Incorporating Efficiency in Evaluation

Eugene Kharitonov[†‡], Craig Macdonald[‡], Pavel Serdyukov[†], Iadh Ounis[‡]
[†]Yandex, Moscow, Russia
[‡] School of Computing Science, University of Glasgow, UK
[†]{kharitonov, pavser}@yandex-team.ru
[‡]{craig.macdonald, iadh.ounis}@glasgow.ac.uk

ABSTRACT

The trade-off between retrieval efficiency and effectiveness naturally arises in various domains of information retrieval research. In this paper, we argue that the users' reaction to the system efficiency should be incorporated into the cascade model of the user behaviour and that an effectiveness-aware modification of a state-of-the-art evaluation metric can be derived. We propose to base the study on the anonymised browser/toolbar log data, and our preliminary experiments demonstrate that the users' reaction to the delays observed in the dataset is consistent with one reported previously.

1. INTRODUCTION

The problem of measuring the efficiency and effectiveness trade-off arises in a variety of retrieval settings. For instance, machine-learned rankers of commercial search engines can leverage thousands of regression trees, operating in space of hundreds of features [2]. Increasing the number of the regression trees is likely to improve the ranking effectiveness, but it inevitably reduces the system efficiency. Indeed, while improving effectiveness improves the users' experience with the search engine, the overall users' satisfaction might fall due to the system efficiency decrease. Completely different approaches to trade-off the system effectiveness for the system efficiency are considered in [6, 8, 9]. The central idea behind that work is to introduce additional tools to trade-off the retrieval effectiveness for the retrieval efficiency.

Since from the user's point of view, network and query processing delays are indistinguishable, in this work we use a generalised notion of efficiency that incorporates the time it takes the user's browser to send the HTTP request to the search engine, time needed for the search engine to process the query, and the time required for the browser to receive the result page over the network.

The question arises how to measure the overall user's satisfaction with the system performance, incorporating effectiveness and efficiency in a single metric. The commonly used approaches are not experimentally justified to be aligned with the user's satisfaction. For instance, Wang et al. [9] used the harmonic mean of effectiveness and efficiency metrics. Later, the same authors used a linear combination of efficiency and effectiveness metrics [8], as this presents an easier optimisation target for machine learning. However, the contrasting approaches suggest that a founded measure that represents both efficiency and effectiveness remains an open problem. At the same time, the metric used can considerably influence the trade-off optimum and it is vital to ensure the progress in that domain of research. Before discussing the proposed approach to build a such metric, in the next section we describe a dataset that is used in our research and a preliminary study of the influence of the search delays on the user clicking behaviour.

SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013), August 1, 2013, Dublin, Ireland.

2. DATASET AND EXPERIMENTS

As a dataset in our experiments we use the anonymised user behaviour data obtained from Yandex' Browser and Toolbar[1]. This data contains the delays that users experience while accessing Yandex result pages as well as information about the users' search result clicking behaviour. We believe that studying the influence of the natural delays occurring in the user's everyday interaction with the search engine is promising due to availability of massive amounts of data. In contrast, previous studies [1, 7] introduced an artificial delays to the user's interactions and thus are limited both in time and in user span.

The search result page access data obtained from the browser and the toolbar can be linked to the anonymised query log data, providing us with the data required to model the user behaviour in presence of the delays: search result pages with user clicks and delays associated. The experimental dataset is sampled from the logs spanning the time period 13th-27th May 2013. To reduce the sparseness of the user-related statistics, we remove all users with less than 5 sessions during the two-week period. In order to reduce influence from the previous search context, we consider only first queries in the sessions. The resulting dataset includes 3M users, 230K unique query-search result page pairs, and 10M sessions.

Since the perception of the delay might be extremely user-centric (e.g. some users are used to their slow Internet connection), for each user we calculate the median delay they experience while using Yandex and subtract it from all their session delays, thus reducing the personal bias.

In order to understand how delays influence the users' behaviour, we perform the following experiment. Using the available click log data, we simulate eight search engines with exactly the same effectiveness but different efficiency. The simulation procedure is inspired by the procedure used by Chapelle et al. [3] to evaluate effectiveness measures. For each unique combination of a query and a search result page (SERP) we sort all sessions associated according to the delay time experienced by the user. Next, we split the obtained session list in ten sub-lists of equal length, so that the first sub-list contains sessions with the smallest delay (i.e. the fastest 10% of sessions), while the last sub-list contains the

[1]browser.yandex.com/ and element.yandex.com/

sessions with the highest delay. We discard the first and the last sub-lists as they are likely to contain various outliers. The rest of the sub-lists are considered as "simulated" search engines. Notably, these sub-lists contain exactly the same query-SERP pairs, as well as the same number of sessions associated with them, with only the delays varied. In Figure 1 we report relative increase in the abandonment rate (AR) with respect to the first simulated search engine, as the relative delay increases. An increase of approximately 121% in the abandonment rate is reached when the delay time is increased by 540%.

