Authors:	Järvelin Kalervo, Price Susan L, Delcambre Lois ML, Nielsen Marianne L
Name of article:	Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions
Name of work:	Advances in Information Retrieval. Proceedings of the 30th European Conference on Information Retrieval (ECIR 2008)
Editors of work:	Macdonald Craig et al.
Year of publication:	2008
ISBN:	978-3-540-78645-0
Publisher:	Springer
Pages:	4-15
Series name and number:	Lecture Notes in Computer Science 4956/2008
ISSN:	0302-9743
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URL: http://www.springerlink.com/content/nw6385u2333wl7r2/fulltext.pdf
URN: http://urn.fi/urn:nbn:uta-3-868
DOI: http://dx.doi.org/10.1007/978-3-540-78646-7_4

Preprint from: Järvelin, K. & Price, S. & Delcambre, L. & Nielsen, M. (2008). Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions. In: Ruthven, I. & al. (Eds.), Proc. of the 30th European Conference on Information Retrieval (ECIR 2008), Glasgow, April 2008. Heidelberg: Springer, Lecture Notes in Computer Science vol. 4956, pp. 4-15. *Please cite the final version. (c) Springer.*

Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions

Kalervo Jarvelin[1], Susan L. Price[2], Lois M. L. Delcambre[2], and Marianne Lykke Nielsen[3]

[1]University of Tampere, Finland
[2]Portland State University, USA
[3]Royal School of Library and Information Science, Denmark
kalervo.jarvelin@uta.fi, prices@cs.pdx.edu, lmd@cs.pdx.edu, mln@db.dk

Abstract. IR research has a strong tradition of laboratory evaluation of systems. Such research is based on test collections, pre-defined test topics, and standard evaluation metrics. While recent research has emphasized the user viewpoint by proposing user-based metrics and non-binary relevance assessments, the methods are insufficient for truly user-based evaluation. The common assumption of a single query per topic and session poorly represents real life. On the other hand, one well-known metric for multiple queries per session, instance recall, does not capture early (within session) retrieval of (highly) relevant documents. We propose an extension to the Discounted Cumulated Gain (DCG) metric, the Session-based DCG (sDCG) metric for evaluation scenarios involving multiple query sessions, graded relevance assessments, and open-ended user effort including decisions to stop searching. The sDCG metric discounts relevant results from later queries within a session. We exemplify the sDCG metric with data from an interactive experiment, discuss how the metric might be applied, and present research questions for which the metric is helpful.

1 Introduction

IR research has a strong tradition of laboratory evaluation of IR systems. Such research is based on test collections, pre-defined test topics, and standard evaluation metrics. While recent research has emphasized the user viewpoint by proposing user-based metrics and non-binary relevance assessments, the methods are insufficient for truly user-based evaluation. Much of the evaluation literature assumes a single query per topic and session, which poorly represents real life.

User-based IR research seeks to attain more realism in IR evaluation [3]. For example, precision at recall = 10% or precision at various document cut-off values (DCV) both seek to account for searchers who choose to scan only a subset of the complete result list. The Discounted Cumulated Gain (DCG) [4] [5] takes a different approach by discounting the value of documents ranked further down in the result list. DCG also supports evaluation by graded relevance assessments. But these metrics as well as traditional IR evaluation metrics assume one query per topic/session. In real life, interactive searchers often issue multiple queries using reformulation [1] and/or relevance feedback until they are satisfied or give up. Evaluation metrics that assume one query per topic are insufficient when the searcher's reformulation effort matters.

The TREC conferences introduced *instance recall* for evaluating interactive experiments [7]. This metric allows multiple queries per session as it rewards for the number of distinct relevant answers identified in a session of a given length. However, it does not reward a system (or searcher) for finding pertinent documents early in the session nor does it help to analyze which queries in a sequence are the most effective. The experiments based on instance recall set a fixed session time and a recall-oriented task. In real life, some tasks are precision-oriented due to time pressure. Stopping decisions often depend on the task, the context, personal factors, and the retrieval results [8]. In the present paper we address issues in session-based evaluation.

