# Tuning Parameters of the Expected Reciprocal Rank

Yury Logachev, Lidia Grauer, Pavel Serdyukov
Yandex
Leo Tolstoy st. 16, Moscow, Russia
ylogachev@yandex-team.ru, lidia@yandex-team.ru, pavser@yandex-team.ru

## ABSTRACT

There are several popular IR metrics based on an underlying user model. Most of them are parameterized. Usually parameters of these metrics are chosen on the basis of general considerations and not validated by experiments with real users. Particularly, the parameters of the Expected Reciprocal Rank measure are the normalized parameters of the DCG metric, and the latter are chosen in an ad-hoc manner. We suggest two approaches for adjusting parameters of the ERR model by analyzing real users behaviour: one based on a controlled experiment and another relying on search log analysis. We show that our approaches generate parameters that are largely different from the commonly used parameters of the ERR model.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

Information retrieval measures, evaluation

## 1. INTRODUCTION

One of the most challenging problems in the field of Web Search is choosing appropriate metric for learning and evaluating retrieval algorithms. Chapelle et al. suggested the Expected Reciprocal Rank (ERR) metric [2], which received wide recognition in the community, mainly due to its plausible underlying user model. The ERR is based on a cascade model which has a set of parameters, each of which corresponds to the relevance grade of a document. Each parameter means the probability of getting completely satisfied after reaching a document with a certain relevance grade. It is assumed that once the user is satisfied with a document, she terminates the search and documents below this result are not examined regardless of their position. Chapelle suggested a method of setting these parameters using the gain parameters of the DCG metric: $R(g) = \frac{2^g-1}{2^{g_{max}}}$ where $g \in \{0, \ldots, g_{max}\}$ are the relevance grades. Thereby, commonly used parameters of the ERR metric for a 5-grade scheme with grades *Perfect*, *Excellent*, *Good*, *Fair*, *Bad* are respectively $\approx 0.94, 0.44, 0.19, 0.06, 0$. The same set of parameters (also for a 5-grades scheme) was used at TREC 2010/2011 [3] and de facto became a standard. We argue that these parameters should be adjusted more accurately and set by analyzing real users' behaviour.

We suggest two experiments for setting parameters of the ERR metric: a controlled experiment and a clickthrough experiment. Both experiments aim to directly estimate these probabilities based on the assumptions of the ERR model. Two parameters sets that we have adjusted are very different from the original parameters of the ERR metric. We hope that the ERR model with the parameters adjusted with our approaches simulates user's behavior more accurately.


*WWW 2012 Companion,* April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

## 2. PARAMETERS ESTIMATION

Both experiments that we describe are based on the same idea. For each relevance grade the corresponding parameter is the estimation of the probability of the user to get satisfied with a document with a given relevance grade if she has seen it. Note that in the context of the cascade model this probability depends only on the relevance grade of the document. Thus in our experiments we estimated this probability as the frequency of being satisfied after seeing the first ranked document with a given relevance grade. It was important due to the fact that in reality users might accumulate partial information from the previously scanned documents what may affect their decisions and we wanted to avoid such a bias in our calculations. For each query and a document that we analyze, we created relevance judgements using a 5-grade system (*Perfect*, *Excellent*, *Good*, *Fair*, *Bad*). Judges used descriptions for each grade, which are very similar to the ones used at TREC Web Track [3] (some documents were also labeled as *Junk*, as at TREC, but we do not analyze them here). The controlled experiment closely simulates the user model behind the ERR: editors look at the documents (not snippets) one by one. The clickthrough experiment is more realistic, because the actual query logs are analyzed and it is not assumed that users open and read every document linked from the search engine result page (SERP) that they make their judgement about.

**Controlled experiment.** In order to create the collection of search tasks, we randomly sampled 2000 unique queries from the logs of a commercial Web search engine and retrieved top-10 documents for these queries. We randomized these sets of documents and presented them to the group of 40 paid users. These users were asked to act as if these queries were their own and think out their own subtopics for each of "their" queries. For example, for query [madonna] there was subtopic "I want to listen to Madonna's last music track". After choosing the query subtopic, users

**Table 1: Results of the experiments**

| | Controlled | | Clicks | DCG-based |
|---|---|---|---|---|
| | R(g) | CI-95% | R(g) | R(g) |
| Bad | 0.06 | (0.05;0.066) | 0.18 | 0 |
| Fair | 0.21 | (0.194;0.236) | 0.23 | 0.06 |
| Good | 0.54 | (0.513;0.563) | 0.27 | 0.19 |
| Excellent | 0.69 | (0.653;0.717) | 0.38 | 0.44 |
| Perfect | 0.74 | (0.712;0.773) | 0.59 | 0.94 |

scanned the documents rankings presented to them by reading one document after another. After reading each document, they could make one of three decisions: "I have found the needed information", "I have not found the needed information and got tired", "I have not found the needed information, but I will continue". So, they could stop their search either by finding the information on the chosen subtopic or by admitting that they are tired and no more interested in the current search task. Each user had to complete 200 search tasks. Although the search tasks consisted of 10 documents, only the first document in the sequence and the corresponding user decision was accounted for our experiment. Once we had the distributions of user decisions for the first documents, we were able to calculate the frequency of cases when users found the needed information in the documents of each grade and stopped their search. The resulting probabilities of user satisfaction for different grades are presented in Table 1, together with their confidence intervals.

