Correlation Between System and User Metrics in a Session

Jiepu Jiang
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst
jpjiang@cs.umass.edu

James Allan
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst
allan@cs.umass.edu

ABSTRACT

We investigate the correlations between system-oriented evaluation metrics and a few user experience metrics for a search session. The system-oriented metrics include session-based DCG (sDCG), normalized sDCG (nsDCG), estimated session nDCG (esNDCG), and a few variants of these metrics. We also look into statistics (e.g., the mean, maximum, and minimum values) of individual queries' nDCG scores, as well as the first and the last query's nDCG in a session. These system-oriented metrics are compared with users' self-rated search performance and task difficulty for a session. Experimental results show that nsDCG and esNDCG have reasonable but weak correlations with the user metrics, while the worst and the last query's nDCG in a session have comparably strong correlations. This suggests future work may better measure users' search experience in a session by modeling each query in the session differently.

Keywords

Search session; evaluation metric; user experience

1. INTRODUCTION

A challenge in the information retrieval community is to develop robust and reusable automatic evaluation approaches for interactive search, usually spanning multiple queries (a *search session*). One of the most critical issues is to find and design metrics for the quality of a search session. Such metrics can provide guidance to the design and optimization of search techniques for a session.

Existing approaches include two types. The first type directly predicts user experience based on behavioral signals. Previous studies predicted search success [1, 5], frustration [4], satisfaction [6, 9], and task difficulty [2, 15]. This approach requires user interaction as inputs and can only be performed in an online manner. The second type extends the Cranfield-style evaluation to a search session. It relies on relevance judgments and session-level evaluation metrics [8, 12] to assess search quality. The second approach is reusable, but it is unclear how well existing metrics correlate with user perceptions on search quality.

In this paper, we examine the issue of the second approach by correlating system-oriented evaluation metrics for a session with user-rated search performance and task difficulty.

2. METRICS

2.1 sDCG

Järvelin et al. [8] proposed the session-based discounted cumulated gain (sDCG) metric. sDCG sums up discounted cumulated gain (DCG) [7] for each query, but penalizes the contribution of later queries in a session. Järvelin et al. [8] believe that results retrieved by later queries in a session are less valuable, because query reformulation costs effort.

Here we examine the version of sDCG used by Kanoulas et al. [12] in order to be consistent with the normalized sDCG metric (discussed in Section 2.2). It is calculated as in Equation 1. For a session of n queries, sDCG sums up the DCG for each query, and applies a discount factor $1/\log_{bq}(i + bq - 1)$ to later queries in a search session. The DCG of the ith query q_i is calculated as in Equation 2, where $rel(q_i, r)$ stands for the relevance grade for the rth result on q_i's search result page (SERP). Following Kanoulas et al.'s work [12], we set $b = 2$ and $bq = 4$. When $b = 2$, the DCG in Equation 2 is identical to the one usually adopted for evaluating a single query's performance.

$$\mathrm{sDCG}(q_1, q_2, ..., q_n) = \sum_{i=1}^{n} \frac{\mathrm{DCG}(q_i)}{\log_{bq}(i + bq - 1)} \tag{1}$$

$$\mathrm{DCG}(q_i) = \sum_{r=1}^{k} \frac{2^{rel(q_i,r)}}{\log_b(r + b - 1)} \tag{2}$$

2.2 nsDCG

Kanoulas et al. [12] proposed a normalized version of the sDCG metric—normalized session DCG (nsDCG). A version of this metric tailored for two queries was used in the TREC 2010 session track [11]. nsDCG assumes an ideal ranked list for each query (q_{ideal}) where the judged relevant results are sorted by their relevance grades in a descending order. A session achieves the ideal performance if each query retrieves results with relevance grades equivalent to those in the ideal ranked list. The ideal session's sDCG is used for normalization, as in Equation 3.

$$\mathrm{nsDCG} = \frac{\mathrm{sDCG}(q_1, q_2, ..., q_n)}{\mathrm{sDCG}(q_{\text{ideal}}, q_{\text{ideal}}, ..., q_{\text{ideal}})} \tag{3}$$

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA
© 2016 ACM. ISBN 978-1-4503-3751-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2854946.2855005

2.3 esNDCG

Kanoulas et al. [12] also proposed the estimated session family of metrics. This family of metrics model different possible scan paths of users in a session. For example, a user may first examine two results on the first query's SERP, and then reformulate to the second query and examine four results. In this case, the scan path consists of the six examined results from the two queries' SERPs. The estimated session metrics evaluate each scan path by treating it as a virtual ranked list of results, and applying conventional IR evaluation metrics (such as nDCG and average precision) to assess the quality of the scan path. The metrics evaluate a search session by summing up each possible scan path's quality scores, weighted by the probability of the scan path.