We approach session-based IR evaluation with the view that, for a given real search situation, (1) a searcher's information need may be muddled as there is no predefined topic to search on, (2) the initial query formulation may not be optimal, (3) his/her need may remain more or less stable, (4) he/she may switch focus, (5) he/she may learn as the session progresses, (6) highly relevant documents are desired, and (7) stopping decisions depend on search tasks and may vary among individuals [3]. Moreover, it is reasonable that (8) examining retrieval results involves a cost, (9) providing feedback or revising the query involves a cost, and (10) costs should be reflected as penalties in the evaluation. A metric allowing systematic testing under these conditions is needed.

We extend the Discounted Cumulated Gain (DCG) into a new, session-based metric for multiple interactive queries. DCG assigns a gain value to each retrieved document of a ranked result and then cumulates the gains from the first document position onwards to the rank of interest in each test design. DCG allows flexible evaluation under various evaluation scenarios through relevance weighting (see also [6] [9]) and document rank-based discounting. Unlike many traditional effectiveness measures, such as MAP, DCG can easily be extended to a session-based DCG (sDCG) metric, which incorporates query sequences as another dimension in evaluation scenarios and allows one to further discount relevant documents found only after additional searcher effort, i.e., feedback or reformulation. The contributions of this paper are to:

- Define the sDCG metric and describe a method for concatenating results from multiple queries into a single discounted cumulated gain for a session.
- Discuss the research questions that this metric can help answer for which there are no suitable existing metrics.
- Provide guidelines for when and how the sDCG metric can be applied.
- Exemplify the metric with data from a real interactive experiment.
- Discuss the contributions of the metric and the challenges of evaluating it.

Section 2 modifies the DCG metric slightly, making it more elegant and principled and then defines sDCG. Section 3 discusses the features, uses, and evaluation of sDCG and illustrates use of sDCG with data from an interactive IR experiment. Section 4 discusses our findings. The Appendix presents mathematical formulas used in defining sDCG. The focus of this paper is on methodological aspects, not on empirical findings per se, which instead serve as an illustration.

2 Cumulated Gain Based Metrics for Queries and Sessions

2.1 Discounted Cumulated Gain

Järvelin and Kekäläinen [4] [5] argue that highly relevant documents are more valuable than marginally relevant documents and that the searcher may reasonably be assumed to scan the result from its beginning up to some point before quitting. Accordingly, they define the cumulated gain (CG) metrics to produce a gain vector based on the ranked retrieved list, where each document is represented by its (possibly weighted) relevance score up to a ranked position *n* set for the experiment. The authors argue that the greater the ranked position of a relevant document, the less valuable it is for the searcher, because the searcher is less likely to examine the document due to time, effort, and cumulated information from documents already seen. This leads to "correcting" the readings provided by cumulated gain by a rank-based discount factor, the logarithm of the rank of each document. The normalized (discounted) cumulated gain is calculated as the share of ideal performance an IR technique achieves. The Appendix gives formulas for cumulated gain, and for discounting and normalizing it. The benefits of the CG, DCG, and nDCG metrics were discussed thoroughly in comparison to several earlier metrics in [5]. This discussion is not repeated here.

Compared with [5], the definition of the DCG presented here contains a notable modification making it more elegant and principled in discounting early relevant documents. The original formulation employed CG up to the rank of the base of discounting logarithm and only thereafter discounted the value of relevant documents. The formulation presented here is simpler and systematic in discounting the value of all relevant documents including the early ones not discounted by the original DCG.

2.2 Discounting over a Query Sequence within a Session

A session consists of a sequence of queries, each producing a ranked result. Each query formulation requires some effort by the searcher and therefore the results gained by reformulated queries are progressively less valuable. A DCG vector representing the *q*th query in sequence is discounted by a factor, which is based on the position of the query. The base of the logarithm *bq* may be set to model varying searcher behavior: a small base, say *bq* = 2, for an impatient or busy searcher, who is unlikely to reformulate queries or issue novel ones, and a larger base, say *bq* = 10, for a patient searcher willing to probe the document space with several reformulations. sDCG uses the DCG metric to discount the gain within each query and further discounts its gain by a factor dependent on the sequence number of the query within the session. Let DCG be the ordinary DCG vector for the result of the *q*th query. The session-based discounted vector for the *q*th query is:

$$\mathrm{sDCG}(q) = (1 + \log_{bq} q)^{-1} * \mathrm{DCG} \qquad (1)$$

where $bq \in \mathbf{R}$ is the logarithm base for the query discount; $1 < bq < 1000$
q is the position of the query.

