Linking Wikipedia to the Web

Rianne Kaptein[1] Pavel Serdyukov[2] Jaap Kamps[1]

[1] University of Amsterdam [2] Delft University of Technology
The Netherlands
kaptein@uva.nl p.serdyukov@tudelft.nl kamps@uva.nl

ABSTRACT

We investigate the task of finding links from Wikipedia pages to external web pages. Such external links significantly extend the information in Wikipedia with information from the Web at large, while retaining the encyclopedic organization of Wikipedia. We use a language modeling approach to create a full-text and anchor text runs, and experiment with different document priors. In addition we explore whether social bookmarking site Delicious can be exploited to further improve our performance. We have constructed a test collection of 53 topics, which are Wikipedia pages on different entities. Our findings are that the anchor text index is a very effective method to retrieve home pages. Url class and anchor text length priors and their combination leads to the best results. Using Delicious on its own does not lead to very good results, but it does contain valuable information. Combining the best anchor text run and the Delicious run leads to further improvements.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]

General Terms: Experimentation, Measurement, Performance.

Keywords: Link Detection, Entity Search, Wikipedia.

1. INTRODUCTION

Wikipedia is a natural starting point for information on almost any topic. As a result, Wikipedia is one of the top ranked results for all queries matching an article's title. But where to go if you want to know more? Can we point searchers directly to other relevant web pages? For this purpose, many Wikipedia pages contain 'External Links' to web pages. According to the guidelines,[1] the links in the External Links section should link to sites that contain neutral and accurate material that is relevant to an encyclopedic understanding of the subject. For example, pages about entities should link to its official home page, and pages about media to a site hosting a copy of the work.

However, only some 45% of all Wikipedia pages have an 'External links' section. Hence, our research question is:

- ▷ Can we automatically find external links for Wikipedia pages?

To evaluate how well we can find external links for Wikipedia pages, we construct a test collection by removing the currently existing links in Wikipedia, and using these links as our ground truth. This is similar to the INEX Link-the-Wiki task [2] where the task consists of finding links between Wikipedia pages. Our task is to find links from Wikipedia pages to external web pages. We use the Clueweb category B, consisting of 50 million English web pages as our test collection to find the external web pages.

To validate that Wikipedia's external links indeed correspond to official home pages, we use the assessments of the 2009 TREC entity ranking task. These assessments contain 60 relevant Wikipedia pages with at least one linked website in the Clueweb collection. When we consider the entity as a query, and urls found in 'External links' as ranked pages a Mean Reciprocal Rank of 0.768 is attained for finding the home pages. That is, there is a high level of agreement between the External Links in Wikipedia and the independent judgment of a TREC assessor on what constitutes the home page for an entity.

2. EXTERNAL LINK DETECTION

2.1 Task and Test Collection

Our task is defined as follows: Given a topic, i.e. a Wikipedia page, return the external web pages which should be linked in the 'External Links' section. We have created a topic set by reusing relevant entities found in the TREC Entity Ranking task. The topic set contains 53 topics with 84 relevant home pages. A topic can have more than one relevant home page, because the Clueweb collection contains duplicate pages, i.e. pages with the same normalized url. We match the urls of the existing external links on the Wikipedia pages with the urls in the Clueweb collection. External links on entity pages are split into two parts, the first external link is a home page, the other links are usually informational pages. In our experiments we only use the home pages.

2.2 Link Detection Approaches

We experiment with three approaches. First, our baseline approach is a language model with a full-text index. Secondly, we make an anchor text index, which has proved to work well for home page finding [1]. We experiment with different document priors for both indexes. We construct priors for the document length, anchor text length, and the url class [3]. To determine the url class, we first apply a number of url normalization rules, such as removing trailing slashes, and removing suffixes like 'index.html'. Since we have no training data, we cannot estimate prior probabilities of url classes based on the distribution of home pages in the training collection. Instead we use only two url classes: root pages (a domain name not followed by any directories) receive a prior probability a 100 times larger than non-root pages, which is a conservative prior compared to the previous work [3]. Our third approach exploits information of social bookmarking site *Delicious*. Delicious ranks search results by relevance, taking into account bookmark titles, notes, and tags, among other things. We send a search re-

[1] http://en.wikipedia.org/wiki/Wikipedia:External_links

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Table 1: Language Modeling Results

	Full-text		Anchor	
Prior	MRR	$Suc@5$	MRR	$Suc@5$
None	0.0385	0.0364	0.5865	0.7091
Doc. length	0.0085°	0.0000	0.4178$^{\bullet}$	0.5455$^{\bullet}$
Anchor length	0.0853°	0.1636	0.6131	0.6909
Url class	0.2348$^{\bullet}$	0.2727$^{\bullet}$	0.6545	0.7273
Anch.length + Url	**0.2555**$^{\bullet}$	**0.2909**$^{\bullet}$	**0.6774**$^{\circ}$	**0.7636**

Significance of increase or decrease over “None” according to t-test, one-tailed, at significance levels 0.05($^{\circ}$), 0.01(◍), and 0.001($^{\bullet}$).

quest to the site and match the first 250 results with the urls in the Clueweb collection to create a ranking. As retrieval score we use the (inverted) ranks. To make combinations with our language model runs we normalize all scores using the Z-score and make a linear combination of the normalized scores.

