Finding Entities or Information Using Annotations

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

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

ABSTRACT

User-generated content often provides more information than just textual data, i.e. tags or annotations are added to label data, in discussions users can comment on the data, and links exist not only between pages, but also between users and annotations. In this paper we explore the use of annotations in the form of categories to find entities or information in Wikipedia. Main differences between entity ranking and ad hoc retrieval are the assessment criteria, and the provision of target categories for entity ranking topics. From analyzing the relevance assessment sets we can see that entity ranking results have more focused categories. The provided target category is however not always the most informative category. Furthermore we show that techniques for entity ranking can also be applied to ad hoc topics and automatically assigned target categories are good surrogates for manually assigned categories. Although using category information leads to larger improvements on entity ranking topics, significant improvements can also be achieved on ad hoc topics.

1. INTRODUCTION

Different networks of user-generated content are emerging on the web . These networks have in common that besides the actual contents meta-information can be added. For example tags or annotations are added to label data, users can add comments or start discussions. A large part of the information is linked, creating (social) networks of connected pages, users, tags etc. New IR models are needed that can exploit the particular structure of these networks. Unfortunately there are no appropriate, widely used test collections available that incorporate all the above mentioned types of information. In this paper we will focus on the use of annotations in the form of categories to find entities or information, allowing us to make use of the INEX Wikipedia test collection. Link information is also available for this collection, but it seems category information is more useful, especially for finding entities [11, 12], and will therefore be our focus here.

Wikipedia is a highly structured resource and includes an extensive collection of categories that are used to categorize Wikipedia pages. Our version of Wikipedia from 2006 includes already 150,000 categories. One of the challenges in exploiting the category information is that Wikipedia categories are created and assigned by different human editors, and are therefore not consistent. With 150,000 categories it is not a trivial task to assign the correct categories to a Wikipedia page. Some categories that should be assigned can be missing, and too general or too specific categories can be assigned to a page. A Wikipedia page is usually assigned to multiple categories. Wikipedia guidelines are to place articles only in the most specific categories they reasonably fit in. Peer reviewing is employed to improve the quality of pages and categorizations [13]. Categories are organized in a loose hierarchy. Some cycles of linked categories exist, but the guideline is to avoid them. Wikipedia takes some measures to prevent that similar categories coexist. If two similar categories are discovered, one category is chosen and whenever people try to use the other category, they are redirected to the chosen category. For example if someone tries to assign or find the Category:Authors, he is referred to the category:Writers. Also if some different spelled versions of the same category exists, category redirects are used, i.e. “Ageing” redirects to “Aging”, and “Living People” redirects to “Living people”. This system is in use not only for categories, but also for pages. Our test collection does not include these redirect pages, so we have to find some way to deal with these inconsistencies. Wikipedia’s category information can provide valuable information when searching for entities or information, but we have to take into account that the data is noisy.

Entity ranking is becoming a popular task in the information retrieval field. Evaluation initiatives like INEX and TREC are running entity ranking tracks. People or expert search is an example of an entity ranking task looking for one specific type of entity that is drawing quite some attention [2]. An issue in all entity ranking tasks is how to represent entities, returning only the name of the entity is not enough. People need to see some evidence, for example surrounding text, why this entity is relevant to their query. Since in this paper we restrict ourselves to entity ranking in Wikipedia, which is also is done in the INEX entity ranking track, we can find an easy way to represent entities. Namely, by representing them as Wikipedia pages, and by defining Wikipedia categories as entity types. Using this representation, the entity ranking task is similar to ad hoc retrieval since both tasks return Wikipedia pages in response to a keyword query. In the entity ranking track large improvements are achieved when the category information is exploited. In this paper we want to investigate the differences and similarities between entity ranking and ad hoc retrieval in Wikipedia to see if we can exploit category information for ad hoc retrieval as well.

