Finding Entities in Wikipedia Using Links and Categories

Rianne Kaptein[1] and Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

Abstract. In this paper we describe our participation in the INEX Entity Ranking track. We explored the relations between Wikipedia pages, categories and links. Our approach is to exploit both category and link information. Category information is used by calculating distances between document categories and target categories. Link information is used for relevance propagation and in the form of a document link prior. Both sources of information have value, but using category information leads to the biggest improvements.

1 Introduction

In the entity ranking track, our aim is to explore the relations and dependencies between Wikipedia pages, categories and links. For the entity ranking task we have looked at some approaches that proved to be successful in previous entity ranking and ad hoc tracks. In these tracks it has been shown that link information can be useful. Kamps and Koolen [2] use link evidence as document priors, where a weighted combination of the number of incoming links from the entire collection and the number of incoming links from the retrieved results for one topic is used. Tsikrika et al. [4] use random walks to model multi-step relevance propagation from entities to their linked entities. For the entity ranking track specifically also the category assignments of entities can be used. Vercoustre et al. [5] use the Wikipedia categories by defining similarity functions between the categories of retrieved entities and the target categories. The similarity scores are estimated using lexical similarity of category names. We combined and extended the aforementioned approaches.

2 Model

In this section we describe how we use category information for entity ranking and list completion, how we exploit link information and finally how we combine these sources of information.

Category information. Although for each 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

S. Geva, J. Kamps, and A. Trotman (Eds.): INEX 2008, LNCS 5631, pp. 273–279, 2009.
© Springer-Verlag Berlin Heidelberg 2009

or other similar categories. Therefore, simply filtering on the target categories is not sufficient. Also, since Wikipedia pages are usually assigned to 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. For the title and contents distance, we need to calculate the probability of a term occurring in a category. To avoid a division by zero, we smooth the probabilities of a term occurring in a category with the background collection:

$$P(t_1, ..., t_n|C) = \sum_{i=1}^{n} \lambda P(t_i|C) + (1 - \lambda)P(t_i|D)$$

where C, the category, consists either of the category title to calculate title distance, or of the concatenated text of all pages belonging to that category to calculate contents distance. D is the entire wikipedia document collection, which is used to estimate background probabilities. We estimate $P(t|C)$ with a parsimonious model [1] that uses an iterative EM algorithm as follows:

$$\text{E-step:} \qquad e_t = tf_{t,C} \cdot \frac{\alpha P(t|C)}{\alpha P(t|C) + (1 - \alpha)P(t|D)}$$

$$\text{M-step:} \qquad P(t|C) = \frac{e_t}{\sum_t e_t}, \text{ i.e. normalize the model}$$

The initial probability $P(t|C)$ is estimated using maximum likelihood estimation. We use KL-divergence to calculate distances, and calculate a category score that is high when the distance is small as follows:

$$S_{cat}(C_d|C_t) = -D_{KL}(C_d|C_t) = -\sum_{t \in D} \left(P(t|C_t) * \log \left(\frac{P(t|C_t)}{P(t|C_d)} \right) \right)$$

where d is a document, i.e. an answer entity, C_t is a target category and C_d a category assigned to a document. The score for an answer entity in relation to a target category $S(d|C_t)$ is the highest score, or shortest distance from any of the document categories to the target category.

In contrast to Vercoustre et al. [5], where a ratio of common categories between the categories associated with an answer entity and the provided target categories is calculated, we take for each target category only the shortest distance from any answer entity category to a target category. So if one of the categories of the document is exactly the target category, the distance and also the category score for that target category is 0, no matter what other categories are assigned to the document. Finally, the score for an answer entity in relation to a query topic $S(d|QT)$ is the sum of the scores of all target categories:

$$S_{cat}(d|QT) = \sum_{C_t \in QT} \operatorname*{argmax}_{C_d \in d} S(C_d|C_t)$$

Besides the entity ranking task, the second task in the entity ranking track is list completion. Instead of the target category, for each topic a few relevant examples entities

are given. We treat all categories assigned to the example entities as target categories. Our approach for using the category information is the same as before. But to get the final score of an article in relation to a topic, we use two variants. The first one is:

$$S_{Sum}(d|QT) = \sum_{ex \in QT} \sum_{C_{ex} \in ex} \operatorname*{argmax}_{C_d \in d} S_{cat}(C_d|C_{ex})$$

In the second variant $S_{Max}(d|QT)$, instead of summing the score of each example category, we only take the maximum score i.e. shortest distance for all example categories of the entity examples to one of the categories of the document. Furthermore, we apply explicit relevance feedback based on the text of the example entities to expand the query.

Link information. We implement two options to use the link information: relevance propagation and document link degree prior. For the document link degree prior we use the same approach as in [2]. The prior for a document d is:

$$P_{Link}(d) = 1 + \frac{Indegree_{Local}(d)}{1 + Indegree_{Global}(d)}$$

The local indegree is equal to the number of incoming links from within the top ranked documents retrieved for one topic. The global indegree is equal to the number of incoming links from the entire collection.

