Evaluating Heterogeneous Information Access (Position Paper)

Ke Zhou
University of Glasgow
zhouke@dcs.gla.ac.uk

Tetsuya Sakai
Microsoft Research Asia
tesakai@microsoft.com

Mounia Lalmas
Yahoo! Labs Barcelona
mounia@acm.org

Zhicheng Dou
Microsoft Research Asia
zhichdou@microsoft.com

Joemon M. Jose
University of Glasgow
jj@dcs.gla.ac.uk

ABSTRACT

Information access is becoming increasingly heterogeneous. We need to better understand the more complex user behaviour within this context so that to properly evaluate search systems dealing with heterogeneous information. In this paper, we review the main challenges associated with evaluating search in this context and propose some avenues to incorporate user aspects into evaluation measures.

1. INTRODUCTION

The performance of search systems is evaluated with *user-oriented* and *system-oriented* measures. The former are obtained through user studies conducted to examine and reflect upon various aspects of user behaviours. The latter rely on reusable test collections (i.e. document, query and relevance judgements) to assess the search quality.

Recent work [5, 9, 11] have attempted to combine those two types of measures, by modelling users and their behaviour within test collection based system-oriented evaluation metrics. For example, Smucker et al. [11] proposed a *time-biased gain* (TBG) framework that explicitly calibrates the time of various (user) aspects in the search process. Another attempt from Sakai et al. [9] proposed a unified evaluation framework (U-measure) that is free from *linear traversal* assumption and can evaluate information access other than ad-hoc retrieval (e.g. multi-document summarisation, diversified search). Recently, Chuklin et al. [5] proposed a common approach to convert click models into system-oriented evaluation measures. Although the above mentioned evaluation frameworks can potentially handle more complex search tasks, they all have been tested on traditional homogeneous search scenarios (newswire or general web search).

The web environment is becoming increasingly heterogeneous. We have now many search engines, so-called *verticals*, each targeting a specific type of information (e.g. image, news, video, etc.). *Aggregated search* [2, 14] is concerned with retrieving search results from a heterogeneous set of search engines, and is a topic of investigation in both the academic community and commercial world. Due to the heterogeneous nature of information in aggregated search, numerous challenges have arisen.

In this paper, we argue that, compared with traditional homogeneous search, evaluation in the context of heterogeneous information is more challenging and requires taking into account more complex user behaviours and interactions. Specifically, we require evaluation approaches that not only model user behaviours but also adapt to how users interact with an heterogeneous information space.

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

2. CHALLENGES

There are three main challenges in incorporating user behaviours within an evaluation framework for heterogeneous information access. We discuss each challenge and current research endeavours for each below.

2.1 Non-linear Traversal Browsing

Presenting heterogeneous information is more complex than the typical single ranked list (e.g. ten blue links) employed in homogeneous ranking. There are three main types of presentation designs: (1) results from the different verticals are blended into a single list (of blocks), referred to as *blended*; (2) results from each vertical are presented in a separate (e.g. horizontal paralleled) panel (tile), referred to as *non-linear blended*; and (3) vertical results can be accessed in separate tabs, referred to as *tabbed*. A combination of all three is also possible. In addition, results from different vertical search engines can be grouped together to form a coherent "bundle" for a given aspect of the query (e.g. a bundle composed of a news article along with videos and user comments as a response to a query "football match"). Finally, the results presented on the search page can contain visually salient snippets (e.g. image).

Different presentation strategies and visual saliencies imply different patterns of user interaction. For example, a user could follow a non-linear traversal browsing pattern. Through eye-tracking studies [13] and search log analysis [7, 12], recent studies have shown that in a *blended* presentation, users tend first to examine results from one vertical (vertical bias), in particular those with visual salient snippets, and results nearby. In addition, when the vertical results are not presented at the top of the search result page, users tend to scan back to re-examine previous web results either bottom-up and top-down. When presenting in a *non-linear blended* style (two parallel panels/columns), a recent eye and mouse tracking study [8] showed that users tend to firstly focus on examining top results on the first column and then jump to the right panel afterwards. For a *tabbed* presentation of vertical results, the user browsing behaviour

is still poorly understood.

2.2 Diverse Search Tasks

User search tasks are more complex with heterogeneous information access than traditional homogeneous ranking. A search task's vertical orientation can affect user behaviours. Previous research [12] showed that the strength of a user's search task's orientation towards a particular source (vertical) type is different, and this affects user's search behaviour (for example, click-through rates). Recently, Zhou et al. [15] found that when assessing vertical relevance, a search task's vertical orientation is more important than the topical relevance of the retrieved results.

