Towards Designing Better Session Search Evaluation Metrics

Mengyang Liu, Yiqun Liu*, Jiaxin Mao, Cheng Luo, Shaoping Ma
Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

User satisfaction has been paid much attention to in recent Web search evaluation studies and regarded as the ground truth for designing better evaluation metrics. However, most existing studies are focused on the relationship between satisfaction and evaluation metrics at query-level. However, while search request becomes more and more complex, there are many scenarios in which multiple queries and multi-round search interactions are needed (e.g. exploratory search). In those cases, the relationship between session-level search satisfaction and session search evaluation metrics remain uninvestigated. In this paper, we analyze how users' perceptions of satisfaction accord with a series of session-level evaluation metrics. We conduct a laboratory study in which users are required to finish some complex search tasks and provide usefulness judgments of documents as well as session-level and query level satisfaction feedbacks. We test a number of popular session search evaluation metrics as well as different weighting functions. Experiment results show that query-level satisfaction is mainly decided by the clicked document that they think the most useful (maximum effect). While session-level satisfaction is highly correlated with the most recently issued queries (recency effect). We further propose a number of criteria for designing better session search evaluation metrics.

KEYWORDS

Session Search; User Satisfaction; Search Evaluation

ACM Reference Format:
Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Shaoping Ma. 2018. Towards Designing Better Session Search Evaluation Metrics. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210097

1 INTRODUCTION

Search evaluation is one of the central concerns in information retrieval (IR) studies. Most existing evaluation metrics (e.g., RBP [8], ERR [1], etc.) are based on result lists of a single query. However, session-level evaluation has also received more and more attention in recent years.

*Corresponding author

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210097

Most of the session-level evaluation metrics (e.g. Session-based DCG (sDCG) [4] and Expected Utility (EU) [10]) are designed in a manner similar to query-level evaluation metrics. They are built on the cascade hypothesis [2] which assumes the user views search results from top to bottom and the user's attention will gradually decay during the browsing process. Although this assumption makes sense for single query evaluation, whether it is valid at session-level has not been verified. The metrics that have a decaying weighting function emphasize the *primacy effect* that the first documents examined by users will be more influential for the overall session-level satisfaction. On the other hand, the *recency effect*, that users tend to begin recall with the end of search session, suggests the documents at the end of the session are more important. In other words, while the assumption behind a decaying weighting function is that the primacy effect is the major effect, an increasing weighting function favors the hypothesis that the recency effect is more important.

In order to investigate whether the primacy effect or the recency effect is more important for session-level evaluation, we conducted a laboratory user study to collect a dataset that contains search logs for 675 sessions and the corresponding satisfaction feedback from users. Using this dataset, we try to answer the following research questions:

- **RQ1** Does the primacy effect or recency effect has a stronger influence on user's session-level satisfaction?
- **RQ2** What is the contribution difference of query satisfaction at different positions to session satisfaction?

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes the user study and experimental settings. In Section 4, we present data analysis to address **RQ1** and **RQ2**. We found the flaws of existing session evaluation metrics and propose a number of criteria for designing better session search evaluation metrics. Finally, we give our conclusions and future work in Section 5.

2 RELATED WORK

Some recent studies concentrated on session-level evaluation methods. sDCG [4] is a extended version of the Discounted Cumulative Gain (DCG) [3], it assumes the documents at lower position and retrieved by later query are less likely to be read by users, and therefore, have weaker influence on session-level satisfaction. EU [10] take the novelty of results into account, the gain of a result will be discounted if the same content has been encountered in previous results.

There are many existing studies focusing on the estimation of user satisfaction. Wang et al. [9] proposed a model in which user's action-level satisfaction was considered as a latent factor that affects the session-level satisfaction. Jiang et al. [5] compared user's feedback in two experimental settings, which include an in situ

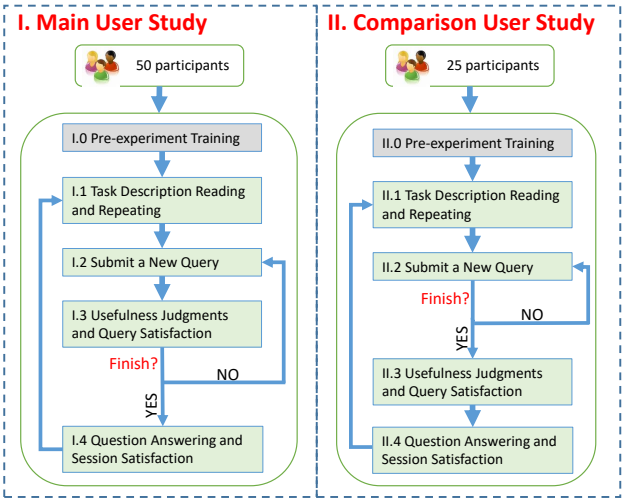


Figure 1: User study.