Each session-based discounted vector sDCG(q) is a vector representing query performance for one query in the session. Thus it may be normalized like any ordinary DCG vector by the ideal vector and such normalized vectors can be concatenated to represent an entire session.

3 Application of sDCG: An Example

The sDCG metric evaluates entire interactive multiple query sessions. Because sessions are products of the search task, the searcher, the retrieval system, and the collection, experiments can be designed to evaluate any combination of these. For example, if search tasks and searchers are appropriately randomized and the collection is held constant, as we do below, one may evaluate the performance of search systems in interactive sessions. Current single query evaluation metrics require unnatural tricks (like freezing) in evaluation because there is no user in place to act.

We believe that something like the sDCG is needed because it has reasonable and intuitive behavior:

- documents at equivalent ranks are valued more highly if returned by an earlier query
- there is smooth discounting of both document rank and query iteration
- the parameters are understandable and reflect recognizable searcher and setting characteristics

Setting the parameters for each evaluation case must be based on either general findings on searcher behavior, specific findings on searcher behavior in the context of interest, or simulation scenarios where the sensitivity of retrieval performance to a range of searcher behaviors is of interest. The sDCG metric allows the experimenter to adjust the evaluation to reflect each setting evaluated.

The important contribution of sDCG is the new information and insight that other metrics do not deliver. We assess the validity of the metric by referring to its behavior in the light of available knowledge on real-life interactive searching. Note that there are no existing session-based metrics to compare to as a standard. For example, instance recall measures very different phenomena and requires tasks of specific kind. There are no existing test collections for interactive searching with multiple queries per session. One needs a searcher to interact with the system and collection, and to produce the queries. Thus, one cannot test the metric on a known collection against a known metric to see if it produces the expected system ranking as might be done with metrics for traditional laboratory-based IR using the TREC collections.

We now illustrate use of the new metric by analyzing data from a user-based experiment. We introduce the test subjects, task and systems in Section 3.1. In Section 3.2, we use sDCG to analyze query effectiveness by query rank across the test searchers and systems. We also consider session effectiveness across the test searchers and systems, rewarding sessions for early finding of highly relevant documents in Section 3.3.

3.1 Sample Data

We show sample data from an empirical, interactive searching study that compared two search systems. Thirty domain experts (family practice physicians) each completed the same four realistic search scenarios that simulated a need for specific information required to make a decision in a short time frame of several minutes. Each scenario formed a separate session. The searchers had a mean of 21 years of experience in medicine and were also experienced searchers, with a mean of 7 years of Internet searching experience and over 2 year's experience with the test collection, sundhed.dk (Table 1). On a Likert scale (1-5), the average of their self-assessed searching skills was 2.4.

Table 1. Searcher features (*N*=30)

Feature	Average	Standard Deviation
Experience using Internet search engines (years)	7.2	± 2.8
Experience in using sundhed.dk (years)	2.4	± 1.4
Searching experience (low=1; high=5)	2.4	± 0.9
Professional experience in medicine (years)	21.4	± 7.6

We asked searchers to simulate a real-life situation by searching only as long as they would in a real setting. The searchers entered queries and examined results until either finding relevant documents that, in their opinion, satisfied the information need in the scenario or until they judged the search a failure. We also asked them to make graded relevance judgments when they viewed a document. All documents judged relevant by at least one user were judged by an independent domain expert to develop the reference standard we used for the calculations we show in this paper.

The two search systems, Systems 1 and 2, operated over the same collection of nearly 25,000 documents. System 1 used a combination of full text and keyword indexing. System 2 used the existing indexing plus a new form of supplemental document indexing, Semantic Components [8], that affected both the query language and document ranking. Each participant searched on two scenarios with each experimental search system, resulting in 15 sessions per scenario–system combination. The order of exposure to the search scenarios and the systems was randomized (a Latin Square design [2]). A more detailed description of the searching study, using traditional metrics, is available [8].

