UvA-DARE (Digital Academic Repository)

Semantic Entities

Van Gysel, C.; de Rijke, M.; Worring, M.

[Link to publication](#)

Semantic Entities

(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

Entity retrieval has seen a lot of interest from the research community over the past decade. Ten years ago, the expertise retrieval task gained popularity in the research community during the TREC Enterprise Track [10]. It has remained relevant ever since, while broadening to social media, to tracking the dynamics of expertise [1–5, 8, 11], and, more generally, to a range of entity retrieval tasks.

In the talk, which will be given by the second author, we will point out that existing methods to entity or expert retrieval fail to address key challenges: (1) Queries and expert documents use different representations to describe the same concepts [6, 7]. Term mismatches between entities and experts [7] occur due to the inability of widely used maximum-likelihood language models to make use of *semantic* similarities between words [9]. (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 [13]. (3) Supervised methods for entity or expertise retrieval [5, 8] were introduced at the turn of the last decade. However, the acceleration of data availability has the major disadvantage that, in the case of supervised methods, manual annotation efforts need to sustain a similar order of growth. This calls for the further development of *unsupervised* methods. (4) According to some entity or expertise retrieval methods, a language model is constructed for every document in the collection. These methods lack *efficient* query capabilities for large document collections, as each query term needs to be matched against every document [2].

In the talk we will discuss a recently proposed solution [12] that has a strong emphasis on *unsupervised model construction*, *efficient query capabilities* and, most importantly, *semantic matching* between query terms and candidate entities. We show that the proposed approach improves retrieval performance compared to generative language models mainly due to its ability to perform *semantic matching* [7]. The proposed method does not require any annotations or supervised relevance judgments and is able to learn from raw textual evidence and document-candidate associations alone.

The purpose of the proposal is to provide insight in how we avoid explicit annotations and feature engineering and still obtain semantically meaningful retrieval results. In the talk we will provide a comparative error analysis between the proposed semantic entity retrieval model and traditional generative language models that perform exact matching, which yields important insights in the relative strengths of semantic matching and exact matching for the expert retrieval task in particular and entity retrieval in general.

We will also discuss extensions of the proposed model that are meant to deal with scalability and dynamic aspects of entity and expert retrieval.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

Entity search; distributional semantics.

REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *IPM*, 45:1–19, 2009.

[2] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Found. & Tr. in Information Retrieval*, 6 (2-3):127–256, 2012.

[3] R. Berendsen, M. de Rijke, K. Balog, T. Bogers, and A. van den Bosch. On the assessment of expertise profiles. *JASIST*, 64(10):2024–2044, 2013.

[4] Y. Fang and A. Godavarthy. Modeling the dynamics of personal expertise. In *SIGIR*, pages 1107–1110, 2014.

[5] 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.

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

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

[8] C. Moreira, B. Martins, and P. Calado. Using rank aggregation for expert search in academic digital libraries. In *Simpósio de Informática, INForum*, pages 1–10, 2011.

[9] R. Salakhutdinov and G. Hinton. Semantic hashing. *Int. J. Approximate Reasoning*, 50(7):969–978, 2009.

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

[11] D. van Dijk, M. Tsagkias, and M. de Rijke. Early detection of topical expertise in community question and answering. In *SIGIR*, 2015.

[12] C. Van Gysel, M. de Rijke, and M. Worring. Unsupervised, efficient and semantic expertise retrieval. In *Submitted*, 2015.

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