Result Diversity and Entity Ranking Experiments: Anchors, Links, Text and Wikipedia

Rianne Kaptein[1] **Marijn Koolen**[1] **Jaap Kamps**[1,2]

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

[2] ISLA, Informatics Institute, University of Amsterdam

Abstract: In this paper, we document our efforts in participating to the TREC 2009 Entity Ranking and Web Tracks. We had multiple aims: For the Web Track's Adhoc task we experiment with document text and anchor text representation, and the use of the link structure. For the Web Track's Diversity task we experiment with using a top down sliding window that, given the top ranked documents, chooses as the next ranked document the one that has the most unique terms or links. We test our sliding window method on a standard document text index and an index of propagated anchor texts. We also experiment with extreme query expansions by taking the top n results of the initial ranking as multi-faceted aspects of the topic to construct n relevance models to obtain n sets of results. A final diverse set of results is obtained by merging the n results lists. For the Entity Ranking Track, we also explore the effectiveness of the anchor text representation, look at the co-citation graph, and experiment with using Wikipedia as a pivot. Our main findings can be summarized as follows: Anchor text is very effective for diversity. It gives high early precision and the results cover more relevant sub-topics than the document text index. Our baseline runs have low diversity, which limits the possible impact of the sliding window approach. New link information seems more effective for diversifying text-based search results than the amount of unique terms added by a document. In the entity ranking task, anchor text finds few primary pages , but it does retrieve a large number of relevant pages. Using Wikipedia as a pivot results in large gains of P10 and NDCG when only primary pages are considered. Although the links between the Wikipedia entities and pages in the Clueweb collection are sparse, the precision of the existing links is very high.

1 Introduction

Modern Web search requires the combination of traditional topical relevance with other features such as authority, recency, or diversity. In practice combining indicators of these different features is hard: features may be sparse, have different strengths, or have radically different score distributions. This can easily lead to disappointing results with straightforward combination methods—even if the features are inherently useful. We propose a new 'sliding window' approach that allows for combining relevance with another feature. Given an initial ranked list, we use a sliding window of n documents, where the window size controls the relative importance of the original relevance ranking. Of the documents within the window, we select the document with the highest score on the new feature, and then slide the window down the ranking. Assume we have an indicator of diversity and set $n = 10$, then the first ranked document will be the most diverse from the top 10 of the original ranking, then we add the 11th ranked document to the window, and again select the most diverse one. Etcetera. The approach is robust in the sense that i) the relevance ranking is used as a basis and is guaranteed to be broadly respected, and ii) the exact scores of the feature are treated independently of the relevance scores, thereby avoiding unfortuitous effects in the combination.

For the Adhoc Task, we made a number of runs using the document-text and propagated anchor-texts. We also aimed for multi-faceted results by using the top 10 retrieved pages as different aspects of the topic. For each aspect a separate relevance model is created, and the resulting runs are merged into a final ranking having a more diverse set of results. For our Diversity Task experiments we apply the above sliding window approach to different ad hoc runs. We assume that the diversity topics are fairly broad, with hundreds or thousands of relevant documents. The initial ranked list will have very high precision in the first hundred or hundreds of results, and we opt to conservatively re-rank them using a window size of 10. Specifically, we look at two new features: a link filter and a term filter. Documents co-citing or co-cited by the same set of documents are topically related and contain similar content. Our assumption is that a document with

many unseen links contains unseen information about the topic, thereby diversifying the results. Hence, we select the document that introduces the most unseen links to the results so far. Alternatively, we filter or re-rank the results list based on term overlap. By boosting documents that contain many terms that do not occur in the results seen so far, we aim to maximize the amount of new information added to the top of the ranked list.

Entity ranking on the Web is a difficult task with many pitfalls. Before any entities can be ranked, they first have to be recognized as entities and classified into the correct entity type. Our hypothesis is that effective entity ranking on the web can be achieved by exploiting the available structured information to make sense of the great amount of unstructured web information. We propose to use Wikipedia to avoid the problem of entity recognition, and to simplify the entity type classification. Wikipedia is an excellent resource for entity ranking because of its elaborate category structure. The TREC entity ranking track investigates the problem of related entity finding, where entity types are limited to people, organizations and products. The people, organization and product entity types can easily be mapped to Wikipedia categories. Successful methods for entity ranking in Wikipedia have been explored in the entity ranking task that runs since 2007 at INEX (Initiative for the Evaluation of XML retrieval) [3, 4] . We investigate the relations between the TREC and the INEX entity ranking task, and try to carry over methods that have proven effective at the Wikipedia task. To retrieve web pages outside of Wikipedia we make use of link information, in particular the external links already present on the Wikipedia pages. The effectiveness of the Wikipedia pivot approach is compared to the effectiveness of two other retrieval methods: using a propagated anchor-text index, and using co-citation information.

