Using Wikipedia's Category Structure for Entity Search

Rianne Kaptein[1] Jaap Kamps[2]

[1] TNO, Delft, The Netherlands*
[2] University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

In this paper we investigate how the category structure of Wikipedia can be exploited for Entity Ranking. In the last decade, the Web has not only grown in size, but also changed its character, due to collaborative content creation and an increasing amount of structure. Current Search Engines find Web pages rather than information or knowledge, and leave it to the searchers to locate the sought information within the Web page. A considerable fraction of Web searches contains named entities. We focus on how the Wikipedia structure can help rank relevant entities directly in response to a search request, rather than retrieve an unorganized list of Web pages with relevant but also potentially redundant information about these entities. Our results demonstrate the benefits of using topical and link structure over the use of shallow statistics. This paper is a compressed version of [1].

1. INTRODUCTION

Searchers looking for entities are better served by presenting a ranked list of entities directly, rather than an unorganized list of Web pages with relevant but also potentially redundant information about these entities. The goal of the entity ranking task is to return entities instead of documents or text as are returned for most common search tasks. Entities can be for example persons, organizations, books, or movies.

A resource that is large enough to generate meaningful statistics, and contains interpretable semantic structure is Wikipedia. The nature and structure of Wikipedia presents new opportunities to solve problems that were thought to require deep understanding capabilities and where bottlenecks such as high cost and scalability where applicable in the past. Combining the benefits of the structured information and the large scale of Wikipedia, creating the opportunity to use probabilistic methods, we can now efficiently process all of the information contained in Wikipedia.

In this paper is motivated by the following main research question: *How can we exploit the structure of Wikipedia to retrieve entities?* We start by looking at how we can retrieve entities inside Wikipedia, which is also the task in the INEX entity ranking track. INEX[1](Initiative for the Evaluation of XML retrieval) is an information retrieval evaluation forum that provides an IR test collection to evaluate the task of entity ranking using Wikipedia as its document collection. Our first research question is: *How can we exploit category and link information for entity ranking in Wikipedia?*

Since a requirement for a relevant result in entity ranking is to retrieve the correct entity type, category information is of great importance for entity ranking. Category information can also be regarded in a more general fashion, as extra context for your query, which could be exploited for ad hoc retrieval. Our second research question is therefore: *How can we use entity ranking techniques that use category information for ad hoc retrieval?*

Since usually ad hoc queries do not have target categories assigned to them, and providing target categories for entity ranking is an extra burden for users, we also examine ways to assign target categories to queries. Our third research question is: *How can we automatically assign target categories to ad hoc and entity ranking queries?*

2. RETRIEVAL MODEL

In this section we describe our retrieval model, how we use category information for entity ranking, how we combine these sources of information, and how we assign categories to query topics automatically.

Exploiting Category Information Although for each entity ranking topic one or a few target categories are provided, relevant entities are not necessarily associated with these provided target categories. Relevant entities can also be associated with descendants of the target category or other similar categories. Therefore, simply filtering on the target categories is not sufficient. multiple categories, not all categories of an answer entity will be similar to the target category. We calculate for each target category the distances to the categories assigned to the answer entity. To calculate the distance between two categories, we tried three options. The first option (*binary distance*) is a very simple method: the distance is 0 if two categories are the same, and 1 otherwise. The second option (*contents distance*) calculates distances according to the contents of each category, and the third option (*title distance*) calculates a distance according to the category titles. We use KL-divergence to calculate distances between categories, and calculate a category score that is high when the distance is small.

Combining information Finally, we have to combine our different sources of information. Our first source of information is a standard language model for retrieval, which calculates the probabilities of occurrence of the query terms

*Work done while at the University of Amsterdam.
[1]https://inex.mmci.uni-saarland.de/

DIR 2013, April 26, 2013, Delft, The Netherlands.

Table 1: 2007 ER Topics using Category Information

Category representation	Weight	MAP	P10
Baseline		0.1840	0.1920
Binary	0.1	0.2145^{-}	0.1880^{-}
Contents	0.1	0.2481$^{\circledcirc}$	0.2320°
Title	0.1	0.2509°	0.2360°
Contents	0.05	**0.2618**$^{\circledcirc}$	**0.2480**$^{\circledcirc}$
Title	0.05		

in a document. This standard language model also serves as our baseline retrieval model. We explore two possibilities to combine information. First, we make a linear combination of the document, link and category score. All scores and probabilities are calculated in the log space, and then a weighted addition is made.

Alternatively, we can use a two step model. Relevance propagation takes as input initial probabilities as calculated by the baseline document model score. Instead of the baseline probability, we can use the scores of the run that combines the baseline score with the category information.

