Improving Offline and Online Web Search Evaluation by Modelling the User Behaviour

Eugene Kharitonov
Yandex, Moscow, Russia
School of Computing Science, University of Glasgow, UK
kharitonov@yandex-team.ru

ABSTRACT

Measurements are fundamental to any empirical science and, similarly, search evaluation is a vital part of information retrieval (IR). Evaluation ensures the progressive development of approaches, tools, and methods studied in this field. Apart from the scientific perspective, the evaluation approaches are also important from the practical perspective. Indeed, the evaluation experiments enable commercial search engines to make data-driven decisions while developing new features and working on the quality of the user experience. Thus, it is not surprising that evaluation has gained a huge attention from the research community and such an interest spans almost fifty years of research [3]. The Cranfield experiments [3] evolved into the widely used offline system evaluation approach. Despite its convenience and popularity, the offline evaluation approach has several limitations [8]. These limitations resulted in the development and recent growth in popularity of the online user-based evaluation approaches such as interleaving and A/B testing [1].

There can be a considerable interplay between the user modelling, the online and the offline evaluation approaches. The online evaluation methods *interpret* the user's behaviour observed during the evaluation experiment to infer if the tested change leads to improvements in the user satisfaction. In contrast, some of the modern offline evaluation approaches proposed for document retrieval are devised to *predict* how the users will behave once the tested search algorithms are deployed and whether the users will be satisfied. Only recently has the offline evaluation methods started to see their foundation in the user modelling, as highlighted by modern effectiveness metrics, such as the ERR metric [2]. Similarly, some interleaving methods have recently started to be formulated in terms of the user click models [7].

We propose to strengthen these ties between the modelling of the user behaviour and the evaluation, and use the user modelling approach to address a number of applications in IR systems evaluation. Specifically, the statement of this thesis is that the user behaviour modelling can aid in addressing a number of applications in information retrieval evaluation. We hypothesise that the understanding of the user searching behaviour can help to devise novel offline evaluation metrics, new online evaluation approaches, and improve the existing online evaluation methods. We also demonstrate how the offline evaluation methods can developed to be in agreement with the online evaluation paradigm, thus effectively bridging the gap between the offline and the online approaches.

In particular, based on the user modelling approach, we aim to study how the offline evaluation methods can be extended to novel applications such as the query auto-completion mechanisms [5]. Our next goal is to investigate how the user model-based offline evaluation can be improved to account for non-effectiveness features, such as search efficiency [4]. We also aim to investigate how the interleaving evaluation methods can be improved based on the user behaviour modelling. More precisely, we study how to maximise the sensitivity of the interleaving methods by modelling the user behaviour [6], and how the applicability of interleaving can be extended to other domains, such as image search. Overall, we hope that our research will contribute to improving the evaluation approaches in IR.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: search evaluation, offline evaluation, online evaluation, user modelling

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2610379.

1. REFERENCES

[1] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM TOIS*, 30(1):6, 2012.

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

[3] C. Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194, 1967.

[4] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Incorporating efficiency in evaluation. In *SIGIR 2013 MUBE Workshop*.

[5] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. User model-based metrics for offline query suggestion evaluation. In *SIGIR 2013*.

[6] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Using historical click data to increase interleaving sensitivity. In *CIKM 2013*.

[7] F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In *WSDM 2013*.

[8] E. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, LNCS, pages 355–370. 2002.