In addition to the keyword query, entity ranking topics include one or a few target categories, that indicate the category that the search results should belong to. In the entity ranking task it is very well possible that relevant pages are not assigned to the target category. The category can either be a few steps away in the category graph, or similar categories can be relevant. Another issue is that some of the target categories provided in the entity ranking topics are redirected, e.g. “Movies”. These categories in principle should not contain any pages, and are not included in the category graph.

The entity ranking techniques that will be described in this paper, are able to deal with these issues.

Since a requirement for a relevant result in entity ranking is to retrieve the correct entity type, category information is of great importance. 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 first research question is therefore:

- Can we use entity ranking techniques that use category information for ad hoc retrieval?

Since usually ad hoc topics 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 second research question is:

- Can we automatically assign target categories to ad hoc and entity ranking topics?

This paper is organized as follows. In the next section we discuss related work. In section 3 we look at the differences between entity ranking and ad hoc retrieval. We analyze relevance assessment sets of different topic sets. Section 4 describes the models used to exploit category information, and how categories are assigned automatically to topics. In section 5 we describe our experiments. Finally in section 6 we draw our conclusion.

2. RELATED WORK

The social network site Delicious[1] is annotated by users and provides category information in the form of informal tags. Much of the early work on social annotations uses this resource, we will discuss two of these papers here. In Wu et al. [14] a global semantic model is statistically derived from the social annotations. The semantic model helps to disambiguate tags and groups synonymous tags together in concepts. Furthermore, the derived semantic can be used to search and discover semantically related web resources. Two aspects of social annotations that can benefit web search are explored in Bao et al. [3]. These aspects are: the annotations are usually good summaries of corresponding web pages and the count of annotations indicates the popularity of web pages. Their approach is able to find the latent semantic association between queries and annotations, and successfully measures the quality (popularity) of a web page from the web users perspective.

Since 2007 an entity ranking track is run in INEX [5]. Using category information is essential in this track, and almost all participants use the category information in some form. We mention here two interesting approaches. Our approach is closely related to Vercoustre et al. [12] where Wikipedia categories are used for entity ranking and ad hoc retrieval 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. To categorize the ad hoc topics, the query title is sent to an index of categories that has been created by using the names of the categories, and the names of all their attached entities. Their model works well for entity ranking, but when applied to ad hoc topics the entity ranking approach performs significantly worse than the basic full-text retrieval run. Another well performing approach in the entity ranking track of 2007 comes from Tsikrika et al. [11]. After relevance propagation, the entities that do not belong to a set of allowed categories are filtered out the result list. The best allowed category set included the target categories with their child categories up to the third level.

[1]http://delicious.com/

Besides in INEX, topical categories have been used in TREC for ad hoc retrieval. The topics in TREC 1 and 2 include a topical domain in the query topic descriptions, which can be used as topical context. It has been shown that these topical domains can successfully be used as query context for ad hoc retrieval [1]. In this paper the automatic and the manual assignment of topical categories is compared. Category models are created by using the relevant documents or the top 100 documents retrieved for the in-category queries. The top terms in the category models are used to expand the query. Automatic query classification is done by calculating KL- divergence scores. Although the accuracy of the automatic query classification is low, the effectiveness of retrieval is only slightly lower than when the query topic category is assigned manually.

The search engine ESTER combines full-text and ontology search [4]. ESTER is applied to the English Wikipedia, combined with the YAGO ontology, which contains about 2.5 million facts and was obtained by a combination of Wikipedias category informations with the WordNet hierarchy. The interactive search interface suggests to the user possible semantic interpretations of his/her query, thereby blending entity ranking and ad hoc retrieval.

3. ENTITY RANKING VS. AD HOC

The difference between entity ranking and ad hoc retrieval in general is that instead of searching for relevant text, you are searching for relevant entities. Entities can be of different types, a popular type of entity ranking is people search, other entity types can be movies, books, cities, etc. One of the difficulties in entity ranking is how to represent entities. Some supporting evidence in addition to the entity id or name is needed to confirm that an entity is relevant. The INEX entity ranking track uses Wikipedia pages to represent entities, and assumes that all entities have a corresponding page in Wikipedia [5].