Table 2. The number of sessions (*N*=60 per system) issuing exactly 1 - 11 queries across four search tasks in Systems 1 and 2 and the average number of queries per session

Sessions in	Number of Queries											Avg per
(*N*=60)	1	2	3	4	5	6	7	8	9	10	11	Session
System 1	28	7	10	4	5	4	1	1	0	0	0	2.53
System 2	21	7	9	11	2	4	2	2	0	1	1	3.18

The test searchers constructed 1 to 11 queries for their search tasks for a total of 343 queries. The number varied by system and topic – most searchers quit as soon as they had a reasonably good result. Table 2 shows how many sessions constructed exactly 1 to 11 queries in each system and the average number of queries per session.

In the illustration below, we have several parameters at our disposal:

- Relevance assessments and weights: we use a four point scale (scores 0 to 3, from non-relevant to highly relevant) given by a domain expert; weighted as 0-1-10-100.
- The rank-based document discount – the log base b: a reasonable range is $1.5 \leq b \leq 10$ to reflect impatient to patient searcher; we use 2 to reflect impatient searchers or time-constrained task scenarios.
- The rank-based query discount – the log base *bq*: we believe that a reasonable range is $1.5 \leq b \leq 10$ to reflect impatient to patient searchers; we use 4 to reflect a impatient searchers or time-constrained task scenarios.
- Stopping – gain vector length: when looking at the effectiveness of queries by their rank, we examine the top-100. When we analyze the gain of an entire session, we concatenate the top-10 of each query in sequence, assuming that a searcher would rather switch to reformulation than continue scanning beyond 10 documents.

We test only some value combinations in the present study.

3.2 Effectiveness by Query Order

Most searchers quit after finding one reasonably good result; only a few continued beyond that point: sometimes they found more (or the same) relevant documents. This searcher behavior is shown clearly in Figure 1, which compares the discounted average gain of the last query in each session (LQ) to the average of the preceding ones (Non-last Q) in Systems 1 and 2. Until their last query the searchers gain little. The last query for searchers in System 2 tends to be somewhat better than in System 1. The last query performance levels off at rank 20 in both systems. The last query was nearly always the best – but not always; a corresponding analysis could be made on the best query of each session, but this does not change the basic result.

The initial query performance suggested by Figure 1 seems poor from the laboratory IR perspective. However, laboratory IR tests typically employ verbose, well-specified topic texts for automatic query construction whereas our sample data reflects real life: human professionals performing a typical search task as best they can. There are many possible explanations for the initial queries not delivering reasonable results. We observed the following problems in our study:

- Errors in using the syntax of the underlying search engine. Capitalizing a search key that is not a proper name when the engine is sensitive to capitalization.
- Using search keys that do not cover the topic appropriately or from the right angle.
- Applying an attribute/metadata based filter (e.g., location criterion) that was too restrictive when combined with content keywords.
- Incorrect controlled metadata value (e.g., wrong value for information type).

Other common reasons include typos, overly specific query formulations (which may return nothing), and far too broad formulations (which are too sparse for relevant documents) – all of which would require query reformulations. sDCG makes such performance variations visible. It shows the magnitude of the gain change due to reformulation and suggests which reformulation to focus on in the analysis. Thus sDCG helps us to analyze which initial formulations and reformulations work best. The query sequence discount penalizes systems that require more queries than others.