Here we use nDCG to evaluate the quality of a scan path (to be consistent with other metrics examined in this paper). The metric is thus called esNDCG. Equation 4 computes esNDCG, where ω is a scan path, and $P(\omega)$ is ω's probability. $P(\omega)$ depends on two parameters: P_{ref}, the probability that users will reformulate to the next query after viewing the current SERP, instead of stopping and exiting the session, and P_{down}, the chances that users, after examining a result, will continue to examine the next one on the SERP.

$$\mathrm{esNDCG} = \sum_{\omega} P(\omega)\mathrm{nDCG}(\omega) \tag{4}$$

We also examine a variant of esNDCG that uses normalized cumulated gain (nCG) to evaluate the quality of a scan path. We call this metric esNCG. nCG is similar to nDCG, except that it does not apply the position-based discount to results at different ranks. The motivation of examining esNCG is that the estimated session family metrics already penalize lower ranked results and later queries in a session by modeling scan path (controlled by P_{ref} and P_{down}). Lower ranked results and those from later queries are less likely to be involved in a scan path. In such cases, it seems redundant to further penalize lower ranked results in each scan path.

2.4 sDCG/q

sDCG/q is an alternative way of normalizing sDCG. A session's sDCG is normalized by simply the number of queries in the session, instead of an ideal session's sDCG. Equation 5 computes sDCG/q, where n is the number of queries.

This metric comes from Jiang et al.'s work for predicting user satisfaction in a session [9]. They reported in a dataset that a similar metric highly correlates with user satisfaction in a session. Their metric does not discount later queries in a session (and is thus referred to as sCG/#queries in their article [9]). Another difference is that they computed sCG as simply the sum of all queries' ratings by external annotators rather than based on relevance judgments. The purpose of their work is to verify that user satisfaction can be modeled as the ratio of search outcome to effort. They measured search outcome by sCG, and effort by the number of queries in a session (n). The latter is motivated by Azzopardi's economic model of search interaction [3], where the cost of a search session is proportional to the number of queries in that session. Here our metric is different in that we calculate sDCG and sDCG/q based on relevance judgments, but they summed up assessors' query ratings.

$$\mathrm{sDCG/q} = \frac{\mathrm{sDCG}(q_1, q_2, ..., q_n)}{n} \tag{5}$$

2.5 Alternatives Without Query Discount

sDCG, nsDCG, sDCG/q, and esNDCG all penalize results retrieved by later queries in a session. However, Jiang et al. [9] showed that a variant of sDCG/q without discounting later queries better correlates with user satisfaction. Therefore, we also examine alternatives of these metrics that do not penalize later queries in a session.

We compute a variant of sDCG that does not discount later queries as in Equation 6. For nsDCG and sDCG/q, we replace sDCG by $\mathrm{sDCG}_{\mathrm{no\_query\_discount}}$ to remove the query discount component. It seems unclear how to exclude the query discount component in esNDCG, because both P_{ref} and P_{down} can affect the discounting. Thus, we do not consider its variant without query discount in this study.

$$\mathrm{sDCG}_{\mathrm{no\_query\_discount}} = \sum_{i=1}^{n} \mathrm{DCG}(q_i) \tag{6}$$

2.6 Individual Queries' nDCG

A seemingly reasonable idea for evaluating a search session is to consider statistics of individual queries' quality in the session. For example, a session may be satisfactory if each query retrieved good results. However, we know few use of these statistics as evaluation metrics for a session. We evaluate the quality of individual queries using nDCG [7], and then use the sum, mean, maximum, and minimum of the queries' nDCG scores in a session as metrics for that session's quality. In addition, we also use the first and the last queries' nDCG scores as indicators for the whole session's quality. This is suggested by Huffman et al.'s work [6] that showed search satisfaction in a session can be predicted using the first query's quality, while we examine both the first and the last query.

2.7 User Metrics

These system-oriented metrics are compared with two user-oriented metrics measured using the following two questions after users finished a search session. Responses to the first and the second questions are referred to as user-rated performance and task difficulty in this study.

- **Performance**: *how well do you think you performed in this task?* Options are: *very well* (5), *fairly well* (4), *average* (3), *rather badly* (2), and *very badly* (1).
- **Task Difficulty**: *how difficult do you think the task is?* Options are: *very difficulty* (5), *difficulty* (4), *average* (3), *easy* (2), and *very easy* (1).

3. DATASET

We use data from an existing user study[1] [10] to examine the system-oriented metrics for a session. We adopt this dataset because it collected both relevance judgments and users' ratings on their search experience when using an interactive search system. We restrict our scope to user-rated performance and task difficulty because the user study only collected these two user experience measures.

The original purpose of that user study was to compare search activity patterns in four types of tasks that vary in search goal (clear or amorphous) and product (factual or

[1] The dataset and source code can be accessed at https://github.com/jiepujiang/ir_metrics.