The rest of this paper is organized as follows. In Section 2, we describe the experimental set-up. In Section 3, we discuss our experiments for the Web Track and our Entity Ranking experiments are discussed in Section 4. Finally, we summarize our findings in Section 5.

2   Experimental Set-up

For both the Entity Ranking and Web Tracks we only used the category B of the ClueWeb collection, and Indri [6] for indexing. Stopwords are removed and terms are stemmed using the Krovetz stemmer. We built the following indexes:

Text: contains document text of all documents in ClueWeb category B.

Anchor: contains the anchor text of all documents in ClueWeb category B. All anchors are combined in a bag of words. 37,882,935 documents (75.43% of all documents) have anchor text and therefore at least one incoming link.

Web only: contains document text of all non-Wikipedia documents in ClueWeb category B. This consists of all documents in part en0000 to en0011.

Wikipedia only: contains document text of all Wikipedia documents in ClueWeb category B. This consists of all documents in part enwp00 to enwp03.

For all runs, we use Jelinek-Mercer smoothing, which is implemented in Indri as follows:

$$P(r|D) = \frac{(1 - \lambda) \cdot tf_{r,D}}{|D|} + \lambda \cdot P(r|C) \qquad (1)$$

where D is a document in collection C. We use little smoothing ($\lambda = 0.15$), which was found to be very effective for large collections [7, 8].

For ad hoc search, pages with more text have a higher prior probability of being relevant [9]. Because some web pages have very little textual content, we use a linear document length prior $\beta = 1$. That is, the score of each retrieved document is multiplied by $P(d)$:

$$P(d) = \frac{|d|^\beta}{\sum |d'|^\beta} \qquad (2)$$

Using a length prior on the anchor text representation of documents has an interesting effect, as the length of the anchor text is correlated to the incoming link degree of a page. The anchor text of a link typically consists of one or a few words. The more links a page receives, the more anchor text it has. Therefore, the length prior on the anchor text index promotes web pages that have a large number of incoming links and thus the more important pages.

3   Web Track

We submitted runs for both the Adhoc and Diversity Tasks. We experiment with using the anchor text of web pages as alternative document representation. The effectiveness of anchor text for locating relevant entry pages is well established [2, 9] but for ad hoc search it seems less useful [5, 8]. Given the fairly large coverage of the anchors–more than 75% of the documents in the collection have at least one incoming link–and the high density of the link graph–we extracted over 1.5 billion collection-internal links–the anchor index could give high early precision, which is required for the Diversity task. As anchor text provides a document representation that is disjoint from the document text, documents that have very similar anchor text might have more *dissimilar* document text. This could be useful for generating a diverse results list.

The ClueWeb collection also contains a snapshot of the English Wikipedia, which is very different in nature from the rest of the World Wide Web. We want to directly compare the results from Wikipedia against results from the rest

of the Web. Because Wikipedia has encyclopedic articles on single topics, it plausibly has lower redundancy of information than the Web. This might have a significant impact on the diversity of retrieved Wikipedia pages, as each page should have unique content and a list of Wikipedia pages should naturally be diverse.

3.1 Diversifying Retrieval Results

We use two methods to diversify search results. The first method post-processes the initial ranked list using a top down filter and the second method is an extreme form of relevance feedback.

3.1.1 Filtering using sliding windows

To make results in the ranked list more diverse, we experiment with a top-down filtering method using a sliding window of n documents. We keep the highest ranked result as is and choose from the next n documents the one that maximises diversity according to some diversity indicator. We then slide the window down one step in the list and repeat the process. All official runs use a window of size $n = 10$. This filter allows us to easily test the utility of different document features before spending a lot of time finding the proper way to combine the most effective features. Because the filter is relatively conservative–documents can move up at most $n - 1 = 9$ ranks–the initial relevance ranking is broadly respected and we avoid low ranked off-topic documents with extremes scores on some feature from infiltrating the higher ranks. If a certain feature is not useful for a certain task–in this case diversity–the sliding window approach guarantees its impact will be small. If we find an effective feature, we can easily make its impact bigger by increasing the size of the sliding window. As diversity indicators we use the number of new terms or new links introduced by the next document.