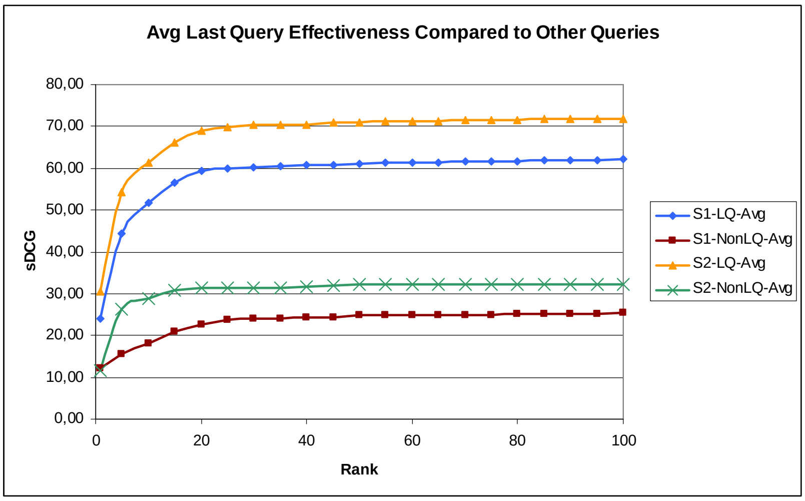

Figure 1. Average session-based discounted query gain for last vs. non-last queries in Systems 1 and 2 across sessions (*b*=2; *bq*=4).

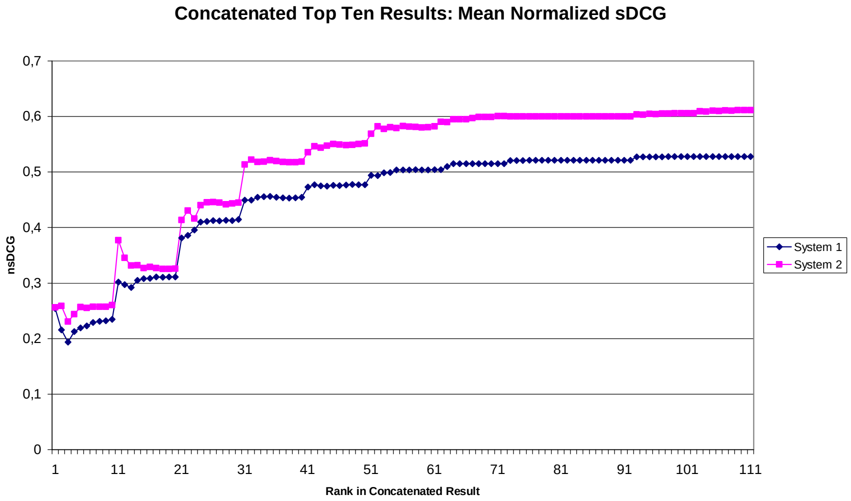

Figure 2. Session performance by normalized sDCG based on concatenated Top-10 results averaged across all sessions in Systems 1 and 2 (*b*=2; *bq*=4).

3.3 Effectiveness by Sessions

In order to analyze the effectiveness of sessions in the two systems, we now represent each session by concatenating the discounted gains from each query result. This allows the analysis of individual sessions as well as of average sessions across searchers using a specific system. Having relevance assessments by an external judge allows us to analyze the gain vectors up to the last ranks of query results. Thus we may examine performance differences, assuming that the searchers would have scanned Top-X documents, regardless of whether they did in the actual experiment.

In order to *normalize* concatenated sDCG vectors, one needs the corresponding ideal vector and an approach to handle duplicate results. When the searchers are assumed to scan Top-X documents of each query in a sequence, we propose the ideal sDCG vector to be constructed as follows: First one constructs the ideal vector (see appendix) for a single query. Second, the Top-X components of this ideal result are concatenated *n* times to represent a query sequence of *n* queries of a session. Each repeated result is discounted using formula (1). This is justified because, in an ideal situation, the searcher issues only one optimal query, which retrieves sufficiently many relevant documents in the optimal order. Each new query in a real session is another attempt at the ideal result. Some documents are, however, returned multiple times by different queries in a session. One must therefore decide whether to cumulate value only the first time a document is returned or every time it is returned. In order to compare systems, and because of using the reference judgments, we chose the latter option because, in our study, some searchers overlooked relevant documents in early queries but recognized them in later ones and each appearance of such a document is a chance provided by the system for the user to recognize it.

Figure 2 reports normalized average performance analyses for concatenated Top-10 query results. In Figure 2, each lot of 10 ranks along the X-axis represents the discounted and normalized Top-10 sDCG of one query, from Q_1 to Q_{11}. The gains are summed progressively so that the gain for Q_n represents the total gain (in the Top-10 ranks of each query) from the beginning of the session. If a searcher stops at Q_n then the gain for that session is held constant up to Q_{11}, i.e., no more gain is amassed. We see that across all the sessions, System 2 has better average performance.