one and a context-independent one. We also use two similar experimental settings for comparison in our study, the experimental details will be introduced in Section 3. Mao et al. [7] found that users' usefulness feedback reflects users' satisfaction better than relevance. They compared a series of evaluation metrics based on user's click sequence, but they did not investigate whether these metrics were suitable for session evaluation.

In this study, we mainly focus on investigating how fine-grained perceptions contribute to session-level satisfaction so that we can know how to design a session evaluation metric.

3 DATA COLLECTION

To investigate the relationship between users' session satisfaction and query satisfaction, we conducted a laboratory user study which consists of two parts (see Figure 1) : I. Main User Study and II. Comparison User Study. We collect the three kinds of feedback from participants in these two parts: (1) Document-level usefulness feedback. (2) Query-level satisfaction feedback. (3) Session-level satisfaction feedback.

3.1 Main user study

In our main user study, we recruited 50 participants aged from 18 to 27. 24 participants were female, and other 26 participants were male. All the participants were familiar with basic usage of web search engines. Each participant needed to complete 9 tasks. These search tasks were designed based on the following criteria. Firstly, the task should be easily interpreted by all participants. Secondly, the task should not be a trivial one, since we mainly focus on search sessions with multiple interactions.

An experimental search engine system is developed for the user study. When users submit queries to this system, it crawls corresponding results from a major commercial search engine. In the crawled SERPs, all query suggestions, ads, and sponsor search results are removed to reduce the potential impacts on users' behavior. When performing tasks, participants can freely formulate queries.

We made sure that each participant understood the experimental process through a pre-experiment training task. After the training stage, each participant was asked to perform 9 tasks in a random order. As shown in Figure 1-(I), the experimental procedure contains:

(I-1) In the first stage, the participant should read and memorize the task description in an initial page, and she is asked to repeat the task description without viewing it to ensure she has remembered it.

(I-2) Next, the participant can submit a query and click on the results to collect information as they usually do with commercial search engines.

(I-3) After finishing the current query, she is asked to mark whether each document was useful for her at an evaluation page (0: not at all, 1: somewhat, 2: fairly, 3: very useful). She is also asked to give a 5-level graded satisfaction feedback on this query. If she wants to find more information, she can go back to step (I-2) and submit a new query. She can end the search whenever she thinks enough information has been found, or she can find no more useful information.

(I-4) Finally, she is required to give a search answer and an overall 5-level graded satisfaction feedback of the whole search session of the task.

3.2 Comparison user study

The usefulness judgement and query satisfaction feedback stage in User-Study-I may affect participants' search behavior and subsequent session satisfaction feedback, therefore, we designed another comparison user study. As shown in Figure 1-(II), only the second and third step was changed. The participants were asked to give usefulness feedback for all documents at once when they finish a task instead of after each query.

We recruited another 25 participants aged from 18 to 26 to take part in this experiment. 13 participants were female, and other 12 participants were male. This part of the data will only be used for comparative analysis in Section 4.4.

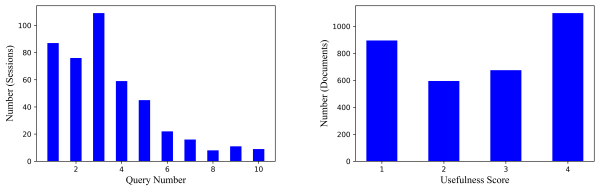
4 DATA ANALYSIS

In this section, we first examine the relationship between users' query-level satisfaction and the corresponding usefulness judgments of the clicked documents. Then we analyze whether the primacy effect or recency effect has a stronger influence on user's session-level satisfaction. Furthermore, we investigate how users' query-level satisfaction contribute to the corresponding session-level satisfaction.

4.1 Data distribution

As mentioned in Section 3, we collected 1,548 users' 5-level graded query satisfaction feedback and 3,276 4-level graded usefulness feedback. We show the distributions of the number of queries in a session, usefulness feedback, query-level and session-level satisfaction feedback in Figure 2.

