Entity Ranking using Wikipedia as a Pivot

Rianne Kaptein[1] kaptein@uva.nl
Pavel Serdyukov[2] p.serdyukov@tudelft.nl
Arjen de Vries[2,3] arjen@acm.org
Jaap Kamps[1,4] kamps@uva.nl

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam, The Netherlands
[2] Delft University of Technology, The Netherlands
[3] Centrum Wiskunde & Informatica, The Netherlands
[4] ISLA, Informatics Institute, University of Amsterdam, The Netherlands

ABSTRACT

In this paper we investigate the task of Entity Ranking on the Web[1] Searchers looking for entities are arguably better served by presenting a ranked list of entities directly, rather than a list of web pages with relevant but also potentially redundant information about these entities. Since entities are represented by their web homepages, a naive approach to entity ranking is to use standard text retrieval. Our experimental results clearly demonstrate that text retrieval is effective at finding relevant pages, but performs poorly at finding entities. Our proposal is to use Wikipedia as a pivot for finding entities on the Web, allowing us to reduce the hard web entity ranking problem to easier problem of Wikipedia entity ranking. Wikipedia allows us to properly identify entities and some of their characteristics, and Wikipedia's elaborate category structure allows us to get a handle on the entity's type.

1. INTRODUCTION

Just like in document retrieval, in entity ranking the document should contain topically relevant information. However, it differs from document retrieval on at least three points: i) returned documents have to represent an entity, ii) this entity should belong to a specified entity type, and iii) to create a diverse result list an entity should only be returned once. The main goal of this paper is to demonstrate how the difficult problem of web entity ranking can often be reduced to the easier task of entity ranking in Wikipedia.

Our proposal is to exploit Wikipedia as a pivot for entity ranking. For entity types with a clear representation on the web, like living persons, organisations, products, movies, we will show that Wikipedia pages contain enough evidence to reliably find the corresponding web page of the entity. For entity types that do not have a clear representation on the web, returning Wikipedia pages is in itself a good alternative. So, to rank (web) entities given a query we take the following steps:

1. Associate target entity types with the query
2. Rank Wikipedia pages according to their similarity with the query and target entity types
3. Find web entities corresponding to the Wikipedia entities

We evaluate our approach using the entity ranking test collection created in the TREC 2009 Entity Ranking track [1].

2. ENTITY RANKING ON THE WEB

To investigate whether the hard problem of web entity ranking can be in principle reduced to the easier problem of Wikipedia entity ranking we look at the coverage of relevant TREC entities in Wikipedia. We find that the overwhelming majority of relevant entities (160 out of 198) of the TREC 2009 Entity ranking track are represented in Wikipedia, and that 85% of the topics have at least one relevant Wikipedia page. We also find that with high precision and coverage relevant web entities corresponding to the Wikipedia entities can be found using Wikipedia's "external links", and that especially the first external link is a strong indicator for primary homepages.

Furthermore we examine the value of entity type information for entity retrieval in Wikipedia. We find that entity types are valuable retrieval cues. Automatically assigned entity types are effective, but less so than manually assigned types. We can exploit the structure of Wikipedia to significantly improve entity ranking effectiveness.

In the remainder of this section we examine our research question: Can we improve web entity retrieval by using Wikipedia as a pivot? We compare our entity ranking approach of using Wikipedia as a pivot to the baseline of full-text retrieval.

We experiment with three approaches for finding webpages associated with Wikipedia pages:
1. External links: Follow the links in the External links section of the Wikipedia page.
2. Anchor text: Take the Wikipedia page title as query, and retrieve pages from the anchor text index. A length prior is used here.
3. Combined: Since not all Wikipedia pages have external links, and not all external links of Wikipedia pages are part of the Clueweb category B collection, we can not retrieve webpages for all Wikipedia pages. In case less than 3 webpages are found, we fill up the results to 3 pages using the top pages retrieved using anchor text.

2.1 Experimental Setup

In this experimental section we discuss experiments with the TREC Entity Ranking topics. We use the Indri search engine. We have created separate indexes for the Wikipedia part and the Web part of the Clueweb Category B. Besides a full text index we have also created an anchor text index. On all indexes we applied the Krovetz stemmer, and we generated a length prior. All runs are created with a language model using Jelinek-Mercer smoothing with a collection λ of 0.15.

[1]This paper is a compressed version of Kaptein, R., Serdyukov, P., Kamps, J., and de Vries, A. P. (2010). Entity ranking using Wikipedia as a pivot. In Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010), pages 69-78. ACM Press, New York USA.

DIR'2011, February 4, 2011, Amsterdam.

Table 1: TREC Web Entity Ranking Results

Run	Full Text	Wikipedia Link	Wikipedia Cat+Link
Rel. WP	**73**	**73** $^{-}$	57°
Rel. HP	**244**	69$^{\bullet}$	70$^{\bullet}$
Rel. All	**316**	134$^{\bullet}$	121$^{\bullet}$
NDCG Rel. WP	**0.2119**	**0.2119** $^{-}$	0.1959 $^{-}$
NDCG Rel. HP	**0.1919**	0.0820$^{\bullet}$	0.0830$^{\bullet}$
NDCG Rel. All	**0.2394**	0.1429$^{\bullet}$	0.1542$^{\bullet}$
Primary WP	78	78 $^{-}$	**96**$^{\bullet}$
Primary HP	6	29°	**34**$^{\circ}$
Primary All	86	107°	**130**$^{\bullet}$
P10 pr. WP	0.1200	0.1200 $^{-}$	**0.1700**$^{\circ}$
P10 pr. HP	0.0050	0.0300°	**0.0400**$^{\bullet}$
P10 pr. All	0.1200	0.1300 $^{-}$	**0.1850**$^{\bullet}$
NDCG pr. WP	0.1184	0.1184 $^{-}$	**0.1604**$^{\bullet}$
NDCG pr. HP	0.0080	0.0292 $^{-}$	**0.0445**$^{\circ}$
NDCG pr. All	0.1041	0.1292 $^{-}$	**0.1610**$^{\bullet}$

Significance of increase or decrease over full text according to t-test, one-tailed, at significance levels 0.05($^{\circ}$), and 0.01($^{\bullet}$).