Figure 3 shows a clear trend in the data for one subset, Scenario D, with session length up to 6 queries. There are two pairs of graphs, the upper ones representing concatenated sDCG results averaged across all sessions for Scenario D and the lower ones representing the same for the subset of sessions issuing at least 4 queries. We did not normalize the data because all the queries are for the same scenario and therefore share the same ideal vector. It is clear that multi-query sessions are initially very ineffective. Graphs of the kind reported in Figure 3 may be created for any number of queries in a session and any query results length of interest. Additionally, one may experiment with the discounting parameters and observe their effects. Such graphs also show the performance up to *n* queries for any number of queries less than the final query.

Figure 4 displays some raw data, individual session performance by concatenated Top-10 results up to six queries in System 1 and across all 15 sessions of one search scenario. We have padded short sessions to the same length. When each graph turns horizontal, the searcher most often quit searching (and only sometimes found nothing more). We have not marked the actual quitting points on the graphs. This figure clearly demonstrates the great variability among sessions and that the last query was far more effective than earlier ones. Such a display is a very useful tool early in the evaluation at the individual session level.

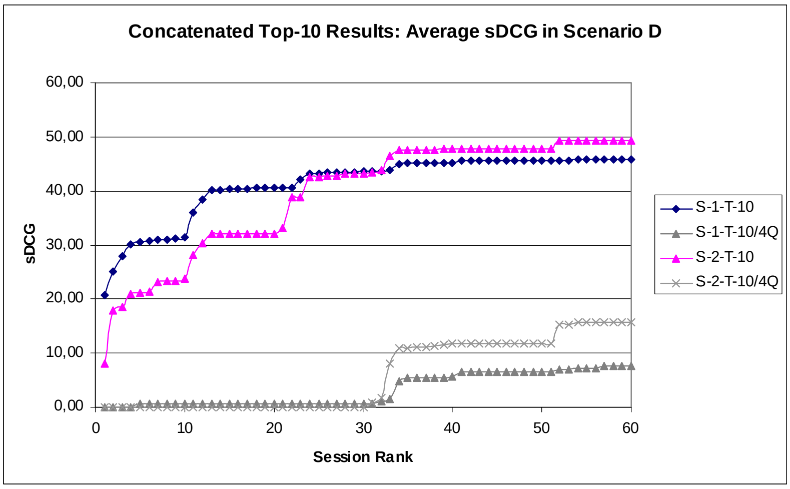

Figure 3. Session performance by concatenated Top-10 results averaged across all sessions of Scenario D in Systems 1 (S-1-T-10) and 2 (S-2-T-10) and across the sessions that constructed at least four queries (S-1-T-10/4Q and S-2-T-10/4Q) (*b*=2; *bq*=4).

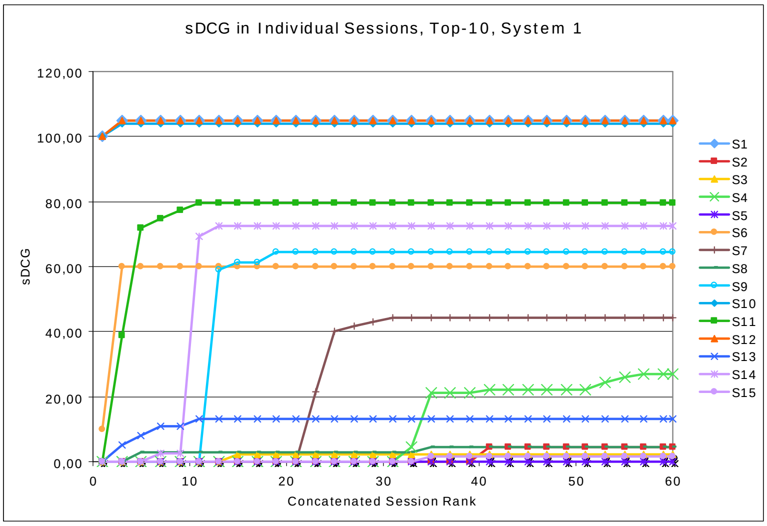

Figure 4. Individual session performance by concatenated Top-10 results in System 1 across all 15 sessions, Scenario D (*b*=2; *bq*=4).