From figure 2(a), we can see that sessions with three queries have the largest frequency, and sessions with no more than five queries accounts for 85.1%. An almost evenly distribution of usefulness judgments and query satisfaction can be seen in figure 2(b) and figure 2(c). As for session satisfaction shown in figure 2(d), only 6.6% sessions have a satisfaction score no more than 2. The results



(a) Query number distribution (b) Usefulness judgements distribution

(c) Query satisfaction distribution (d) Session satisfaction distribution

Figure 2: Distribution of usefulness and query satisfaction.

Table 1: Correlation of different metrics with query-level satisfaction (All correlations are significant at p < 0.001).

Metrics	All Queries (1532)	Queries with Clicks (1330)
cCG	0.552	0.508
cDCG	0.698	0.670
cERR	0.704	0.665
cMin	0.639	0.615
cMax	**0.831**	**0.838**

indicate that most users have found some satisfying information within several queries.

4.2 Query-level satisfaction

To investigate the relationship between users' query-level satisfaction and the usefulness judgments. We calculated five different metrics (*cCG*, *cDCG*, *cERR*, *cMin*, *cMax*) based on users' usefulness judgments of click sequence. As shown in Equation(1)-(4), $CS = (d_1, d_2, ..., d_{|CS|})$ is the click sequence in which each element d_r is a clicked document, u_r is the usefulness judgment of d_r and we use $(2^{u_r} - 1)$ to represent the gain of it.

$$cCG = \sum_{r=1}^{|CS|} gain(d_r) = \sum_{r=1}^{|CS|} 2^{u_r} - 1 \quad (1)$$

$$cDCG = \sum_{r=1}^{|CS|} \frac{gain(d_r)}{log_2(r+1)} = \sum_{r=1}^{|CS|} \frac{2^{u_r} - 1}{log_2(r+1)} \quad (2)$$

$$cERR = \sum_{r=1}^{|CS|} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r = \sum_{r=1}^{|CS|} \frac{1}{r} \prod_{i=1}^{r-1} (1 - \frac{2^{u_i} - 1}{2^3}) \frac{2^{u_i} - 1}{2^3} \quad (3)$$

$$cMin/cMax = (min/max)(u_1, u_2, ..., u_{|CS|}) \quad (4)$$

The Pearson's correlation coefficient between these metrics and query satisfaction for all queries and queries with at least one click are shown in Table 1. The results show that all metrics have significant correlations with query satisfaction. *cMax* has the strongest correlation ($r = 0.831$) with query satisfaction, which suggests that the maximum usefulness of clicked documents can best reflect user's query-level satisfaction.

Table 2: The r-th query's weight of different session weighting functions (N is the query number of a session).

Metrics	$w_r(0 < r \leq N/2)$	$w_r(N/2 < r \leq N)$
Decrease_weight	$1/r$	$1/r$
Increase_weight	r	r
Equal_weight	1	1
Middle_high	r	$N + 1 - r$
Middle_low	$1/r$	$1/(N + 1 - r)$

Table 3: Correlation of different metrics with session-level satisfaction. (The two columns represent metrics based on query satisfaction or cMax of query respectively).

Metrics	Satisfaction Based	cMax Based
Decrease_weight	0.644	0.463
Increase_weight	0.765	0.560
Equal_weight	0.724	0.532
Middle_high	0.696	0.513
Middle_low	0.732	0.532

4.3 Comparison of different weighting function

Since existing session evaluation metrics often include a weighting function, we investigate whether the primacy effect or recency effect has a stronger influence on user's session-level satisfaction.

As shown in Table 2, we used five types of weighting functions to weight user's query-level satisfaction. Each metric can be calculated with Equation(5) in which s_i represents the user's satisfaction feedback on i-th query. To let the metrics comparable across sessions with different length, we use a normalized query weight w_i^* in this Equation.

$$M = \sum_{i=1}^{N} w_i^* * s_i \quad (5)$$

The w_i^* can be calculated with Equation(6). Here, N is the query number, w_r is the r-th query's original weight which has been shown in Table 2.

$$w_r^* = \frac{w_r}{\sum_{r=1}^{N} w_r} \quad (6)$$

Table 3 shows the Pearson's correlation coefficient between these five metrics and session satisfaction. We can see that the metric with decreasing weights has the lowest correlation coefficient with session satisfaction while the metric with increasing weight has the strongest correlation. We also use the *cMax* of all queries instead of query satisfaction feedback from users to calculate the five metrics and show their correlation with session satisfaction in Table 3. Similarly, user's session-level satisfaction has a higher correlation with the metric that has an increasing weight. These results suggests that it is the recency effect but not the primacy effect that has a stronger influence on user's session-level satisfaction.