For our experiments we use the Indri toolkit. We build two indexes: an anchor text and a full text index. Both indexes are stemmed with the Krovetz stemmer. We have created document priors for document length, anchor text length, and url class. For all our runs we apply Dirichlet document smoothing. To construct the query we always use the title of the Wikipedia page. We use Mean Reciprocal Rank (MRR) and Success at 5 ($Suc@5$) to evaluate our runs.

2.3 Link Detection Results

Results of our experiments using the language modeling approach are shown in Table 1. The anchor text index leads to much better results than the full-text index. Home pages often contain a lot of links, pictures, and animations, and not so much actual text, so it was to be expected that the anchor text index is more effective. For the same reason, applying a document length prior deteriorates the results: longer documents are not more likely to be a relevant home page.

The two other document priors do lead to improvements. The full-text index run has much more room for improvement, and indeed the priors lead to a major increase in performance, e.g. using the url class prior increases the MRR from 0.0385 to 0.2348. The improvements on the anchor text runs are smaller. The anchor text length prior does not affect the results much. A reason for this can be that the Dirichlet smoothing also takes into account the document length, which equals the anchor text length for the anchor text run. Despite its simplicity, the url class prior leads to significant improvements for both the full-text and the anchor text runs. Since we did not have training data available, we did not optimize the url class prior probabilities, but used a conservative prior on only two classes. Combining the full-text runs and the anchor text runs does not lead to improvements over the anchor text run. We experimented also with including different parts of the Wikipedia page in the query, such as the first sentence and the page categories, but none of these runs improved over using only the title of the page. By analyzing the failure cases, we identify three causes for not finding a relevant page: the external link on the Wikipedia page is not a home page, the identified home page is redirected or varies per country, and the Wikipedia title contains ambiguous words or acronyms.

Besides the internal evidence, we also looked for external evidence to find home pages. The results of the run using Delicious, and a combination with the best anchor text run can be found in Table 2. The Delicious run performs better than the full-text run, but not as good as the anchor text run. One disadvantage of the Delicious run is that it does not return results for all topics. Some topics with long queries do not return any results, other topics do return results, but none of the results exists in the Clueweb collection. For 49 topics Delicious returns at least one result, for 41 topics at least one Clueweb page is returned. Around half of all returned results are part of the Clueweb collection. When we combine the Delicious run with the best anchor text run we do get better results, so Delicious is a useful source of evidence. Most of the weight (0.9) in the combination is on the anchor text run though. The Delicious run retrieves 68 relevant home pages, which is more than the 58 pages the anchor text run retrieves. The Delicious run however contains more duplicate pages, because it searches for all pages matching the normalized url retrieved by searching Delicious. In the combination of runs, pages found both by Delicious and by the anchor text run, end up high in the ranking.

Table 2: Delicious Results

Run	MRR	$Suc@5$
Delicious	0.3597	0.4000
Comb	**0.7119**	**0.7818**
Anchor	0.6774	0.7636

When we compare our results to previous home page finding work, we can make the following remarks. Most differences can be attributed to the test collections. Clueweb is crawled in 2009, and in comparison to older test collections the full-text index performs much worse. Modern home pages contain less relevant text and more pictures, photos and animations, making the full-text index less informative. The anchor text index on the other hand, performs better than ever before. The Clueweb collection is larger than previous collections, and has a higher link density, so there is more anchor text available for more pages.

3. CONCLUSION

In this paper, we investigate the task of finding external links for Wikipedia pages. We have constructed a test collection of topics about different entities, with their corresponding relevant home pages. Two language modeling approaches, one based on a full-text index, and one based on an anchor text index have been investigated. In addition a run based on the Delicious bookmarking site is made. All anchor text runs perform much better than the full-text index runs. Useful document priors are the anchor text length and the url class. Delicious on itself does not perform so well, but it is a useful addition when it is combined with an anchor text run. We can conclude our system is effective at predicting the external links for Wikipedia pages.

Acknowledgments The test collection for Wikipedia external link detection is available at http://staff.science.uva.nl/~kamps/effort/data. This research was supported by the Netherlands Organization for Scientific Research (NWO, under project # 612.066.513). We thank Arjen de Vries for useful discussions on related issues.

REFERENCES

[1] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *SIGIR ’01*, pages 250–257, 2001.

[2] D. W. Huang, Y. Xu, A. Trotman, and S. Geva. Overview of INEX 2007 link the wiki track. In *Focused Access to XML Documents*, pages 373–387, 2008.

[3] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR ’02*, pages 27–34, 2002.