A main difference between the INEX entity ranking and ad hoc retrieval tasks lies in the assessments. In ad hoc retrieval, a document is judged relevant if any piece of the document is relevant. In the entity ranking track, a document can only be relevant if the document is of the correct entity type, resulting in far less relevant documents. The correct entity type is specified during topic creation as a target category.

3.1 Topics

For our experiments we use different INEX topic sets from the 2007 ad hoc and entity ranking tracks. The ad hoc assessments are based on highlighted passages. Since we only do document retrieval and do not return document elements or passages, we have to modify the ad hoc assessments. In our experiments, a document is regarded as relevant if some part of the article is regarded as relevant, i.e. highlighted by the assessor [8]. Ad hoc topics consist of a title (short keyword query), an optional structured query, a one line description of the search request and a narrative with more details on the requested topic and the task context. Entity ranking topics do not have an optional structured query, but they do include a few relevant example entities, and one or a few target categories. The example entities are used in a list completion task, that we do not consider in this paper. We only use the topic titles, and the target categories of the entity ranking topics.

We run our experiments on the following topic sets:

- Set A: Entity ranking topics 60-100, consisting of 25 assessed genuine entity ranking topics.
- Set B: Entity ranking topics 30-59, consisting of 19 assessed entity ranking topics derived from ad hoc topics

Table 1: Topic examples

Set	Topic Title	Description	Target Category
A	Paul Auster novels	I want a list of novels written by Paul Auster	685: novels
B / C_1	Van Gogh paintings	Find lots of paintings from Van Gogh	87939: work of vincent van gogh
C_2	Image processing segmentation wavelet	Find information about the use of wavelets for image processing	

- Set C: Ad hoc topics 414-543, consisting of 99 assessed ad hoc topics.
  - C_1: 19 Ad hoc topics that have been used to create the entity ranking topics 30-59.
  - C_2: The remaining 80 ad hoc topics.

Set A consists of genuine entity ranking topics, set C_2 consists of genuine ad hoc topics. Set B and set C_1 consist of the same topics, but with different relevance assessments, i.e. entity ranking assessments for set B and ad hoc assessments for set C_2. These different topic sets allow us to explore the relations between ad hoc retrieval and entity ranking. Example topics of each set are given in Table 1.

3.2 Relevance Assessments

In order to gain some information on category distributions within the retrieval results, we analyze the relevance assessment sets. We show some statistics in Table 2. As expected, the ad hoc topics contain more relevant pages. The relevance assessment set of topic set B, contains all relevant pages from topic set C_1. Of these pages 41.4% are relevant for the entity ranking task.

For each topic we determine the most frequently occurring category in either all pages or only the relevant pages. Then we calculate what percentage of pages is assigned to this majority category. For the ad hoc topic sets the categories are the most diverse, only around 6-7% of the pages belong to the same category. The categories in the entity ranking topic sets are more focused, with 16.3% of pages in set A belonging to the majority category, and even 31.6% of the pages in set B.

The majority categories in the relevant pages are quite large within these relevant pages, around 60% for the entity ranking topics, and still around 32% for the ad hoc topics. What is interesting for the entity ranking topics, is that this percentage is much higher than the percentage of relevant pages belonging to the target category. This means that there are categories other than the target category, which are good indicators of relevance. It seems that in many cases the target category can be more specific, e.g. to our example topic 'Paul Auster novels' category '685: novels' is assigned. The majority category in the relevant pages is '68456: books by paul auster'. This category is far more specific, and using it probably leads to better results.

For all topic sets we see that from the relevant pages a far higher percentage belongs to the majority category than non-relevant pages. This might imply that category information can not only be beneficial for entity ranking topics, but also ad hoc topic results could be improved if the right target categories can be found.