4 Discussion

Standard single query metrics, such as MAP, and interactive metrics, such as instance recall, are insufficient when IR systems and interaction are studied from a more realistic session-based perspective. We extended the Discounted Cumulated Gain metric (DCG) [5] to a session-based sDCG metric by applying a query discount over a sequence of queries. The salient features of the new metric are:

- It uses graded relevance assessments and can reward highly relevant documents.
- It supports experimentation with relevance weighting from liberal, flat weights to sharp weighting of highly relevant documents.
- It supports experimentation using document and query discounts that can adjust the gain of documents retrieved late in a query or session. This supports modeling of various searcher/task scenarios regarding searcher impatience vs. persistence and regarding task/context dependent constraints and time limitations.
- For individual sessions, it supports identifying unsuccessful and successful reformulations, aiding searcher behavior analysis and system development.
- By selecting the assumed result scanning length (or recording observed searcher behavior) before stopping, it supports representing entire sessions by gain vectors that can be compared to each other or to ideal performance.
- sDCG can be normalized for comparisons across search scenarios. However, direct or averaged comparisons without normalization may be very informative as well and normalization is unnecessary when all queries are for the same scenario.

The sDCG metric and its application may appear complex compared to standard IR testing. This is unavoidable. Involving users and multiple query sessions introduces new variables into the test setting. The document and query discount parameters are important for realism because initial queries may not be successful and searching time may be limited. The new parameters allow assessing performance over a range of searcher/task scenarios. The complexity is a strength that allows bringing realism to evaluation and does not assume that all searchers or contexts are alike. Setting the parameters in each evaluation case depend on the evaluation purposes and should be based on relevant findings on searcher behavior or simulation scenarios exploring a range of searcher behaviors.

Initial queries may fail due to the widely known vocabulary problem. Our data shows that domain professionals have different interpretations and, consequently, construct differently behaving queries – even when facing the same scenario. This does not happen when the topics of test collections are used directly as queries. Using the sDCG we can evaluate systems and interfaces that may or may not be helpful in supporting good queries. For example, a plain engine with a keyword search box may be excellent in ranking documents for any given query. However, a domain specific interface may support the searcher in (re)formulating a much better query. New tools, such as sDCG, are essential for evaluating such interfaces.

sDCG can accept relevance scores derived from users, from independent relevance assessors, or from pre-existing test collections. For concatenating top-N results, there are three important issues to consider:

- Scanning length: In this study, we assumed most users consider the top ten results. Ideally we want to know how far each user scanned each result list, using eye-tracking data or having searchers mark the last document considered. Click data is a limited surrogate because it doesn't indicate which documents were rejected based on title or summary.
- Short results lists: Some queries return few or no hits. Concatenating short results lists "as is" ignores the time spent looking at a short or empty list and reformulating the query. In this study we padded concatenated lists to a fixed

length (10) but a hybrid approach might be more appropriate, assessing a minimum penalty per query.

- Duplicates: Documents are often returned multiple times by different queries in a session. In truly user-based evaluation, the searcher is free to score duplicates as relevant or non-relevant. For example, in our study some searchers overlooked relevant documents in early queries but recognized them in later ones. If judgments are supplied by external judges or a test collection, this option is not available. One must decide whether to cumulate value only the first time a document is returned or every time it is returned. We chose the latter option.

In the present paper we do not perform statistical testing as we are only using the data to illustrate the proposed metric. However, statistical testing may be applied on the sDCG findings. Appropriate statistical tests depend on the study design, and the sDCG metric may be used in many quite different designs. The final concatenated sDCG gain of a session, the sDCG gain of the best or last query, or the average gain within a session, are possible choices for evaluation. The *final concatenated gain* is insensitive to early stopping in some sessions, as the gain does not change after stopping. The *average gain,* i.e., the average position value of an sDCG vector (see [5] for the formula), represents the overall performance of all the queries in a session as a single number. When averaging gain across multiple search scenarios and using independent relevance judgments, normalizing (see Appendix) adjusts for the differing number of relevant documents across scenarios and represents performance in the range [0, 1] in relation to the ideal. Thence statistical testing is not affected by outliers, scenarios returning many highly relevant documents.