Secondly, search task's complexity can also have a major effect on user behaviours. This is because users can access results from different verticals to accomplish their search tasks in multiple search sessions. Arguello et al. [2] showed that more complex search tasks require significantly more user interaction and more examination of vertical results. Finally, Bron et al. [6] found that user's preference for aggregated search presentation (*blended* and *tabbed*) changes during multi-session search tasks.

2.3 Coherence, Diversity and Personalization

Another important consideration when evaluating heterogeneous information access is coherence. This refers to the degree to which results from different verticals focus on a similar "sense" of the query (can they form a bundle?). Recent research [3] showed that query-senses associated with the *blended* vertical results can influence user interaction with web search results.

The diversity of the results is another interesting problem. It has been shown to be considerably different, and that users often have their own *personalized* vertical diversity preferences [14]. Finally, Santos et al. [10] showed that for an ambiguous or multi-faceted query, user's intended information need varies considerably across different verticals.

3. AVENUES OF RESEARCH

We need an approach that models the above mentioned user behaviours and incorporates them into system-oriented measures to evaluate heterogeneous information access. This requires two main lines of research: (1) understanding and modelling users behaviours, and (2) incorporating these into the evaluation. We elaborate on each below.

The first line of research aims to give insights on the user perspectives and provide better models of user behaviours. Although there have been studies aiming at better understanding the behaviour of users in aggregated search, the problem of evaluating heterogeneous information access is far from solved. There remains a large gap between understanding user behaviours in this context and incorporating this understanding into the evaluation measures. There has been attempts at building models of aggregated search clicks [7, 13] which could be incorporated in measures, e.g. to account for search task's vertical orientation and vertical visual saliency. However, many aspects still lack investigation (e.g. coherence, diversity). We propose to follow current research endeavours and investigate models to capture user aspects, in particular those poorly accounted for in the evaluation. To achieve this, we must collect data on user behaviours for aggregated search through laboratory experiments, crowd-sourcing or accessing search engine logs.

The second line of research aims to incorporate these new models into a general evaluation framework that can accurately capture the variations in user behaviours. There are few powerful evaluation frameworks that we could use for this, for instance, TBG [11] and U-measure [9] as mentioned in Section 1. Zhou et al. [14] also recently proposed a general evaluation framework to model utility and effort in aggregated search. In addition, we could follow Chuklin et al [5] and convert obtained aggregated search click models into system-oriented evaluation. Preference-based evaluation approach is another direction that is worth of attention, for instance, Chandar et al. [4] and Arguello et al. [1].

4. CONCLUSIONS

This paper advocates the need to incorporate user behaviours into system-oriented measures for evaluating heterogeneous information access. We listed challenges and proposed some avenues for shaping future research in this direction. A new track at TREC, FedWeb[1], is studying information access for heterogeneous information, and is the perfect forum to carry some of the research avenues discussed in this paper.

5. REFERENCES

[1] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In ECIR 2011.

[2] J. Arguello, W. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display, and user interaction in aggregated search. In SIGIR 2012.

[3] J. Arguello, and R. Capra. The effect of aggregated search coherence on search behavior. In CIKM 2012.

[4] P. Chandar, and B. Carterette. Preference based evaluation measures for novelty and diversity. In SIGIR 2013.

[5] A. Chuklin, P. Serdyukov, and M. Rijke. Click model-based information retrieval metrics. In SIGIR 2013.

[6] M. Bron, J. Grop, F. Nack, L.B. Baltussen, and M. Rijke. Aggregated search interface preference in multi-session search tasks. In SIGIR 2013.

[7] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: enabling user click modeling in federated web search. In WSDM 2012.

[8] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In WWW 2013.

[9] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In SIGIR 2013.

[10] R. L. T. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In ICTIR 2011.

[11] M. Smucker, and C. Clarke. Time-based calibration of effectiveness measures. In SIGIR 2012.

[12] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In CIKM 2010.

[13] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In SIGIR 2013.

[14] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In SIGIR 2012.

[15] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Which vertical search engines are relevant? Understanding vertical relevance assessments for web queries. In WWW 2013.

[1] Federated web search: https://sites.google.com/site/trecfedweb/.