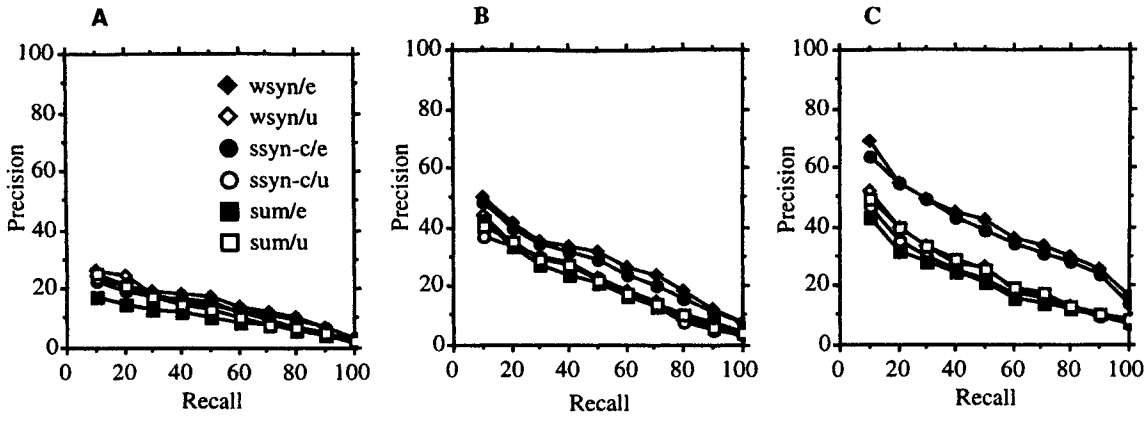

Figure1. P-R curves of SUM, SSYN-C, and WSYN queries at relevance levels 1 (A), 2 (B), and 3 (C).

4. The average actual CG and DCG vectors were compared to the theoretically best possible average vectors.

3.5 P-R curves and average precision

Figure 1 presents the P-R curves of the six query types at different relevance levels. At the relevance level 1, the curves are almost inseparable. At the relevance level 2, expanded WSYN and SSYN-C queries are more effective than the other query types. At the relevance level 3, the difference is even more accentuated. The higher the relevance level is, the greater are the differences between the best and the worst query types.

In Table 1 the average precision (AvP) figures are given. It can be seen that QE never enhances the average precision of SUM queries. In contrast, QE always improves the average precision of strongly structured queries. When queries are unexpanded the differences in precision are negligible within each relevance level. The best effectiveness over all relevance levels is obtained with expanded WSYN queries. At the best, the difference in average precision between unexpanded SUM and expanded WSYN queries is at the relevance level 3 (AvP: a change of 15.1 percentage units or an improvement of 58.3 %). In other words, expanded queries with strong structure are most effective in retrieving the most relevant documents.

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

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

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

3.6 Cumulative gain

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

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

3.7 Discounted cumulative gain

Figure 3 shows the DCG vector curves for ranks 1 - 50, the six query types studied above and the theoretically best possible (average) query. The $\log_2$ of the document rank is used as the discounting factor. Figure 3A shows the curves when documents both at the relevance levels 2 and 3 are taken into account. The best possible curve still grows at the rank 50 (it levels off at the rank 90). The two best (synonym structured)