Target Category Assignment Besides using the target categories provided with the entity ranking query topics, we also look at the possibility of automatically assigning target categories to entity ranking and ad hoc topics. From our baseline run we take the top N results, and look at the T most frequently occurring categories belonging to these documents, while requiring categories to occur at least twice. These categories are assigned as target categories to the query topic.

3. EXPERIMENTS

In this section we describe our experiments with entity ranking and ad hoc retrieval in Wikipedia.

Experimental Set-up We experiment with two different tasks. First of all we experiment with the entity ranking task as defined by INEX. We will make runs on the topic sets from 2007 to 2009. Secondly, we experiment with ad hoc retrieval using category information on the ad hoc topic sets from 2007 and compare automatic and manual category assignment for ad hoc and entity ranking topics.

Entity Ranking Results The results on the 2007 entity ranking topic set (ER07b, 19 topics) are summarized in Table 1. The weight of the baseline score is 1.0 minus the weight of the category information. For all three distances, a weight of 0.1 gives the best results. In addition to these combinations, we also made a run that combines the original score, the contents distance and the title distance. When a single distance is used, the title distance gives the best results. The combination of contents and title distance gives the best results overall. For the 2008 and 2009 entity ranking topic sets (not shown here), also significant improvements are achieved when category information is used. Additional improvements to the approach are to rerank the top 2500 documents from the baseline retrieval run, instead of the top 500, which have been reranked for the 2007 runs. Normalizing the scores before combining shows improvements for the 2009 topics.

Ad Hoc Retrieval Results A selection of 19 topics in the ad hoc topic set (AH07a) was transformed into an additional entity ranking topics (set ER07a). There are 80 more judged ad hoc topics (set AH07b). Results for 2007 entity ranking and ad hoc topics expressed in MAP are summarized in Table 2, where "M" stands for manually assigned categories, and "A" for automatically assigned categories.

Table 2: Ad Hoc vs. Entity Ranking results in MAP

	Query	Category	Combi.	Best Score	
Set (M/A)	$\mu = 0.0$	$\mu = 1.0$	$\mu = 0.1$	μ	
ER07a M	0.2804	0.2547^{-}	0.3848$^{\bullet}$	0.2	0.4039$^{\bullet}$
ER07a A	0.2804	0.2671^{-}	0.3607$^{\circledcirc}$	0.1	0.3607$^{\circledcirc}$
ER07b M	0.1840	0.1231^{-}	0.2481$^{\circledcirc}$	0.1	0.2481$^{\circledcirc}$
ER07b A	0.1840	0.1779^{-}	0.2308$^{\circledcirc}$	0.2	0.2221°
AH07a M	0.3653	0.2067°	0.4308$^{\circledcirc}$	0.1	0.4308$^{\circledcirc}$
AH07b M	0.3031	0.1761$^{\bullet}$	0.3297$^{\circledcirc}$	0.05	0.3327$^{\bullet}$

From the four topic sets, the baseline scores of the ad hoc topic sets are higher. There is quite a big difference between the two entity ranking topic sets, where the topics derived from the ad hoc topics are easier than the genuine entity ranking topics. The entity ranking topics benefit greatly from using the category information with significant MAP increases of 44% and 35% for topic sets ER07a and ER07b respectively. When we use the category information for the ad hoc topics with manually assigned categories improvements are smaller than the improvements on the entity ranking topics, but still significant. Comparing manual and automatic assignments of target categories, manually assigned target categories perform somewhat better than the automatically assigned categories. However, for both topic sets using the automatically assigned categories leads to significant improvements over the baseline.

4. CONCLUSION

In this paper we have experimented with retrieving entities from Wikipedia exploiting its category structure. First, we examined whether Wikipedia category and link structure can be used to retrieve entities inside Wikipedia as is the goal of the INEX Entity Ranking task. Category information proves to be a highly effective source of information, leading to large and significant improvements in retrieval performance on all data sets. Secondly, we studied how we can use category information to retrieve documents for ad hoc retrieval topics in Wikipedia. Considering retrieval performance, also on ad hoc retrieval topics we achieved significantly better results by exploiting the category information. Finally, we examined whether we can automatically assign target categories to ad hoc and entity ranking queries. Guessed categories lead to performance improvements that are not as large as when the categories are assigned manually, but they are still significant. Our main conclusion is that the category structure of Wikipedia can be effectively exploited, in fact not only for entity ranking, but also for ad hoc retrieval, and with manually assigned as well as automatically assigned target categories.

Thanks This research was funded by NWO (# 612.066.513).

REFERENCES

[1] R. Kaptein and J. Kamps. Exploiting the Category Structure of Wikipedia for Entity Ranking, In *Artificial Intelligence*, Volume 194, January 2013, Pages 111-129.