For the entity ranking topics we can also determine how many of the pages belong to one of the specified target categories. In fact, only 11.3% of set A pages and 16.7% of set B pages belong to a target category. The runs used to create the pool for topic set B are ad hoc runs, so the target categories have not been taken into consideration here. In topic set A however the target categories were available, but here less pages belong to the target category indicating that target categories themselves are not treated as an important feature in the submitted runs. Considering that 11.1% of the non-relevant pages also belong to the target category, this is a good decision.

Table 2: Relevance assessment sets statistics

Set	A	B	C_1	C_2
Avg. # of pages	485	83	611	612
Avg. % relevant pages	0.04	0.414	0.135	0.089
Pages with majority category of all pages:				
all pages	0.163	0.316	0.066	0.059
relevant pages	0.313	0.426	0.200	0.200
non-relevant pages	0.154	0.167	0.045	0.048
Pages with majority category of relevant pages:				
all pages	0.084	0.281	0.047	0.047
relevant pages	0.590	0.630	0.318	0.316
non-relevant pages	0.064	0.074	0.016	0.028
Pages with target category:				
all pages	0.113	0.167		
relevant pages	0.277	0.387		
non-relevant pages	0.111	0.048		

Over all kinds of pages, set B has more focused categories than set A, the genuine entity ranking set. This can be explained by the fact that the pages in set B were already assessed as relevant for the ad hoc topic, so at least topically they are more related.

Now that we have found some indications that category information is indeed useful for entity ranking topics, and could also be useful for ad hoc topics, in the next section we describe how we can make use of the category information.

4. USING CATEGORY INFORMATION

Category information has been proved to be of great value for entity ranking in Wikipedia [11, 12]. Other sources of information like link information also lead to improvements, but these are in general much smaller. Although for each topic one or a few target categories are provided, relevant entities are not necessarily associated with these provided target categories. In section 3.2 we already saw that on our data sets less than 40% of the relevant pages belong to the target category. Simply filtering on pages belonging to the target category already helps, but more can be done. We have to take into account that relevant entities can also be associated with categories linked to or from the target category or other similar categories. Therefore, we define similarity functions between the categories of retrieved entities and the target categories. The similarity scores are estimated using lexical similarity of category contents. This approach is similar to Vercoustre et al. [12], but they use lexical similarity of category names. We think using category contents will make our approach more robust. 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.

4.1 Model

To calculate the distance between two categories we use a language modeling approach [6]. First of all we make a maximum likelihood estimation of the probability of a term occurring in a document belonging to a certain category. To avoid a division by zero, we smooth the probabilities of a term occurring in a category

with the background collection:

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

where C, the category, consists of the concatenated text of all pages belonging to that category. D is the entire Wikipedia document collection, which is used to estimate background probabilities. The final $P(t|C)$ is estimated with a parsimonious model [7] that uses an iterative EM algorithm as follows:

$$\begin{aligned} \text{E-step:} \quad e_t &= tf_{t,C} \cdot \frac{\alpha P(t|C)}{\alpha P(t|C) + (1-\alpha)P(t|D)} \\ \text{M-step:} \quad P(t|C) &= \frac{e_t}{\sum_t e_t}, \text{i.e. normalize the model} \end{aligned} \quad (2)$$

The maximum likelihood estimation of $P(t|C)$ is used as initial probability. Now we can calculate distances between categories. We do this using KL-divergence to calculate a category score that is high when the distance between the categories is small as follows:

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

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 is the highest score, corresponding to the smallest distance, from the scores $S_{cat}(C_t|C_d)$, the scores for the distances from the categories of the document to the target category. In contrast to Vercoustre et al. [12], 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 minimal distance of the distances from the answer entity categories 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 are 0, no matter what other categories are assigned to the document. The category score for an answer entity in relation to a query topic ($S_{cat}(d|QT)$) is the sum of the scores of all target categories:

$$S_{cat}(d|QT) = \sum_{C_t \in QT} \arg\max_{C_d \in d} S_{cat}(C_t|C_d) \quad (3)$$