4.4 Fitting weight analysis

As shown in section 4.2, metrics with different weighting functions have different performance in estimating session satisfaction. Therefore, the contribution of query satisfaction to session satisfaction may vary at different positions.

Table 4: Fitting weight at different query positions (Main user study) .

#Query	Intercept	R1	R2	R3	R4	R5
1	−0.299	1.021				
2	−0.329	0.213	0.856			
3	−0.222	0.211	0.275	0.566		
4	−0.323	0.035	0.131	0.331	0.781	
5	−0.206	0.162	0.391	0.104	0.226	0.274

Table 5: Fitting weight at different query positions (Comparison user study) .

#Query	Intercept	R1	R2	R3	R4	R5
1	−0.396	1.023				
2	0.287	0.104	0.303			
3	−0.094	0.099	0.143	0.300		
4	−0.102	0.184	0.139	0.219	0.261	
5	−0.238	0.163	0.164	0.220	0.316	0.508

To analyze the weight difference between queries, we perform a linear regressions analysis. We standardize query satisfaction and session satisfaction with a z-score transformation respectively and fit different models for sessions with different length. Only the sessions with no more than five queries are considered because there are only a few sessions with more than five queries. As shown in Table 4, we can see that the weight of the last query is higher than those of other query positions in each group of sessions. This proves again that the recency effect has a stronger influence on user's session-level satisfaction than the primacy effect.

In our main user study, the users are asked to give their usefulness judgments and satisfaction feedback while finishing each query. So the satisfaction feedback of the last query is collected just before the session satisfaction. To investigate whether the experimental settings have impacts on the results, we conduct a parallel analysis on the data collected in the comparison user study (Section 3.2). The results of the comparison analysis are shown in Table 5. We find that the weight of the last query is still higher than others. The observed recency effect is not caused by the adjacency of satisfaction feedback of the last query and the whole session.

These results imply that a user is more concerned about the quality of recent queries. If a user experiences a series of unsatisfying queries and finally finds a good one, she may also feel satisfied with her search session.

4.5 Discussion

In this section, we analyze the user study data to address the two RQs. A comparative experiment is designed to ensure our findings are not due to the experimental settings.

From our results, we can see that traditional evaluation metrics may not be suitable for session satisfaction evaluation based on browsing sequence because they ignore the recency effect in users' perception of session-level satisfaction. Furthermore, the experiment results suggest that a session satisfaction evaluation metric should meet the following criteria: (1) The most useful document in a query is the most important; (2) The weighting function between queries should be normalized; (3) The primacy effect is not suitable for session evaluation; (4) The recency effect has a stronger influence on user's session satisfaction.

5 CONCLUSION

In this study, we conducted two laboratory studies in which users need to give their usefulness judgments and satisfaction feedback. We investigated how users' query-level satisfaction contribute to their session-level satisfaction. Based on users' click sequence, we calculated a series of query-level evaluation metrics and found that users' query-level satisfaction mainly decided by the clicked document that is most useful from users' perspective. We tried different weighting function to fit session satisfaction with query satisfaction and found that a decaying function is not appropriate for session evaluation. Furthermore, through a linear regression analysis, we found that users' perceptions of the last query have the greatest impact on their session satisfaction. Finally, we proposed some criteria for designing session evaluation metrics based on the experiment results.

As for future work, we would like to design click-sequence-based evaluation metrics for multi-query sessions. Beyond usefulness and satisfaction, more kinds of measures (e.g. search success [5, 6]) will be investigated in our future work.

6 ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and National Key Basic Research Program (2015CB358700).

REFERENCES

[1] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.

[2] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.

[3] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[4] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval*. Springer, 4–15.

[5] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 405–414.

[6] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship Between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1533–1542. https://doi.org/10.1145/3178876.3186065

[7] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 463–472.

[8] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.

[9] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan, and Ryen W White. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 123–132.

[10] Yiming Yang and Abhimanyu Lad. 2009. Modeling expected utility of multi-session information distillation. In *Conference on the Theory of Information Retrieval*. Springer, 164–175.