Unsupervised, Efficient and Semantic Expertise Retrieval

Extended abstract

Christophe Van Gysel
cvangysel@uva.nl

Maarten de Rijke
derijke@uva.nl

Marcel Worring
m.worring@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We introduce an unsupervised discriminative model for the task of retrieving experts in online document collections [5]. We exclusively employ textual evidence and avoid explicit feature engineering by learning distributed word representations in an unsupervised way. We compare our model to state-of-the-art unsupervised statistical vector space and probabilistic generative approaches. Our proposed log-linear model achieves the retrieval performance levels of state-of-the-art document-centric methods with the low inference cost of so-called profile-centric approaches. It yields a statistically significant improved ranking over vector space and generative models in most cases, matching the performance of supervised methods on various benchmarks.

1. INTRODUCTION

The expertise retrieval task gained popularity in the research community during the TREC Enterprise Track [4]. Existing methods fail to address key challenges: (1) Queries and expert documents use different representations to describe the same concepts [2, 3]. (2) As the amount of available data increases, the need for more powerful approaches with greater learning capabilities than smoothed maximum-likelihood language models is obvious [6]. (3) The acceleration of data availability has the major disadvantage that, in the case of supervised methods [1], manual annotation efforts need to sustain a similar order of growth. This calls for the further development of *unsupervised* methods. Our proposed solution has a strong emphasis on *unsupervised model construction*, *efficient query capabilities* and *semantic matching* between query terms and candidate experts.

2. A LOG-LINEAR MODEL FOR EXPERT SEARCH

We propose an unsupervised log-linear model with efficient inference capabilities for the expertise retrieval task. We show that our approach improves retrieval performance compared to vector space-based and generative language models, mainly due to its ability to perform semantic matching [3]. Our method does not require supervised relevance judgments and is able to learn from raw textual evidence and document-candidate associations alone. The purpose of this work is to provide insight in how discriminative language models can improve performance of core retrieval tasks compared to maximum-likelihood language models.

3. DISCUSSION

We evaluated our model on the W3C, CERC and TU benchmarks and compared it to state-of-the-art vector space-based entity ranking (based on LSI and TF-IDF) and language modeling (profile-centric and document-centric) approaches. The log-linear model combines the ranking performance of the best maximum-likelihood language modeling approach (document-centric) with inference time complexity linear in the number of candidate experts. We observed a notable increase in precision over existing methods. Analysis of our model's output reveals a negative correlation between the per-query performance and ranking uncertainty: higher confidence (i.e., lower entropy) in the rankings produced by our approach often occurs together with higher rank quality.

An error analysis of the log-linear model and traditional language models shows that the two make very different errors. These errors are mainly due to the semantic gap between query intent and the raw textual evidence. Some benchmarks expect exact query matches, others are helped by our semantic matching. An ensemble of methods employing exact and semantic matching generally outperforms the individual methods. This observation calls for further research in the area of combining exact and semantic matching.

REFERENCES

[1] Y. Fang, L. Si, and A. P. Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *SIGIR*, pages 683–690, 2010.

[2] G. E. Hinton. Learning distributed representations of concepts. In *8th Annual Conference of the Cognitive Science Society*, volume 1, page 12, Amherst, MA, 1986.

[3] H. Li and J. Xu. Semantic matching in search. *Found. & Tr. in Information Retrieval*, 7(5):343–469, June 2014.

[4] TREC. Enterprise Track, 2005–2008.

[5] C. Van Gysel, M. de Rijke, and M. Worring. Unsupervised, efficient and semantic expertise retrieval. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1069–1079. International World Wide Web Conferences Steering Committee, 2016.

[6] V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

This is an extended abstract of an Article published in the 25th International World Wide Web Conference.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2882974.