The second use of link information is through relevance propagation from initially retrieved entities, as was done last year in the entity ranking track by Tsikrika et al. [4].

$$P_0(d) = P(q|d)$$
$$P_i(d) = P(q|d)P_{i-1}(d) + \sum_{d' \to d}(1 - P(q|d'))P(d|d')P_{i-1}(d')$$

Probabilities $P(d|d')$ are uniformly distributed among all outgoing links from the document. Documents are ranked using a weighted sum of probabilities at different steps:

$$P(d) = \mu_0 P_0(d) + (1 - \mu_0) \sum_{i=1}^{K} \mu_i P_i(d)$$

For K we take a value of 3, which was found to be the optimal value last year. We try different values of μ_0 and distribute $\mu_1 ... \mu_K$ uniformly, i.e. $\mu_1 ... \mu_K = 1/3$.

Combining information. Finally, we have to combine our different sources of information. We start with our baseline model which is a standard language model. We have two possibilities to combine information. We can make a linear combination of the probabilities 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. Instead of the baseline probability, we can use the scores of the run that combines the baseline score with the category information. Similarly, for the link degree prior we can use the top results of the baseline combined with the category information instead of the baseline ranking.

3 Experiments

In this section we describe our experimental results on the training and the test data.

3.1 Training Results

For our training data we use the 25 genuine entity ranking test topics that were developed for the 2007 entity ranking track. For our baseline run and to get initial probabilities we use the language modeling approach with Jelinek-Mercer smoothing, Porter stemming and pseudo relevance feedback as implemented in Indri [3] to estimate $P(d|q)$. We tried different values for the smoothing λ. We found $\lambda = 0.1$ gives the best results, with a MAP of 0.1840 and a P10 of 0.1920. For the document link degree prior we have to set two parameters: the number of top documents to use, and the weight of the document prior. For the number of top documents to use, we try 50, 100, 500 and 1,000 documents. For the weight of the prior we try all values from 0 to 1 with steps of 0.1. Only weights that give the best MAP and P10 are shown in Table 1.[1]

Table 1. Document link degree prior results

# docs	Weight	MAP	P10
Baseline		0.1840	0.1920
50	0.6	0.1898 $^{-}$	**0.2040** $^{-}$
50	0.5	0.1876 $^{-}$	0.2000 $^{-}$
100	0.7	0.1747 $^{-}$	0.2000 $^{-}$
100	0.3	0.1909 $^{-}$	0.1920 $^{-}$
500	0.5	**0.1982** $^{\circ}$	0.2000 $^{-}$
500	0.3	0.1915 $^{-}$	**0.2040** $^{\circ}$
1,000	0.5	0.1965 $^{-}$	0.1960 $^{-}$
1,000	0.4	0.1965 $^{\circ}$	0.2000 $^{-}$

Table 2. Category distances results

Dist.	Weight	MAP	P10
Binary	0.1	0.2145 $^{-}$	0.1880 $^{-}$
Cont.	0.1	0.2481 $^{\circledcirc}$	0.2320 $^{\circ}$
Title	0.1	0.2509 $^{\circ}$	0.2360 $^{\circ}$
Cont. Title	0.05 0.05	**0.2618** $^{\circledcirc}$	**0.2480** $^{\circledcirc}$

The results of using category information are summarized in Table 2. 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.

In our next experiment we combine all information we have, the baseline score, the category and the link information. Firstly, we combine all scores by making a linear combination of the scores and probabilities (shown in Table 3). Secondly, we combine the different sources of information by using the two step model (see Table 4). Link information is mostly useful to improve early precision, depending on the desired results we can tune the parameters to get optimal P10, or optimal MAP. Relevance propagation performs better than the document link degree prior in both combinations.

[1] Significance of increase over the baseline according to the t-test, one-tailed, at significance levels 0.05 ($^{\circ}$), 0.01 ($^{\circledcirc}$), and 0.001 ($^{\bullet}$).

Table 3. Results linear combination

Link Info	Weight	MAP	P10
Prior	0.3	0.2682$^{\circledcirc}$	0.2640$^{\circledcirc}$
Prop.	0.1	**0.2777**$^{\circledcirc}$	**0.2720**$^{\circledcirc}$

Table 4. Results two step model

Link info	Weight	MAP	P10
Prior	0.5	0.2526$^{\circledcirc}$	0.2600$^{\circledcirc}$
Prop.	0.2	0.2588$^{\circledcirc}$	**0.2960** $^{\bullet}$
Prop.	0.1	**0.2767**$^{\circledcirc}$	0.2720$^{\circledcirc}$

For the list completion task, we use the examples for relevance feedback. To evaluate the list completion results, example entities are removed from our ranking. Applying explicit and pseudo relevance feedback leads to the results given in Table 5. Additional

Table 5. Feedback results

RF	PRF	MAP	P10
No	No	0.1409	0.1240
Yes	No	**0.1611**	0.1600
Yes	Yes	0.1341	**0.1960**