Our baseline run uses standard document retrieval on a full text index. The result format of the TREC entity ranking runs differs from the general TREC style runs. One result consists of one Wikipedia page, and can contain up to three webpages from the non-Wikipedia part of the collection. The pages in one result are supposed to be pages representing the same entity.

For our baseline runs we do not know which pages are representing the same entity. In these runs we put one homepage and one Wikipedia page in each result according to their ranks, they do not necessarily represent the same entity. The Wikipedia based runs contain up to three homepages, all on the same entity. When a result contains more than one primary page, it is counted as only one primary page, or rather entity found.

2.2 Experimental Results

Recall from the above that the ultimate goal of web entity ranking is to find the homepages of the entities (called primary homepages). There are 167 primary homepages in total (an average of 8.35 per topic) with 14 out of the 20 topics having less than 10 primary homepages. In addition, the goal is to find an entity's Wikipedia page (called a primary Wikipedia page). There are in total 172 primary Wikipedia pages (an average of 8.6 per topic) with 13 out of the 20 topics having less than 10 primary Wikipedia entities.

The results for the TREC Entity Ranking track are given in Table 1. Our baseline is full text retrieval, which works well (NDCG 0.2394) for finding relevant pages. It does however not work well for finding primary Wikipedia pages (NDCG 0.1184). More importantly, it fails miserably for finding the primary homepages: only 6 out of 167 are found, resulting in a NDCG of 0.0080 and a P10 of 0.0050. Full text retrieval is excellent at finding relevant information, but it is a poor strategy for finding web entities.

We now look at the effectiveness of our Wikipedia-as-a-pivot runs. The Wikipedia runs in this table use the external links to find homepages. The second column is based on the baseline Wikipedia run, the third column is based on the run that uses the manual categories that proved effective for entity ranking on Wikipedia. Considering primary pages, we find more primary Wikipedia pages, translating into a significant improvement of retrieval effectiveness (up to a P10 of 0.1700, and a NDCG of 0.1604). Will this also translate into finding more primary homepages? The first run is a straightforward run on the Wikipedia part of ClueWeb, using the external links to the Web (if present). Recall that we established that primary pages linked from relevant Wikipedia pages have a high precision. This strategy finds 29 primary homepages (so 11 more than the baseline) and improves retrieval effectiveness to an NDCG of 0.0292, and a P10 of 0.0300. The second run using the Wikipedia category information improves significantly to find 34 primary homepages with a NDCG of 0.0445 and a P10 of 0.0400.

Table 2: TREC Homepage Finding Results

Run	Cat+Link	Anchor	Comb.
Rel. HP	70	127	**137**
Rel. All	121	178	**188**
NDCG Rel. HP	0.0830	0.0890	**0.1142**
NDCG Rel. All	0.1542	0.1469	**0.1605**
Primary HP	34	29	**56**
Primary All	130	125	**152**
P10 pr. HP	0.0400	0.0450	**0.0550**
P10 pr. All	**0.1850**	0.1750	**0.1850**
NDCG pr. HP	0.0445	0.0293	**0.0477**
NDCG pr. All	0.1041	0.1472	**0.1610**

Recall again that the external links have high precision but low recall. We try to find additional links between retrieved Wikipedia pages and the homepages by querying the anchor text index with the name of the found Wikipedia entity. This has no effect on the found Wikipedia entities, so we only discuss the primary homepages as presented in Table 2. Ignoring the existing external links, searching for the Wikipedia entities in the anchor text leads to 29 primary homepages. The combined run, supplementing the existing external links in Wikipedia with the automatically generated links, finds a total of 56 primary homepages. For homepages this improves the P10 over the baseline to 0.0550, and NDCG to 0.0447.

3. CONCLUSION

This paper investigates the problem of entity retrieval on the Web. Our main findings are the following. Our first finding is that, in principle, the problem of web entity ranking can be reduced to Wikipedia entity ranking. We found that the majority of entity ranking topics in our test collections can be answered using Wikipedia, and that with high precision relevant web entities corresponding to the Wikipedia entities can be found using Wikipedia's 'external links'. Our second finding is that we can exploit the structure of Wikipedia to improve entity ranking effectiveness. Entity types are valuable retrieval cues in Wikipedia. Automatically assigned entity types are effective, and almost as good as manually assigned types. Our third finding is that web entity retrieval can be significantly improved by using Wikipedia as a pivot. Both Wikipedia's external links and the enriched Wikipedia entities with additional links to homepages are significantly better at finding primary web homepages than standard text retrieval.

REFERENCES

[1] K. Balog, A. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Notebook*. National Institute for Standards and Technology, 2009.