Diversity, Intent, and Aggregated Search

Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
derijke@uva.nl

1. BACKGROUND

Diversity, *intent* and *aggregated search* are three core retrieval concepts that receive significant attention. In search result *diversification* one typically considers the relevance of a document in light of other retrieved documents. The goal is to identify the probable "aspects" of an ambiguous query, retrieve documents for each of these aspects and make the search results more diverse. By doing so, in the absence of any knowledge of users' context or preferences, the chance that the user will find at least one of these results to be relevant to their underlying information need is increased. Those probable "aspects" of a query may refer to lexical ambiguity (e.g., *flash* – Adobe Flash, flash light, flash gordon, flash airlines, flash mob, ...) or to *intentional* ambiguity (e.g., *pizza* – how to make one, where to buy one, images, nutritional value, background, restaurant, ...). The automatic discovery of query intent has become an active research area, with a range of observational and algorithmic studies as outcomes. Understanding the likely intents behind a query can help search engines to automatically route the query to the corresponding vertical search engines so as to obtain particularly relevant results, thus greatly improving user satisfaction. In *aggregated* search the task is to search and assemble information from a variety of sources and to organize the resulting material within a single interface. The result page of a modern search engine often goes beyond a simple ranked list. Many specific intents are addressed by aggregated search solutions: specially presented documents, often retrieved from specific sources, that stand out from the regular organic search results.

2. RECENT ADVANCES

Diversity, intent, and aggregated search give rise to significant research challenges, both algorithmically and in terms of evaluation. In the talk I will highlight recent developments and point out directions for future work. In particular, concerning search result diversification I will run through a new perspective on the problem by casting it as a data fusion problem, following [5], and inferring latent topics of the query for which the result set is being diversified. A second important algorithmic development concerns recent work on *personalized* search result diversification, based on structured learning [6]. Result items can be diverse because they address multiple *intents*, either by belong to multiple collections or genres, or by addressing different informational uses. Increasingly, weakly supervised methods are being used to uncover intents. In addition, complex characteristics of documents such as frames (the way in which an issue is depicted in terms of a 'central organizing idea') are being studied to facilitate new diversification methods for, e.g., news search [7]. Finally, a big challenge in aggregated search is the evaluation of complex result layouts. This is especially true for so-called *interleaving* methods: by mixing results from different result pages interleaving can easily break the desired web layout in which vertical documents are grouped together, and hence hurt the user experience. I briefly describe recently proposed vertical-aware interleaving methods [2].

Making progress on the connection between diversity, intent and aggregated search is at least as important as deepening our understanding in these areas in isolation. I will highlight progress on the interface of these areas by focusing on three examples. In the first, I focus on result pages containing fresh results and propose a way to model user intent distribution and bias due to different document presentation types [1]. In the second, I focus on the fresh vertical prediction task for repeating queries and address the following algorithmic problem: how to quickly and accurately detect fresh intent shifts and adjust the ranking in an online setting [3]. Finally, I consider a scenario where a single intent may be served by multiple verticals, which leads to a new ranking problem [4].

The talk is based on joint work with Björn Burscher, Aleksandr Chuklin, Damien Lefortier, Shangsong Liang, Daan Odijk, Zhaochun Ren, Fedor Romanenko, Anne Schuth, Pavel Serdyukov, Rens Vliegenthart, and Ke Zhou.

3. REFERENCES

[1] A. Chuklin, P. Serdyukov, and M. de Rijke. Using intent information to model user behavior in diversified search. In *ECIR '13*. Springer, March 2013.

[2] A. Chuklin, A. Schuth, K. Zhou, and M. de Rijke. A comparative analysis of interleaving methods for aggregated search. *ACM Transactions on Information Systems*, 33, 2015. To appear.

[3] D. Lefortier, P. Serdyukov, and M. de Rijke. Online exploration for detecting shifts in fresh intent. In *CIKM '14*. ACM, November 2014.

[4] D. Lefortier, P. Serdyukov, F. Romanenko, and M. de Rijke. Blending vertical and web results: A case study using video intent. In *ECIR '14*. Springer, April 2014.

[5] S. Liang, Z. Ren, and M. de Rijke. Fusion helps diversification. In *SIGIR '14*. ACM, July 2014.

[6] S. Liang, Z. Ren, and M. de Rijke. Personalized search result diversification via structured learning. In *KDD '14*. ACM, August 2014.

[7] D. Odijk, B. Burscher, R. Vliegenthart, and M. de Rijke. Automatic thematic content analysis: Finding frames in news. In *SocInfo2013*. ACM, November 2013.

ADCS '14, Nov 27–28 2014, Melbourne, VIC, Australia
ACM 978-1-4503-3000-8/14/11.
http://dx.doi.org/10.1145/2682862.2684462.