Table 6. List Completion results

Dist.	Weight	$S(A\|QT)$	C_t	MAP	P10
Baseline LC				0.1611	0.1600
Cont.	0.1	Sum	No	0.2385$^{\circledcirc}$	0.2520°
Cont.	0.9	Sum	Yes	0.2467$^{\bullet}$	0.2560°
Cont.	0.2	Max	No	0.1845 $^{-}$	0.2360 $^{-}$
Title	0.1	Sum	No	0.2524$^{\circledcirc}$	0.2640°
Title	0.9	Sum	Yes	**0.2641**$^{\bullet}$	**0.2760**$^{\circ}$
Title	0.5	Max	No	0.1618 $^{-}$	0.2080 $^{-}$
Cont. Title	0.05 0.05	Sum	No	0.2528$^{\bullet}$	0.2640°

pseudo relevance feedback after the explicit feedback, only improves early precision, and harms MAP. We take the run using only relevance feedback as our baseline for the list completion task.

When we look at the previous entity ranking task, the largest part of the improvement comes from using category information. So here we only experiment with using the category information, and not the link information. We have again the different category representations, content and category titles. Another variable here is how we combine the scores, either add up all the category scores $S_{Sum}(A|QT)$ or taking only the maximum score $S_{Max}(A|QT)$. Not part of the official task, we also make some runs that use not only the categories of the example entities, but also the target category(ies) provided with the query topic. In Table 6 we summarize some of the best results. The combination of contents and title distance, does not lead to an improvement over using only the title distance. The maximum score does not perform as well as the summed scores. We use all categories assigned to the entity examples as target categories, but some of these categories will not be relevant to the query topic introducing noise in the target categories. When the scores are summed, this noise is leveled out, but when only the maximum score is used it can be harmful. Comparing the list completion and the entity ranking task, the list completion task has a slightly lower baseline score, but the results of both tasks when category information is used, are very similar.

3.2 Test Results

The test data consists of 35 new entity ranking topics.We use the parameters that gave the best results on the training data, i.e. baseline with pseudo-relevance feedback and $\lambda = 0.1$, weights of contents and title category information is 0.1, or 0.05 and 0.05 in the combination. For the link prior we use the top 100 results, and the two-step model is used to combine the information. In Table 7 our results on the test topics are shown. Using the category information leads to an improvement of 100% over the baseline, the score is doubled! Even when we rerank the top 500 results retrieved by the baseline using only the category information, the result are significantly better than the baseline, with a MAP of 0.2405. Since the category information is so important, it is likely that relevant pages can be found outside the top 500. Indeed, when we rerank the top 2500, but still evaluating the top 500, our results improve up to a MAP of 0.3519. Furthermore, we found that on the test data doubling the weights of the category information leads to slightly better results. Similar to the training results, relevance propagation performs better than the link prior, and leads to small additional improvements over the runs using category information.

Table 7. Results on the 2008 test topics

# Results	Category info				Link info		MAP	P10
			Baseline				0.1586	0.2257
500	Title	0.1			No		0.3059$^\bullet$	0.4171$^\bullet$
	Title	0.2			No		0.3164$^\bullet$	0.4400$^\bullet$
			Cont.	0.1	No		0.3031$^\bullet$	0.4086$^\bullet$
			Cont.	0.2	No		0.3088$^\bullet$	0.4200$^\bullet$
	Title	0.05	Cont.	0.05	No		0.3167$^\bullet$	0.4343$^\bullet$
	Title	0.1	Cont.	0.1	No		0.3189$^\bullet$	0.4400$^\bullet$
	Title	0.05	Cont.	0.05	Prior	0.5	0.3196$^\bullet$	0.4371$^\bullet$
	Title	0.05	Cont.	0.05	Prop.	0.1	0.3324$^\bullet$	0.4543$^\bullet$
2500	Title	0.1			No		0.3368$^\bullet$	0.4343$^\bullet$
	Title	0.2			No		0.3504$^\bullet$	0.4514$^\bullet$
	Title	0.2			Prop.	0.1	0.3519$^\bullet$	0.4629$^\bullet$

For the list completion task we submitted two runs. These runs use only the category information, in the form of category titles and summing scores over categories. Curiously, the run using only the examples scores slightly better than the run that uses also the specified target categories, with MAP of 0.325 and 0.323 respectively.

4 Conclusion

We have presented our entity ranking approach where we use category and link information. Category information is the factor that proves to be most useful and we can do more than simply filtering on the target categories. Category information can both be extracted from the category titles and from the contents of the category. Link information

can also be used to improve results, especially early precision, but these improvements are smaller. In future research, we will look in more detail at the list completion task to derive more focused target categories from the example entities.

Acknowledgments. This research is funded by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513).

References

[1] Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. In: Proceedings SIGIR 2004, pp. 178–185. ACM Press, New York (2004)
[2] Kamps, J., Koolen, M.: The importance of link evidence in Wikipedia. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 270–282. Springer, Heidelberg (2008)
[3] Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis (2005)
[4] Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D., de Vries, A.P.: Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In: Focused Access to XML Documents, pp. 306–320 (2007)
[5] Vercoustre, A.M., Pehcevski, J., Thom, J.A.: Using wikipedia categories and links in entity ranking. In: Focused Access to XML Documents, pp. 321–335 (2007)