Besides the category score, we also need a query score for each document. This score is calculated using a language model with Jelinek-Mercer smoothing without length prior:

$$P(q_1, ..., q_n|d) = \sum_{i=1}^{n} \lambda P(q_i|d) + (1-\lambda)P(q_i|D) \quad (4)$$

Finally, we combine our query score and the category score through a linear combination. Both scores are calculated in the log space, and then a weighted addition is made.

$$S(d|QT) = \mu P(q|d) + (1-\mu)S_{cat}(d|QT) \quad (5)$$

4.2 Automatic Assignment

Besides using the target categories provided with the query topics, we also look at the possibility of automatically assigning target categories to entity ranking and ad hoc topics. Since the entity ranking topic assessments heavily depend on the target categories used during assessment, the automatically assigned categories will have to be suitably similar to the provided target categories in order to perform well. The advantage of automatically assigning target categories is that no effort from a user is required.

There are many ways to do automatic topic categorization, for example by using text categorization techniques [9]. For this paper we keep it simple and exploit the existing Wikipedia categorization of documents. From our baseline run we take the top 10 results, and look at the most frequently occurring categories belonging to these documents. The two most frequent categories are assigned as target categories to the query topic, if the category occurs at least two times in the top results.

Entity ranking topics look for a collection of pages belonging to the same entity type, instead of just any type of document. Ad hoc topics look for any type of document as long as it is on the query topic. Not all categories in Wikipedia are entity types, but there are also topical categories, that are assigned to all pages on that particular topic. The automatic assignment of categories is applied in the same way to entity ranking and ad hoc topics, but looking at these categories for the entity ranking topics in almost all cases the category can be considered as a (usually low level) entity type. For the ad hoc topics there is still a considerable number of entity type categories but topical categories occur regularly here as well. In order to compare manual and automatic assignment of categories on the ad hoc topics as well, we have manually assigned target categories to the ad hoc topics. These categories are not necessarily entity types, the category that seems closest to the query topic is selected, i.e. for the query "Steganography and its techniques" the category "Steganography" is assigned as target category.

5. EXPERIMENTS

In this section we first describe our experimental setup before going into details of our results. We compare ad hoc topics to entity ranking topics, and also we look at the effects of using automatically assigned categories instead of manually assigned target categories.

5.1 Experimental Setup

To create our baseline runs incorporating only the query score, we use the Indri search engine [10]. Our baseline model is a language model using Jelinek-Mercer smoothing with a collection λ of 0.1. Also pseudo-relevance feedback is applied, using the top 50 terms from the top 10 documents. The category score is only calculated for the top 1000 documents of the baseline run. These documents are reranked to produce the run that combines query and category score.

We have four topic sets A, B, C_1, and C_2, and two possible topic assignments, manually or automatically. The manual categories are the provided target categories for the entity ranking topics and for the ad hoc topics target categories are manually assigned by the author. For the parameter μ we tried values from 0 to 1, with steps of 0.1. The best values of μ turned out to be on the high end of this spectrum, therefore we added two additional values of μ: 0.95 and 0.98.

5.2 Experimental Results

The results of our experiments expressed in MAP are summarized in Table 3. This table gives the query score, which we use as our baseline, the category score, the combined score using $\mu = 0.9$ and the best score of their combination with the corresponding value of μ.