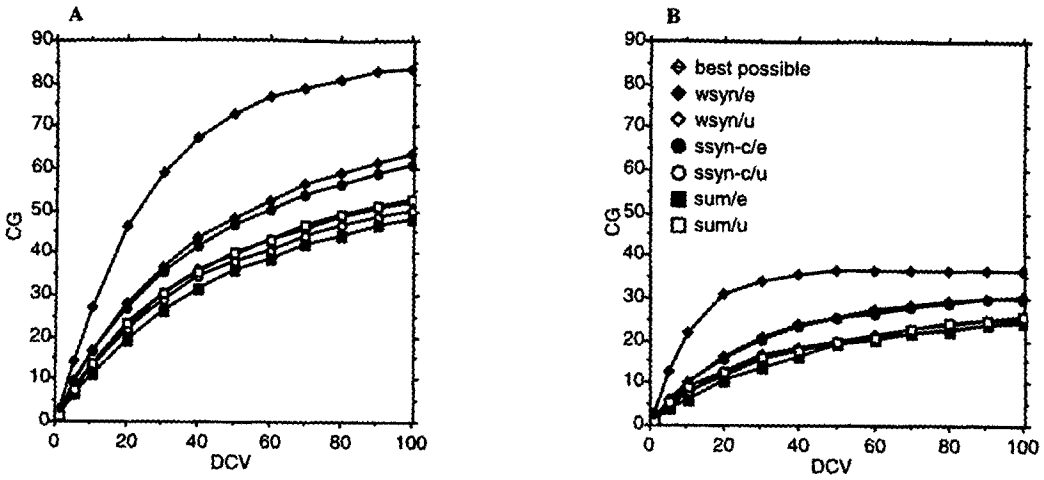

Figure 2. Cumulative gain curves at ranks 1-100, relevance levels 2&3 (A), and 3 (B).

query types hang below by 5 - 9 points (35 - 36 %) from the rank 10 to 50. The difference is growing. The other four query types remain further below by 2 - 4 points (15 - 27 %) from rank 10 to 50. The difference to the best possible curve is 7 - 13 points (47 - 50 %). Beyond the rank 50 the differences between the best possible and all actual curves gradually become stable. Figure 3B shows the curves when documents only at the relevance level 3 considered. The precise figures are different and the absolute diferences smaller. However, the proportional differences are larger. At the greatest, the difference between the best and the remaining query types is 3 points (or one level - 3 document) at the rank 40 and further. It is a consistent and statistically significant difference but are the users able to notice it?

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

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

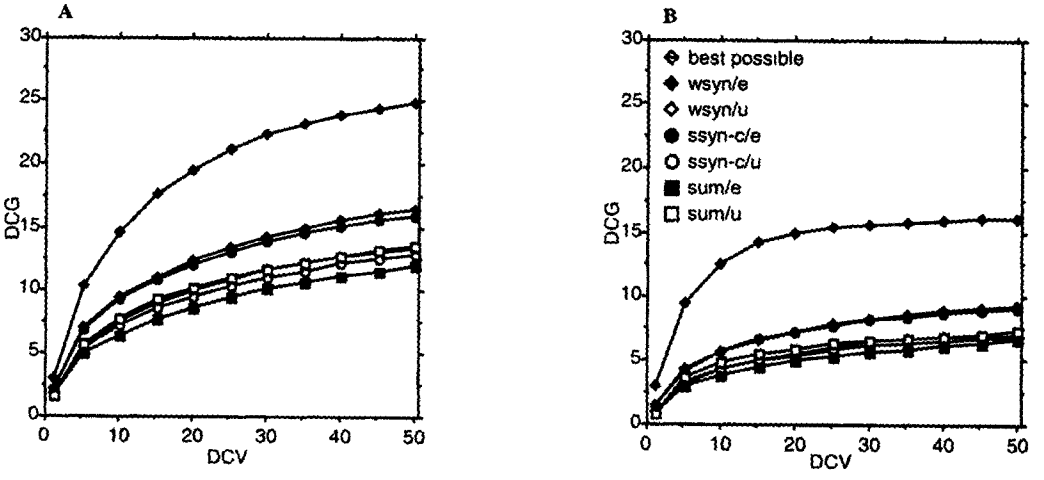

Figure 3. Discounted ($\log_2$) cumulative gain curves ranks 1-50, relevance levels 2&3 (A), and 3 (B).

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

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 methods. We then developed two methods for IR method evaluation, which aim at taking the document relevance degrees into account. One is based on a novel application of the traditional P-R curves and separate recall bases for each relevance level of documents. The other is based on two novel evaluation measures, the CG and the DCG measures, which give the (discounted) cumulative gain up to any given document rank in the retrieval results. Both measures systematically combine document rank (based on its probability of relevance) and degree of relevance.