**Clickthrough experiment.** In order to simulate a more realistic search scenario, we decided to conduct experiments with real users and try to derive the probabilities of satisfaction for documents with different grades by analyzing the average user's search behaviour. Such approach has its pros and cons. From one point of view, some unrealistic assumptions are avoided in such a scenario. For example, we no more assume that the user reads every document before it decides that she is not satisfied with it. In reality, users always rely on snippets (though often mistakenly) and we assume that user models that underlie IR metrics should take this into account. From another point of view, we cannot know for sure if the user is satisfied with the document she interacted. However, we rely on popular ways to determine it with high confidence by analyzing the type and the time of the next user action.

In this experiment, we assume that the user is *satisfied* and finds the needed information in the first ranked document if she clicked on it, did not click any other document below on the same SERP and did not issue another query quickly after that first click. If the user clicked another document below (also by skipping the first ranked document) or issued another query quickly after that first click, it means *dissatisfaction* with the first ranked document. We considered next query as a sign of dissatisfaction if it was requested in less than 30 seconds after the click on the first ranked document (as the user spent less than 30 seconds reading it). This parameter can be tuned, but we followed the recent publications on click-based personalization, which successfully used the same threshold to determine the user-specific relevance of the document to the user who clicked it [1].

We used query logs of a popular search engine for three weeks period. Queries generated by search bots were filtered using a proprietary bot filtering algorithm. We considered only each first query in each session (30 minutes period was used to delimit sessions), if it had at least one document clicked in the organic search results (and hence the user always examined the first ranked document or its snippet) and did not have clicks on any other SERP elements (such as ads). Only first queries were used to avoid the bias of accumulating information on the same information need from the previous queries in the session. We sampled random 3740 unique queries and the corresponding search sessions and asked our judges to assess all result documents (with the same 5-grade system as in the controlled experiment) that were actually shown to the users. As a result we got 181,853,603 search sessions with first queries which had their first ranked document judged.

As long as we have serious imbalance in the number of users for each query, we wanted to avoid the bias towards popular queries and averaged results as follows. Because of this procedure, calculating confidence interval is non-trivial and we did not perform it. For each unique pair <query, $D$>, where $D$ is the top result document, we first estimated weight $\frac{n_s}{N}$, where $N$ - is the number of such pairs in the logs, $n_s$ - is the number of times when a user was satisfied with the document $D$ for that query. Finally, for each grade, these weights were averaged over documents with that grade. The results are presented in Table 1.

# 3. CONCLUSION

We described two methods of estimating parameters of the ERR model. The results of these methods are different due to the limitations of the clickthrough experiment, natural differences between a paid user and a usual user, the fact that paid users judged entire documents, not their query-specific snippets. However, as we see in Table 1, the DCG-based probabilities for different grades are not realistic according to both experiments. Especially, they do not seem so for the lowest (*Bad*) and the highest (*Perfect*) grades. As we see, users often get satisfied with even *Bad* documents and quite not very often get satisfied with *Perfect* ones. Most probably, it may happen simply due to the fact that judges are subjective and never fully agree on judgements for all documents. Potentially, some part of *Bad* or *Perfect* documents could be also *Fair* or *Good* and hence we cannot expect that all such documents will be always useless or almost always useful for each and every user.

We hope that the described approaches may give a start to development of sophisticated and accurate methods of setting parameters of the ERR model. In the future we are going to increase the reliability of our controlled experiment by taking additional factors into account: mistakes of paid users, agreement of judges over different grades, etc. Particularly, when analyzing the decisions of users on *Bad* documents, it is interesting to distinguish the cases when paid users and judges make mistakes or disagree. We wil also look for the ways to improve our guess about the user satisfaction in the clickthrough experiment, probably by utilizing document dwell time in a more sophisticated manner.

# 4. REFERENCES

[1] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *SIGIR*, 2011.

[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. CIKM '09, 2009.

[3] C. L. A. Clarke, N. Craswell, N. Craswell, and G. V. Cormack. Overview of the trec 2010 web track. TREC '10, 2010.