The baseline score on the entity ranking topics is quite low as expected. Using only the keyword query for article retrieval, and disregarding all category information, can not lead to good results since the relevance assessments are based on the category information. For the ad hoc topics on the other hand, the baseline scores are good. They would have been ranked among the top 10 participants

Table 3: Retrieval results in MAP

Cats	Set	Query μ = 1.0	Category μ = 0.0	Comb. μ = 0.9	Best Score μ	
Man	A	0.1840	0.1231 $^{-}$	0.2481$^{\circledcirc}$	0.9	0.2481$^{\circledcirc}$
Man	B	0.2804	0.2547 $^{-}$	0.3848$^{\bullet}$	0.8	0.4039$^{\bullet}$
Man	C_1	0.3653	0.2067°	0.4308$^{\circledcirc}$	0.9	0.4308$^{\circledcirc}$
Man	C_2	0.3031	0.1761$^{\bullet}$	0.3297$^{\circledcirc}$	0.95	0.3327$^{\bullet}$
Auto	A	0.1840	0.1779 $^{-}$	0.2308 $^{-}$	0.8	0.2221°
Auto	B	0.2804	0.2671 $^{-}$	0.3607$^{\circledcirc}$	0.9	0.3607$^{\circledcirc}$
Auto	C_1	0.3653	0.2641°	0.3923 $^{-}$	0.95	0.4081$^{\circledcirc}$
Auto	C_2	0.3031	0.1692$^{\bullet}$	0.3122 $^{-}$	0.95	0.3284$^{\circledcirc}$

Significance of increase or decrease over baseline (query score) according to t-test, one-tailed, at significance levels 0.05($^{\circ}$), 0.01($^{\circledcirc}$), and 0.001($^{\bullet}$).

of the 2007 ad hoc track.

The best value for μ differs per topic set, but for all sets μ is quite close to 1. This doesn't mean however that the category scores are not important, which is also clear from the improvements achieved. The reason for the high μ values is that the category scores are in a larger order of magnitude, because instead of scoring a few query terms, all the terms occurring in the language model of the category are scored. So even with small weights, the category score contributes significantly to the total score. Normalizing the scores can give a realistic estimation of the value of the category information.

Ad Hoc vs. Entity Ranking

From the four topic sets, the baseline scores best on the two ad hoc topic sets C_1 and C_2. 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 topics derived from the ad hoc topics are a selection of the complete ad hoc topic set, and mostly easy topics with a lot of relevant pages are selected. The genuine entity ranking topics are developed by the participants in the INEX entity ranking track who have less insight into topic difficulty.

The entity ranking topics benefit greatly from using the category information with significant MAP increases of 35% and 44% for topic sets A and B respectively. When only the category score is used to rerank the top 1000 results, the scores are surprisingly good, for set B MAP only drops a little with no significant difference from 0.2804 to 0.2655. Apparently the category score really moves up relevant documents in the ranking. 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 with MAP increases of 18% and 10% for set C_1 and C_2 respectively. So, we have successfully applied entity ranking techniques to improve retrieval on ad hoc topics. The improvements are bigger on the ad hoc topics that are later converted into entity ranking topics, indicating that queries that can be labeled as entity ranking topics benefit the most from using category information.

Our approach compares favourably to other approaches on these data sets. Topic sets A and B together form the test data of the 2007 INEX entity ranking track. Our best score on this test data is achieved with $\mu = 0.8$ which leads to a MAP of 0.313. This score is better than any of the official submitted runs, of which the best run achieves a MAP of 0.306. Likewise, topic sets C_1 and C_2 together form the test data of the 2007 INEX ad hoc track. If we compare our best automatic run to the official submitted runs, our combined run including the category score with $\mu = 0.95$ outperforms all official runs in two out of the three tasks in this track.

Manual vs. Automatic Assignment

From the analysis of the relevance assessment sets we already got some indications that there are other categories besides the target category that are dominant in the relevant documents. It turns out that the automatic assignment of target categories works not as good as manually assigning the target categories, but it does come sufficiently close. When we look at the category scores only, the automatically assigned topics perform even better than the manually assigned categories, except on set C_2, the real ad hoc topic set. For the ad hoc topic set C_2, using manually assigned topics leads to an improvement of 8% over the baseline, resulting in a best score of 0,3284, compared to 0,3327 with manually assigned categories. For all topic sets using the automatically assigned categories leads to significant improvements over the baseline.

