Exploiting User Signals and Stochastic Models to Improve Information Retrieval Systems and Evaluation

Maria Maistro
University of Padua, Padua, Italy
maistro@dei.unipd.it

Abstract

Progress and innovation are driven by experiments, but experimentation is useless without an objective evaluation measure that allow researchers to detect the improvements and identify the successful strategies. Due to the experimental nature of Information Retrieval (IR), accurately interpreting the result of a system in terms of user satisfaction is fundamental to push the research in the correct direction. Therefore, measuring systems effectiveness continues to be an active area of research and discussion in the scientific community. It is also the case of this thesis, whose leitmotiv is an investigation of effectiveness measures exploited in different aspects of IR.

Our first aim was to provide a formal and theoretical definition of effectiveness measure. Several evaluation measures have been proposed since the beginning of IR, starting from simple ratios between relevant and retrieved documents to more complex functions discounting each rank positions and accounting for plausible user models. However, even if much research was conducted, a prior question is still just partially fulfilled: what is a general definition of IR evaluation measure? We tackle this challenge and give a formal definition of utility-oriented measurement of retrieval effectiveness [2], based on the representational theory of measurement.

A further complexity of evaluation in IR is represented by relevance. Unfortunately relevance is subjective, the information need is unique, and the user is the only person able to provide a fair and reliable judgement of a document in terms of relevance. Since it is not possible to directly ask to the user to provide relevance judgements when she performs a search, IR evaluation relies on external assessors to judge documents for relevance.

The same query-document pair is assigned to more than one crowd assessor to prevent potential errors caused by wrong labels. This makes necessary to merge possibly discording labels generated by different workers. We propose our upstream approach called Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE) [3]. AWARE is defined as an upstream approach because it directly combines the scores of the evaluation measures computed from the relevance labels of each assessor, instead of merging the labels and then computing the measures. This allows to account for the error introduced by incorrect labels and to develop a framework which estimates performance measures in a way more robust to the potential noise introduced by crowd assessors.

The effectiveness of a system can be measured in terms of the amount of relevance retrieved, however, what about the user perspective? Is it possible to account more for the user-system interactions? These questions suggest that the user and her interactions with the system can be included in the evaluation process.

To this end, we present a novel user model defined on top of a Markov process. Differently from many traditional models, which assume a user linearly scanning the result list, this model allows the user to follow complex paths when browsing the run, as moving backward and forward in the list, skipping some documents or considering already visited documents. Based on this model, we defined two different evaluation measures Markov Precision (MP) [1] and Normalized Markov Cumulated Gain (nMCG) [4], both involving the user in the evaluation process.

MP [1] injects the user model into precision with the invariant distribution of a Markov chain, which is the probability of finding the user in a given rank position after a long time. MP stems from the idea that if a user does not see a document, even if the document is relevant, the evaluation measure should account less for it, while it should account more for documents that have been visited.

By exploiting the same model, we defined nMCG-MART [4] a measure calibrated with real word click log data. nMCG-MART is exploited as objective function in a state of the art Learning to Rank (LtR) algorithm. By using nMCG-MART as objective function, we embed the user dynamic in the learning process and we push the algorithm to rank the results by considering the user experience.

Supervisor: Prof. Nicola Ferro, University of Padua
Available from: `http://paduaresearch.cab.unipd.it/10819/`

References

[1] Ferrante, M., Ferro, N., and Maistro, M. (2014). Injecting User Models and Time into Precision via Markov Chains. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 597–606, New York, NY, USA. ACM.

[2] Ferrante, M., Ferro, N., and Maistro, M. (2015). Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 21–30, New York, NY, USA. ACM.

[3] Ferrante, M., Ferro, N., and Maistro, M. (2017). AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM Transaction on Information Systems*, 36(2):20:1–20:38.

[4] Ferro, N., Lucchese, C., Maistro, M., and Perego, R. (2017). On Including the User Dynamic in Learning to Rank. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1041–1044, New York, NY, USA. ACM.