Combining Document- and Paragraph-Based Entity Ranking

Henning Rode, Pavel Serdyukov, Djoerd Hiemstra
Database Group, University of Twente
Enschede, The Netherlands
{rodeh, serdyukovpv, hiemstra}@cs.utwente.nl

ABSTRACT

We study entity ranking on the INEX entity track and propose a simple graph-based ranking approach that enables to combine scores on document and paragraph level. The combined approach improves the retrieval results not only on the INEX testset, but similarly on TREC's expert finding task.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval;

General Terms

Performance, Experimentation

Keywords

entity ranking, expert finding, graph-based retrieval

1. INTRODUCTION

When people use retrieval systems, they are often not searching for documents or text passages in the first place, but for some information contained inside. Often *named entities* like person names, organizations, or locations play a central role in answering such information needs. The INEX entity ranking track addresses this issue with a testset of entity ranking topics and corresponding judgments on the INEX Wikipedia corpus. All topics specify a target entity type and a topic of interest in a few query terms. The target type is given as a Wikipedia category, e.g. "movies", "trees", or "programming languages". In contrast to other entity ranking tasks, each retrieved entity in the INEX track needs to have its own article in the Wikipedia collection. Obviously, this decision is only suitable for entity ranking within an encyclopedia, where we can assume that most mentioned entities in fact have their own entry. In consequence, a simple baseline run is given by a straightforward article ranking using the query terms that describe the topic of interest. Combined with an appropriate category filtering mechanism that also allows articles of descendant categories, such a baseline can reach already a high retrieval quality [4].

However, the described baseline approach shows no techniques so far that are specific to entity ranking. Looking into the domain of expert finding, which can be considered as a typical entity ranking task with fixed entity type, most approaches establish connections between documents and contained mentions of entities and use these connections to propagate the relevance from initially retrieved documents towards the entities. We want to show in this paper how the relevance propagation approach can be introduced to the setting of the INEX entity ranking track. Furthermore, we will extend the existing propagation model by incorporating text fragments of various sizes.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. EXPLOITING DOCUMENT ENTITY RELATIONS

Entity mentions in Wikipedia articles are often linked towards their own encyclopedia entry. If we use these links to build a query-dependent entity containment graph, consisting of the top k initially retrieved entries and all their included linked entities, we can apply known retrieval models from expert finding. Balog et al. rank expert entities according to the relevance sum of all documents mentioning the expert [1]. In our containment graph model a similar approach can be expressed by calculating a weighted indegree $wIDG$:

$$wIDG(e) = \sum_{e' \in \Gamma(e)} w(e'|q),$$

where $\Gamma(e)$ denotes the set of vertices adjacent to e, and $w(e'|q)$ specifies the relevance weight of e's encyclopedia entry with respect to the query q. Notice that our weighted indegree ranking neither requires a probabilistic scoring model nor normalized associations weights between adjacent entries as in the work of Balog et al. Initial experiments showed that this model does not improve over our initial baseline ranking. It even decreases the retrieval quality considerably. In fact, the direct description of an entity given in its own Wikipedia article is so important for the ranking that it needs to be considered in the retrieval model. Hence, we suggest the following extension of the weighted indegree:

$$PwIDG(e) = \lambda w(e|q) + (1-\lambda) \sum_{e' \in \Gamma(e)} w(e'|q).$$

The factor λ interpolates the initial article relevance with the summed relevance of other articles mentioning entity e. Since the above equation resembles the definition of a *personalized* graph centrality by incorporating the weighting of the vertices themselves, we call it personalized weighted indegree $PwIDG$.

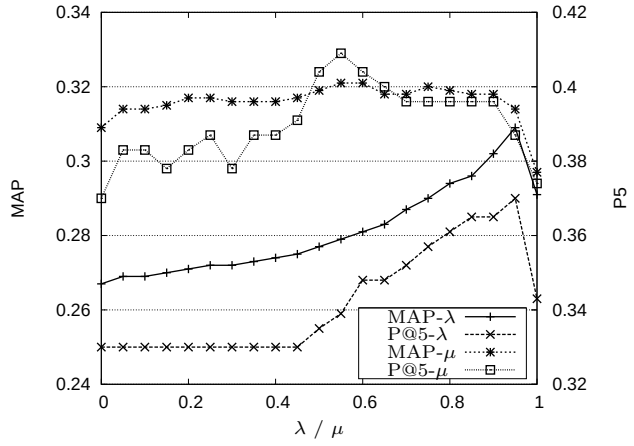

Figure 1: Influence of λ and μ on retrieval quality

2.1 Adding Smaller Sized Text Fragments

Related work on expert finding shows that proximity features help to further improve the retrieval quality. Proximity features have been integrated either in the relevance estimation model itself [3], or by tightening the initially ranked text fragments [5]. Within the framework of entity containment graphs, the latter means to combine paragraph and article level relevance by simply adding vertices of both types to the graph.