Table 1: Correlations between system-oriented metrics and user-rated performance and task difficulty.

Block	Metrics	Performance				Task Difficulty			
		Pearson		Spearman		Pearson		Spearman	
A	Performance	-		-		−0.787	***	−0.788	***
	Difficulty	−0.787	***	−0.788	***	-		-	
	Number of queries	−0.256	*	−0.241	*	0.305	**	0.301	**
B	sDCG	0.009		−0.056		0.065		0.063	
	nsDCG	0.350	**	0.326	**	−0.324	**	−0.300	**
	sDCG/q	**0.401**	***	**0.349**	**	**−0.388**	***	**−0.336**	**
	esNDCG ($P_{\text{ref}} = 0.9$, $P_{\text{down}} = 0.7$)	0.325	**	0.285	*	−0.246	*	−0.224	*
	esNCG ($P_{\text{ref}} = 0.8$, $P_{\text{down}} = 0.7$)	0.357	**	0.335	**	−0.261	*	−0.253	*
C	sDCG (no query discount)	−0.020		−0.104		0.092		0.118	
	nsDCG (no query discount)	0.353	**	0.323	**	−0.332	**	−0.305	**
	sDCG/q (no query discount)	**0.399**	***	**0.330**	**	**−0.374**	***	**−0.315**	**
D	sum nDCG	−0.018		−0.115		0.094		0.136	
	mean nDCG	0.352	**	0.320	**	−0.332	**	−0.302	**
	max nDCG (best query)	0.269	*	0.204		−0.191		−0.177	
	min nDCG (worst query)	0.348	**	**0.358**	**	−0.364	***	−0.379	***
	first query's nDCG	0.259	*	0.227	*	−0.177		−0.156	
	last query's nDCG	**0.371**	***	0.354	**	**−0.436**	***	**−0.419**	***

*, **, and *** indicate the correlation is significant at 0.05, 0.01, 0.001 levels, respectively.
Bold font indicates the strongest correlation of its column in each block.

informational) [14]. These tasks were developed by and used in the TREC 2012 session track [13]. The study recruited 20 subjects. Each worked on four tasks for about 10 minutes using an experimental search system. The system is similar to existing search engines, except that it shows only 9 results per page (to facilitate analysis of eye-movement data). The study collected 80 sessions (4.9 queries per session).

Relevance of results were judged at three levels: *Highly Relevant* (2), *Relevant* (1), or *Non-relevant* (0). Users rated 22 sessions' performance as *very well*, 27 as *fairly well*, 22 as *average*, 7 as *rather badly*, and 2 as *very badly.* They rated 3 sessions' difficulty as *very difficult*, 14 as *difficult*, 25 as *average*, 14 as *easy*, and 24 as *very easy.*

4. RESULTS

We evaluate the collected 80 sessions using the system-oriented metrics, and correlate the metrics' values with user-rated performance and task difficulty. esNDCG and esNCG include two parameters. We set their values by a brute force scan to optimize Pearson's r with user-rated performance. Table 1 reports Pearson's r and Spearman's ρ between the system-oriented metrics and the user metrics. Block A shows that performance and task difficulty have a strong negative correlation. This indicates that search performance and task difficulty are closely related but different. The number of queries, as an indicator of search cost [3, 9], has a positive correlation with task difficulty and a negative one with user-rated performance.

4.1 Existing Session Evaluation Metrics

Block B examines and compares the correlations of sDCG, nsDCG, esNDCG, esNCG, and sDCG/q with user metrics.

sDCG does not have any significant correlations with user-rated performance or task difficulty. We suspect it is because sDCG only measures how much information are retrieved in a session, but ignores the cost—as long as searchers issue more queries, they can find more information. Since sDCG sums up DCG for each query, it naturally correlates with the number of queries in a session ($r = 0.67$). But the latter has a negative correlation with user-rated performance.

The two normalized versions—nsDCG and sDCG/q—both significantly correlate with user-rated performance and task difficulty. It should be noted that the normalization factor of nsDCG (the ideal session's sDCG) also correlates with the number of queries, because it sums up the ideal DCG for each query. Thus, after normalization, both nsDCG and sDCG/q set off the cost factor in sDCG.

Both esNDCG and esNCG show significant correlations with the two user metrics, but esNCG consistently has relatively stronger correlations. This suggests that for the estimated session family metrics, discounting lower ranked results in each scan path may be harmful.