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

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

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

The P-R curves demonstrate that the good performance of the expanded structured query types is due to, in particular, their ability to rank the highly relevant documents toward the top of retrieval results. The cumulative 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 50 to 100 % more documents by the worse query types to collect the same gain collected by the best query types. While it is possible that persistent users go way down the result list, e.g., from 30 to 60 documents, it often is unlikely to happen, and a system requiring such a behaviour is, in practice, much worse than a system yielding the gain within a 50 % of the documents.

The novel CG and DCG measures complement the modified P-R measure. Precision over fixed recall levels hides the user's effort up to a given recall level. The DCV-based precision - recall curves are better but still do not make the value gained by ranked position explicit. The CG and DCG curves provide this directly. The distance to the theoretically best possible curve shows the effort wasted on less-than-perfect or useless documents. The advantage of the P-R measure is that it treats requests with different number of relevant documents equally, and from the system's point of view the precision at each recall level is 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 evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable even in IR experiments, and may reveal interesting phenomena. The dichotomous relevance assessments generally applied may be too permissive, and, consequently, too easily give credit to IR system performance. We believe that, in modern large environments, the proposed modified P-R measure and the novel (D)CG measures should be used whenever possible, because they provide richer information for evaluation.

Acknowledgements.

This study was funded in part by Academy of Finland under the grant number 44703. We thank the FIRE group at University of Tampere, especially Heikki Keskustalo and Eero Sormunen, for helpful comments, and Heikki Keskustalo and Timo Tervola for programming efforts for data analysis.

References

[1] J. Allan, J. Callan, B. Croft, L. Ballesteros, J. Broglio, J. Xu & H. Shu. INQUERY at TREC 5. In E.M. Voorhees & D.K. Harman (Eds.), *Information technology: The Fifth Text Retrieval Conference (TREC-5).* Gaithersburg, MD: National Institute of Standards and Technology, 119–132, 1997.

[2] D.C. Blair, & M.E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM,* 28(3): 289–299, 1985.

[3] P. Borlund & P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval. New York: ACM, 324–331, 1998.

[4] W.J. Conover. *Practical nonparametric statistics* (2nd ed.). New York: John Wiley & Sons, 1980.

[5] R. Green. The expression of conceptual syntagmatic relationships: A comparative survey. *Journal of Documentation,* 51(4): 315–338, 1995.

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

[7] P. Ingwersen & P. Willett. An introduction to algorithmic and cognitive approaches for information retrieval. *Libri,* 45(): 160–177, 1995.

[8] E.M. Keen. The use of term position devices in ranked output experiments. *Journal of Documentation,* 47(1): 1–22, 1991.

[9] J. Kekäläinen. *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval.* Ph.D. dissertation. Department of Information Studies, University of Tampere, 1999.

[10] J. Kekäläinen & K. Järvelin. The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval,* 1(4): 329-344, 2000.

[11] J. Kekäläinen & K. Järvelin. The impact of query structure and query expansion on retrieval performance. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM, 130–137, 1998.

[12] R.M. Losee. *Text retrieval and filtering: Analytic models of performance.* Kluwer Academic Publishers: Boston, 1998.

[13] T.B. Rajashekar & W.B. Croft. Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science,* 46(4): 272–283, 1995.

[14] S.E. Robertson & N.J. Belkin. Ranking in principle. *Journal of Documentation,* 34(2): 93–100, 1978.

[15] T. Saracevic, P. Kantor, A. Chamis & D. Trivison. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science,* 39(3): 161–176, 1988.

[16] S. Smithson. Information retrieval evaluation in practice: A case study approach. *Information Processing & Management,* 30(2): 205–221, 1994.

[17] E. Sormunen. *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases.* Ph.D. dissertation. Department of Information Studies, University of Tampere, 2000.

[18] H.R. Turtle. *Inference networks for document retrieval.* Ph.D. dissertation. Computer and information Science Department, University of Massachusetts, 1990.