Acknowledgements

This work was supported in part by the Academy of Finland, under Project numbers 204978 and 1209960 and by the U.S. National Science Foundation, grant numbers 0514238, 0511050 and 0534762.

References

1. Bates, M.: The design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review 13, 5, 407--424 (1989)
2. Beaulieu, M., Robertson, S., Rasmussen, E.: Evaluating Iinteractive Systems in TREC. Journal of the American Society for Information Science 47, 1, 85--94 (1996)
3. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Dortrecht (2005)
4. Järvelin, K., Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents. In: 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41--48. ACM Press, New York, (2000)
5. Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems, 20, 4, 422--446 (2002)
6. Kekäläinen, J.: Binary and Graded Relevance in IR Evaluations – Comparison of the Effects on Ranking of IR Systems. Inform. Processing & Management 41, 5, 1019--1033 (2005)
7. Over, P.: TREC-7 interactive track report. In: NIST Special Publication 500-242: The Seventh Text Retrieval Conference. NIST, Gaithersburg (1999)

8. Price, S. L., Lykke Nielsen, M., Delcambre, L. M. L., Vedsted, P.: Semantic Components Enhance Retrieval of Domain-Specific Documents. In: 16th ACM conference on Conference on information and knowledge management, pp. 429--438. ACM Press, New York (2007)
9. Voorhees, E.: Evaluation by highly relevant documents. In: 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 74--82. ACM Press, New York (2001)

Appendix

The cumulated gain at ranked position i is computed by summing from position 1 to i, $i \leq 100$. By denoting the value in position i in the gain vector G by G[i], the cumulated gain vector CG is defined as the vector CG where:

$$\mathrm{CG}[i] = \sum\nolimits_{j=1}^{i} \mathrm{G}[j] \qquad (1)$$

For example, assuming G' = <3, 2, 3, 0, 0, 1, 2, 2, 3, 0, … > we obtain CG' = <3, 5, 8, 8, 8, 9, 11, 13, 16, 16, …>. The DCG metric also allows for weighting the relevance scores. For example, one can choose to replace the scores 0-3 by the weights of 0-1-10-100 to reward retrieval of highly relevant documents.

We define the vector DCG as follows:

$$\mathrm{DCG}[i] = \sum\nolimits_{j=1}^{i} \mathrm{G}[j]/(1 + \log_b i) \qquad (2)$$

For example, let b = 4. From G' given above we obtain DCG' = <3, 4, 5.67, 5.67, 5.67, 6.11, 6.94, 8.14, 9.30, 9.30, …>. Note that the formulation is slightly different from the original [9] and more elegant.

The construction of average vectors requires vector sums and multiplication by constants. For this, let V = $<v_1, v_2, \ldots, v_k>$ and W = $< w_1, w_2, \ldots, w_k>$ be two vectors.

$$\begin{gathered} V + W = <v_1 + w_1, v_2 + w_2, \ldots, v_k + w_k> \\ \Sigma_{V \in \vartheta}\ V = V_1 + V_2 + \ldots + V_n \text{ when } \vartheta = \{V_1, V_2, \ldots, V_n\} \\ r{*}V = <r{*}v_1,\ r{*}v_2, \ldots, r{*}v_k> \text{ when } r \text{ is constant} \end{gathered} \qquad (3)$$

The average vector of vectors $\vartheta = \{V_1, V_2, \ldots, V_n\}$, is:

$$\mathit{avg\text{-}vect}(\vartheta) = |\vartheta|^{-1} * \Sigma_{V \in \vartheta}\, V \qquad (4)$$

Given an (average) (D)CG vector V = $<v_1, v_2, \ldots, v_k>$ for an IR technique, and the (average) ideal DCG vector I = $<i_1, i_2, \ldots, i_k>$, the normalized performance vector nDCG is obtained by the function [9]:

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

The ideal vector has the relevance scored of the recall base in descending order. The nDCG vector components have values in the range [0, 1], representing the share of the ideal discounted gain achieved by the DCG vector V.