Except sDCG, other metrics in Block B all show significant positive correlations with user-rated performance. This verifies that existing metrics for a search session's performance are reasonable to some extent. But the correlations remain weak (Pearson's $r \leq 0.4$ and Spearman's $\rho \leq 0.35$), indicating the limited status of existing system-oriented metrics in measuring potential search experience of users.

sDCG/q, a variant of nsDCG that normalizes sDCG simply by the number of queries, has the strongest correlation with both user-rated performance and task difficulty in the collected data. We do not want to over-generalize this finding due to the limited size of the collected data. We require further studies to fully validate this metric.

4.2 Discounting Later Queries in a Session

Block C examines variants of sDCG, nsDCG, and sDCG/q that do not discount results from later queries in a session. Compared with their corresponding metrics in Block B, these variants have very similar correlations with both user-rated performance and task difficulty. For nsDCG and sDCG/q, the differences in r and ρ between Block B and Block C do not exceed 0.02. This indicates that the query discounting component in sDCG, nsDCG, and sDCG/q may not be necessary. At least our collected data do not show clear benefits of discounting results from later queries.

4.3 Individual Queries' Performance

Block D examines the connections between search experience in a session and individual queries' quality.

The sum of all queries' nDCG scores (sum nDCG) does not show any significant correlations with user-rated performance or task difficulty. The reason is similar to that for sDCG. The metric also highly correlates with the number of queries in a session ($r = 0.72$). Similar to sDCG/q, the mean value of individual queries' nDCG in a session (mean nDCG) shows significant correlations with user metrics. It should be noted that in our collected data, "sum nDCG" and sDCG have an almost perfect correlation ($r = 0.98$). The correlations of "mean nDCG" with nsDCG ($r = 0.99$) and sDCG/q ($r = 0.92$) are almost perfect as well. This means that many existing system-oriented session evaluation metrics, such as sDCG and nsDCG, are not much different from the sum or average performance of individual queries.

The worst query's nDCG in a session (min nDCG) has the strongest Spearman's correlation with user-rated performance among all the system-oriented metrics we examined. The Pearson's correlation ($r = 0.348$) is also comparable to those for nsDCG and esNDCG. In contrast, the best query's nDCG in a session (max nDCG) does not show clear correlations with the two user metrics. This indicates that a few underperforming queries in a session may substantially affect user experience, while it is common to find well-performing queries in any session.

The last query's nDCG has significant correlations with both user metrics. In fact, it has the strongest correlation with task difficulty among all system-oriented metrics we examined. In contrast, the first query's nDCG does not correlate much with user metrics. This suggests that failing to formulate effective queries in later stages of a session may be an indicator of task difficulty.

To conclude, results in Block D suggest that further studies may rely on individual queries to evaluate a search session. In addition, some queries such as the worst and the last query may have stronger influence on users' search experience in a session compared with other queries. This suggests that future work may better measure users' search experience in a session by modeling each query differently.

5. CONCLUSION

We examined the correlations between system-oriented evaluation metrics and user-rated performance and task difficulty in a search session. We found that a few existing metrics such as nsDCG and esNDCG have significant but weak correlations with user metrics. This verifies that these metrics are reasonable evaluation surrogates to some extent. However, results also indicate limited value of existing metrics. For example, we found that metrics such as sDCG and nsDCG are not much different from the sum or mean values of individual queries' nDCG. The worst query and the last query's nDCG values generally show stronger correlations with user metrics than existing system-oriented metrics such as sDCG, nsDCG, and esNDCG in our dataset.

This work, however, is limited by the small number of sessions we had evaluated (80). Due to the limited sample size, we also did not examine task effects on these metrics. In addition, we also did not examine a few other metrics in this work, such as the time-biased gain [17] and U-measure [16]. We leave these issues for future work.

6. ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *SIGIR '11*, pages 345–354, 2011.

[2] J. Arguello. Predicting search task difficulty. In *ECIR '14*, pages 88–99, 2014.

[3] L. Azzopardi. Modelling interaction with economic models of search. In *SIGIR '14*, pages 3–12, 2014.

[4] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR '10*, pages 34–41, 2010.

[5] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. In *WSDM '10*, pages 221–230, 2010.

[6] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *SIGIR '07*, pages 567–574, 2007.

[7] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[8] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *ECIR '08*, pages 4–15, 2008.

[9] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM '15*, pages 57–66, 2015.

[10] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *SIGIR '14*, pages 607–616, 2014.

[11] E. Kanoulas, B. Carterette, P. Clough, and M. Sanderson. Overview of the TREC 2010 session track. In *TREC 2010*.

[12] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *SIGIR '11*, pages 1053–1062, 2011.

[13] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. Overview of the TREC 2012 session track. In *TREC 2012*.

[14] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.

[15] C. Liu, J. Liu, and N. J. Belkin. Predicting search task difficulty at different search stages. In *CIKM '14*, pages 569–578, 2014.

[16] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *SIGIR '13*, pages 473–482, 2013.

[17] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *SIGIR '12*, pages 95–104, 2012.