Since the Wikipedia collection contains structured text, we can make use of the given paragraph segmentation and retrieve and score XML <P> elements as well. An entity e is then linked by other entities e' or paragraph vertices p when their text refers to e. For distinction, we denote the set of neighboring paragraph vertices of an entity e by $\Gamma^P(e)$, respectively $\Gamma^E(e)$ for the set of adjacent entities. In order to control the influence of both types of text fragments, a second interpolation factor μ is introduced:

$$\begin{aligned} w(\Gamma(e)) &= \mu \sum_{p \in \Gamma^P(e)} w(p|q) + (1-\mu) \sum_{e' \in \Gamma^E(e)} w(e'|q), \\ PwIDG^*(e) &= \lambda w(e|q) + (1-\lambda) w(\Gamma(e)). \end{aligned}$$

3. EXPERIMENTS

For the initial ranking as well as for the graph generation the PF/Tijah retrieval system [2] was employed. We generated for each of the 46 INEX topics an XQuery that creates an entity containment graph from the top 200 articles retrieved by the title keywords. A standard language modeling retrieval model was employed for the initial scoring of text nodes.

Analyzing the introduced entity ranking models requires to study the influence of the two interpolation factors λ and μ. Figure 1 shows the results of two experiments combined in one graph. In a first test we only retrieved articles and were interested in finding an appropriate setting of λ for calculating a personalized weighted indegree $PwIDG$. The lower two curves in the figure represent the outcome of this experiment. They show that λ needs to be set close to 1, which clearly points out the importance of the entity's own Wikipedia entry. On the other hand, combination with the scores of other articles mentioning the entity clearly improves the retrieval quality. For the second test, we kept λ fixed at its best value, now varying the setting of μ, displayed in the two upper curves of the graph. While the mean average precision improves slightly, the precision on top of the retrieved list P@5 shows a clear maximum when article and paragraph scores are considered equally with a setting of $\mu = 0.55$. The independence of λ and μ assumed by the testing procedure might be not adequate and the fine-tuning on the data set can lead to unrealistic results, but the combination model achieves improvements even without finding the optimal parameter setting. Table 1 shows the outcome of all introduced ranking methods. Besides the INEX collection, we also tested the proposed entity ranking model on TREC's expert finding task from 2007. Since the TREC data does not provide a text description of each expert entity comparable to the Wikipedia entries, we could only evaluate the proposal of combining document and paragraph level retrieval scores. Using the same parameter setting of $\mu = 0.55$, we show that the found improvements are not a matter of overfitting. Our combination model yields a considerable improvement on MAP and a slighter one for P@5.

	INEX 07		TREC 07	
	MAP	P@5	MAP	P@5
baseline	0.291	0.343		
wIDG	0.267	0.330	0.352	0.212
$PwIDG$ $\lambda = 0.95$	0.309	0.370		
$PwIDG^*$ $\lambda = 0.95, \mu = 0.55$	0.321	0.409	0.382	0.220

Table 1: Retrieval quality overview

4. CONCLUSIONS

We demonstrated on the INEX entity ranking testset the advantage of combining the direct Wikipedia article score with the propagated scores from text fragments of different sizes. The outcome demonstrates the potential of our combination model, which beats the best performing INEX run [4] on the entity ranking task. The improving effect of using article and paragraph scores in a joined propagation model could be confirmed on a different testset and collection.

5. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06*, pages 43–50, New York, NY, USA, 2006.

[2] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. PFTijah: text search in an XML database system. In *Proceedings of the Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17, 2006.

[3] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07*, pages 731–740, New York, NY, USA, 2007.

[4] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In *INEX '07*, 2008.

[5] J. Zhu, D. Song, S. Ruger, M. Eisenstadt, and E. Motta. The open university at TREC 2006 enterprise track expert search task. In *TREC '06*, 2006.