During the automatic assignment we use the top 10 results of the baseline run as surrogates to represent relevant documents. So we would expect that if the precision at 10 is high, this would lead to good target categories. However, precision at 10 of the baseline for topic set A, is only 0.2640, but the category score is almost as good as the query score (0.1840 and 0.1779 respectively). On the real ad hoc topics precision at 10 is better with 0.4787, but the category score does not improve accordingly.

The question remains why the combined scores of the automatically assigned categories are worse than the combined scores of the manually assigned categories while their category scores are higher. The automatically assigned categories may find documents that are already high in the original ranking of the baseline run, since the categories are derived from the top 10 results. The manually assigned categories do not necessarily appear frequently in the top results of the baseline, so the category scores can move up relevant documents that were ranked low in the baseline run.

6. CONCLUSION

In this paper we have investigated the use of annotations or categories to find entities and information in the Wikipedia domain, examining the differences and similarities between these two tasks. We started with analyzing the relevance assessment sets for entity ranking and ad hoc topic sets. Less than 40% of the relevant pages belong to a provided target category, so simply filtering on the target category is not sufficient for effective entity ranking. Furthermore, the provided target categories are not always the majority category among the relevant pages, these majority categories are often more lower level categories.

Moving on to our experiments, we found a positive answer on our research question if entity ranking techniques can be used on ad hoc topics. Using category information leads to significant improvements over the baseline for both ad hoc and entity ranking topics. Considering our second research question, automatically assigned categories prove to be good substitutions for manually assigned target categories. Similar to the runs using manually assigned categories, using the automatically assigned categories leads to significant improvements over the baseline for all topic sets.

In future work we want to experiment with different methods to estimate distances between categories. Besides calculating KL-divergence scores on category contents and titles, distances can be estimated by counting collocations of categories on Wikipedia pages, or by calculating distances in the category graph. Another issue that requires some more attention is the automatic assignment of categories. We have only experimented with an approach that uses the majority categories of the top 10 results, but it might well

be possible to obtain better categories with other approaches. Finally, our test collection has some limitations. Wikipedia is a very controlled form of user-generated content, so it is still a question whether a similar approach can be applied to less organized networks of user-generated content. Furthermore, our Wikipedia collection is already a few years old and does not include page and category redirects. In the last years Wikipedia is evolving rapidly so it would be interesting to repeat some experiments using all features of the current Wikipedia.

Acknowledgments.

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

REFERENCES

[1] J. Bai, J.-Y. Nie, H. Bouchard, and G. Cao. Using query contexts in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 15–22. ACM Press, New York NY, 2007.

[2] K. Balog. The SIGIR 2008 workshop on future challenges in expertise retrieval (fCHER). *SIGIR Forum*, 42(2):46–52, December 2008.

[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, 2007.

[4] H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: efficient search on text, entities, and relations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

[5] A. P. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In *Focused Access to XML Documents*, 2007.

[6] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001.

[7] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings SIGIR 2004*, pages 178–185. ACM Press, New York NY, 2004.

[8] J. Kamps, S. Geva, A. Trotman, A. Woodley, and M. Koolen. Overview of the INEX 2008 ad hoc track. In *INEX 2008 Workshop Pre-proceedings*, pages 1–28, 2008.

[9] R. Kaptein and J. Kamps. Web directories as topical context. In *Proceedings of the 9th Dutch-Belgian Workshop on Information Retrieval (DIR 2009)*, 2009.

[10] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[11] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In *Focused Access to XML Documents*, pages 306–320, 2007.

[12] A.-M. Vercoustre, J. Pehcevski, and J. A. Thom. Using wikipedia categories and links in entity ranking. In *Focused Access to XML Documents*, pages 321–335, 2007.

[13] Wikipedia. The Free Encyclopedia, 2008. URL: `http://www.wikipedia.org`.

[14] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, 2006.