Finding all the Needles in a Haystack

A System to Estimate the Costs of e-Discovery and Systematic Reviews

Giorgio Maria Di Nunzio
Dept. of Information Engineering, University of Padua
Padova, Italy
giorgiomaria.dinunzio@unipd.it

ABSTRACT

Systematic review and e-Discovery have a common task in which the objective is to find most (if not all) of the relevant documents in a collection by means of a (semi-)manual screening of the potentially interesting documents [3]. However, the high cost of e-Discovery software and the management of the advanced e-Discovery mechanism are expected to affect the growth of this market (which is expected to reach 17.32 billion dollars by 2023). [1] Moreover, the large and growing number of published studies makes the task of identifying relevant studies in systematic reviews in an unbiased way both complex and time consuming [4]. In this paper, we present an active learning system which combines different sampling approaches in order to estimate a 95% confidence interval of the number of relevant documents while taking into account the monetary costs of running the system itself.

KEYWORDS

Probabilistic Retrieval Models, Continuous Active Learning

SYSTEM OVERVIEW

In [3], the authors proposed a jointly research agenda between e-Discovery and systematic review by focusing on the question of "how much is enough". For e-Discovery, this question is partially answered by the effort required of additional document review compared to the expected impact on legal proceedings [5]. For systematic reviews, there are professional guidelines that describe the application of the scientific method to uncover and minimize bias and error in the selection and treatment of studies [4]. In order to save hours of tedious work, the Cochrane Transform Project is now applying a machine learning process in order to analyze thousands of reports and automatically select those to include in systematic reviews. This process addresses the difficulty in finding reports of studies for inclusion in a review in a reliable way.[2]

In this paper, we discuss some of the points of the research agenda [3] by means of an interactive system for systematic reviews based on an active learning framework [2]. This system monitors the monetary costs of a variable thresholding approach and a mixed strategy for sampling documents. This approach achieved a recall greater than 95% with 25,000 documents less than the best performing systems during an evaluation task for technology assisted reviews in empirical medicine [1].[3] In Figure 1, we show a high level overview: 1) the model ranks the document in the collection and shows the top ranked document to the user; 2) the user gives the feedback on the document (either relevant or not) and this feedback is used to re-train the model and re-rank the document of the collection. The system allows to adjust the proportion of documents that are sampled against those that are selected by the automatic system in order to balance the amount of money we want to spend to estimate the confidence intervals more accurately or get the most relevant information as quick as possible.

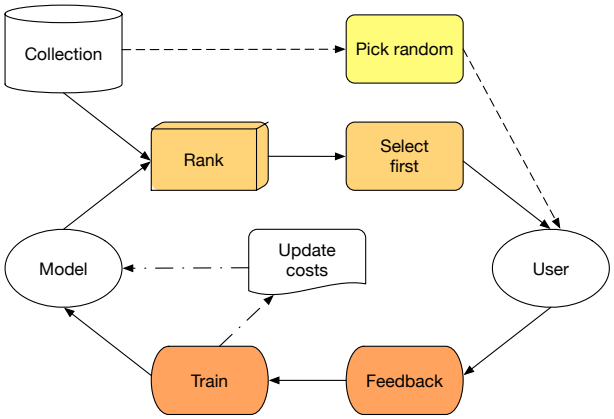

Figure 1: High level overview of the system. The system is continuously updated with the explicit relevance feedback of the user. Randomly picked documents are used to estimate the number of relevant documents in the collection.

REFERENCES

[1] G. M. Di Nunzio. 2018. A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews. In *Proc. of ECIR 2018.* Springer, 672–677. https://doi.org/10.1007/978-3-319-76941-7_61
[2] G. M. Di Nunzio, M. Maistro, and F. Vezzani. 2018. A Gamified Approach to Naïve Bayes Classification: A Case Study for Newswires and Systematic Medical Reviews. In *Companion of the The Web Conference 2018 WWW 2018, Lyon , France, April 23-27, 2018.* 1139–1146. https://doi.org/10.1145/3184558.3191547
[3] M. Lease, G. V. Cormack, A. T. Nguyen, T. A. Trikalinos, and B. C. Wallace. 2016. Systematic review is e-discovery in doctor's clothing. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR).*
[4] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 51 (2014), 242 – 253.
[5] D. W. Oard, J. R. Baron, B. Hedin, D. D. Lewis, and S. Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *Artif. Intell. Law* 18, 4 (2010), 347–386. https://doi.org/10.1007/s10506-010-9093-9

[1] https://www.researchandmarkets.com/research/v5kb4w
[2] http://community.cochrane.org/help/tools-and-software/evidence-pipeline/
[3] https://github.com/CLEFeHealth

DESIRES 2018, August 2018, Bertinoro, Italy