Document Filtering for Long-tail Entities (abstract)

Ridho Reinanda[†]
r.reinanda@uva.nl

Edgar Meij[‡]
edgar.meij@acm.org

Maarten de Rijke[†]
derijke@uva.nl

[†] University of Amsterdam, Amsterdam, The Netherlands
[‡] Bloomberg L.P., London, United Kingdom

ABSTRACT

A knowledge base contains information about entities, their attributes, and their relationships. Modern search engines rely on knowledge bases for query understanding, question answering, and document enrichment [1, 6]. Knowledge-base construction, either based on web data or on a domain-specific collection of documents, is the cornerstone that supports a large number of downstream tasks. In this paper, we consider the task of entity-centric document filtering, which was first introduced at the TREC KBA evaluation campaign [4]. Given an entity, the task is to identify documents that are relevant and vital for enhancing a knowledge base entry of the entity given a stream of incoming documents.

To address this task, a series of *entity-dependent* and *entity-independent* approaches have been developed over the years. Entity-dependent approaches use features that rely on the specifics of the entity on which they are trained and thus do not generalize to unseen entities. Such methods include approaches that learn a set of keywords related to each entity and utilize these keywords for query expansion and document scoring [3, 5] as well as text-classification-based approaches that build a classifier with bag-of-word features for each entity. Signals such as Wikipedia page views and query trends have been shown to be effective, since they usually hint at changes happening around an entity [2]; these signals are typically available for popular entities but when working with long-tail entities, challenges akin to the cold-start problem arise. In other words, features extracted from and working for popular entities may simply not be available for long-tail entities.

In this paper, we are particularly interested in filtering documents for long-tail entities. Such entities have limited or even no external knowledge base profile to begin with. Other extrinsic resources may be sparse or absent too. This makes an entity-dependent document filtering approach a poor fit for long-tail entities. Rather than learning the specifics of each entity, *entity-independent* approaches to document filtering aim to learn the characteristics of documents suitable for updating a knowledge base profile by utilizing signals from the documents, the initial profile of the entity (if present), and relationships between entities and documents [2, 7, 8]. While entity-dependent approaches might be able to capture the distributions of features for each entity better, entity-independent approaches have the distinct advantage of being applicable to unseen entities, i.e., entities not found in the training data. As an aside, entity-independent methods avoid the cost of building a model for each entity which is simply not practical for an actual production-scale knowledge base acceleration system.

Our main hypothesis is that a rich set of *intrinsic* features, based on aspects, relations, and the timeliness of the facts or events mentioned in the documents that are relevant for a given long-tail entity, is beneficial for document filtering for such entities. We consider a rich set of features based on the notion of *informativeness*, *entity-saliency*, and *timeliness*. The intuition is that a document (1) that contains a rich set of facts in a timely manner, and (2) in which the entity is prominent makes a good candidate for enriching a knowledge base profile. To capture informativeness, we rely on three sources: generic Wikipedia section headings, open relations, and schematized relations in the document. To capture entity-saliency, we consider the prominence of an entity with respect to other entities mentioned in the document. To capture timeliness, we consider the time expressions mentioned in a document. We use these features with other basic features to train an entity-independent model for document filtering for long-tail entities.

Our main contributions can be summarized as follows: (1) We propose a competitive entity-independent model for document filtering for long-tail entities with rich feature sets designed to capture informativeness, entity-saliency, and timeliness. (2) We provide an in-depth analysis of document filtering for knowledge base acceleration for long-tail entities. We have shown that the proposed features successfully improve the filtering performance on long-tail entities.

This paper is interesting for the DIR community because it discusses how a document filtering method can be tailored towards the challenging problem of maintaining long-tail entity profiles in knowledge bases. This paper was published at CIKM 2016.

REFERENCES

[1] N. Balasubramanian and S. Cucerzan. Topic pages: An alternative to the ten blue links. In *ICSC '10*. IEEE, 2010.

[2] K. Balog, H. Ramampiaro, N. Takhirov, and K. Nørvåg. Multi-step classification approaches to cumulative citation recommendation. In *OAIR '13*, pages 121–128. Le Centre De Hautes Etudes Internationales D'Informatique Documentaire, 2013.

[3] L. Dietz and J. Dalton. UMass at TREC 2013 Knowledge Base Acceleration track. In *TREC 2013*. NIST, 2013.

[4] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, Z. Ce, R. Christopher, and I. Soboroff. Building an entity-centric stream filtering test collection for TREC 2012. In *TREC 2012*. NIST, 2012.

[5] X. Liu, J. Darko, and H. Fang. A related entity based approach for knowledge base acceleration. In *TREC 2013*. NIST, 2013.

[6] N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In *ACL-IJCNLP 2015*, 2015.

[7] J. Wang, D. Song, C. Lin, and L. Liao. BIT and MSRA at TREC KBA CCR Track 2013. In *TREC 2013*. NIST, 2013.

[8] J. Wang, D. Song, Q. Wang, Z. Zhang, L. Si, L. Liao, and C.-Y. Lin. An entity class-dependent discriminative mixture model for cumulative citation recommendation. In *SIGIR '15*, pages 635–644. ACM, 2015.