From the figure it can be observed that the delays do influence the users' abandoning behaviour. Indeed, there is a statistically significant evidence that lower delays cause smaller abandonment rates. While such a dependency between user's dissatisfaction and higher delays has previously been reported in [1, 7], our obtained results support the validity of the proposed experimental setup, which is, as discussed, different from the one used in the previous works. On analysing the figure, we also notice that the increase in the abandonment rate grows approximately linearly with the increase in the delay time.

Overall, our results exhibit the expected trend between delays and user dissatisfaction. In the next section, we discuss our plan to incorporate the research on the user behaviour into an evaluation metric.

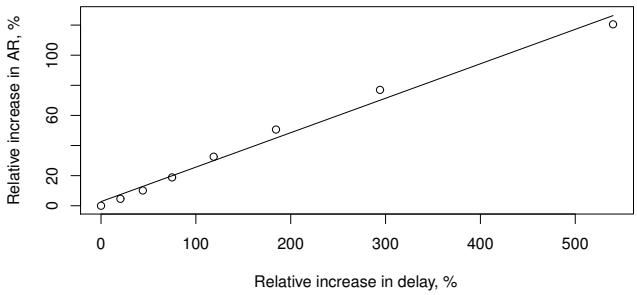
3. PROPOSED APPROACH

Following the work on state-of-the-art effectiveness evaluation metrics, such as Expected Reciprocal Rank (ERR) [3] and Expected Browsing Utility (EBU) [10], we devise our proposed metric basing on the user behaviour modelling. At a first step, we aim to model the user behaviour in presence of the system efficiency variability. In particular, we incorporate the user's tolerance to the system's delays into the cascade model [4] of the clicking behaviour, which underlies ERR. In order to get insight on how this can be achieved, we propose to train the Dynamic Bayesian Network (DBN) click model with abandonment [4] on the sessions from bins with different delays. As a result, we might expect that the following model parameters depend on the retrieval delay: the probability of examining the first document; the probability of abandoning after examining a non-relevant document (the user's tolerance to irrelevant results); the probability of clicking on results (readiness to spend time to check if the landing page is relevant). Furthermore, this delay-aware modification of the cascade model can be used to predict the probability of satisfying the user with document labels and system delay provided. In the final step, the probability of user satisfaction can be used as an evaluation metric, similar to the extension of ERR to address abandonment, proposed by Chapelle et al. [3].

Another possible direction might imply modification of the time-biased gain effectiveness metric proposed in [5]. However, directly accounting for the delay time might be fruitless, as the delays are likely to be considerably smaller than the characteristic time scale considered by this metric (e.g. half of the users continue their search after 224s [5]).

Having proposed a new metric, a question arises how to evaluate its quality, i.e. to check that it indeed represents the users' preferences and optimising it leads to higher users' satisfaction. We propose to leverage the evaluation approach considered by Chapelle et al. [3]. This approach studies the correlation of the considered metrics with the online satisfaction indicators, such as abandonment rate. We believe that the use of the abandonment rate will provide us with a reliable indicator of user satisfaction. A good efficiency-aware metric should outperform its effectiveness-only counterparts, i.e. efficiency-aware ERR should correlate with the online click metrics better than ERR. In addition, we expect it outperform the heuristic approaches used in the literature [7, 8].

Figure 1: Relative change in the abandonment rate (AR), as a function of the relative change in the delay. 95% error bars are smaller than the symbols. Linear fit corresponds to the line $y = 0.028 + 0.229 \cdot x$. $R^2 = 0.98$



4. CONCLUSIONS

In this work, we argued the need for an evaluation metric that combines both retrieval efficiency and retrieval effectiveness in a founded and empirically-verified manner. We proposed to devise such a metric by means of incorporating the users' tolerance to delays into the cascade model of the user behaviour. Further, this model can be used to build an efficiency-aware modification of the ERR effectiveness metric. We suggested to use the search engine click logs as well as the browser/toolbar data as a source of the user preference evidence. Finally, our preliminary experiments demonstrated that this data exhibits the same patterns as reported in previous studies and hence can be used to derive metrics such as those proposed in this paper.

5. REFERENCES

[1] J. D. Brutlag, H. Hutchinson, and M. Stone. User preference and search engine latency. In *JSM Proceedings, Qualtiy and Productivity Research Section*, 2008.

[2] B. B. Cambazoglu, H. Zaragoza, O. Chapelle, J. Chen, C. Liao, Z. Zheng, and J. Degenhardt. Early exit optimizations for additive machine learned ranking systems. WSDM '10.

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

[4] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. WWW '09.

[5] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. SIGIR '12.

[6] N. Tonellotto, C. Macdonald, and I. Ounis. Efficient and effective retrieval using selective pruning. WSDM '13.

[7] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. SIGKDD '09.

[8] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. SIGIR '11.

[9] L. Wang, J. Lin, and D. Metzler. Learning to efficiently rank. SIGIR '10.

[10] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. CIKM '10.