Term Filter (TF): Term overlap is often used to measure document similarity. We use the inverse of this idea to achieve diversity. Given the highest ranked document(s), the next document should add new terms to those the user has already seen in higher ranked documents. From the documents within the sliding window we choose the one that has the most new terms to optimise diversity. A side effect of this feature is that it favours long documents, as they tend to contain more distinct terms.

Link Filter (LF): Another measure of document similarity is co-citation coupling, which is used in citation analysis. The more citations two documents have in common, the more similar their subject matter. We use the same approach as with the term-based filter and choose from the documents in the sliding windows the one that

has the most new incoming or outgoing links. With incoming links we measure how often a document is cited by others that do not cite documents higher in the ranking. With outgoing links we measure how often a document cites web pages that are not cited by documents higher in the ranking. A side effect of using incoming links is that it favours documents with a high indegree, which are typically entry pages of sites or popular pages. A side effect of using outgoing links is that it favours documents with a high outdegree, which are typically list pages or index pages.

3.1.2 Merging results from multiple relevance models

Another method is to use the top n documents as n different aspects of the search topic, and use them for relevance feedback to obtain diverse expanded queries. For each document a separate relevance model is created to obtain n results list, which are then merged into a final ranking. Assuming that each document will give a different relevance model, each query will represent the overall topic in a slightly different context. Our submitted runs use $n = 10$ documents.

3.2 Official Runs

We submitted two runs for the Adhoc Task:

UamsAw7an3: mixture of text and anchor text runs. $s_{mix}(d) = \lambda \cdot s_{text}(d) + (1 - \lambda) \cdot s_{anchor}(d)$ with $\lambda = 0.7$

UamsAwebQE10: full ClueWeb text index. 10 different relevance models are constructed, one from each document in the top 10 results. The results retrieved using the 10 relevance models are merged into a final ranking based on their retrieval scores.

We submitted three runs for the Diversity Task:

UamsDancTFb1: Anchor text index run with length prior $\beta = 1$, term filter applied with $n = 10$.

UamsDwebLFout: Text index run with length prior $\beta = 1$, link filter applied using all outgoing links and $n = 10$

UamsDwebQE10TF: Text index run with length prior $\beta = 1$, each result in the top is used as a separate document for query expansion. Final run is a merge of 10 runs using different relevance models.

3.3 Results

We will first discuss results of our baseline runs to show the relative effectiveness of the various indexes.

3.3.1 Baseline results

For the Adhoc Tasks we report the official statMAP measure and statMPC@30 in Table 1. Clearly, the length prior has a big impact on performance. On the text index, both early and overall precision increase when the length prior is used. On the anchor text index, the overall precision drops slightly when using the length prior, but the early precision vastly improves. Because most documents in the collection have no or only a few incoming links, the anchor text of these documents is poor. Thus, the anchor text run will miss many of the relevant documents, as is reflected by the low average precision. Although we expected the *Anchor* run to do well on early precision, the estimated P@30 of 0.5558 seems very high when compared to similar *Anchor* only runs on the TREC Terabyte tracks [7, 8] where their scores for P@10 and MAP are usually well below those of a full-text run. A possible explanation might be found when considering the way relevance is estimated. If most runs contributing to the assessment pool use a similar document representation, a single run using a very different document representation might a very different set of documents in the top ranks, which have a low sampling probability. A document with a low sampling probability that is judged relevant represents many estimated relevant documents and can result in per topic precision scores above 1.0 for runs that ranks these documents highly, thereby boosting the overall scores significantly. As mentioned before, the document representation of the anchor texts will be very different from the full text representation, and hence result in a very different ranked list. Plausibly, the anchor text model ranks certain relevant documents highly that have a low sampling probability, resulting in an estimated precision well above 1, as is the case for our anchor text run. The high statMPC@30 might be an over-estimation. We removed the pool inclusion probability column from the official *prels* and used standard *trec_eval* to see if the traditional P@30 measure gives similar results (Table 2) and found that the anchor text run has a much lower P@30 than the full-text run. On the other hand, the *Anchor* run has a much higher Mean Reciprocal Rank (MRR) than the *Text* run. The MRR can never be over-estimated, as it simply looks at the rank of the first retrieved relevant document. The precision at rank 30 could be under-estimated if the number of judged results is low.

The *Anchor* run has many more non-judged documents in the top ranks than the *Text* run. At rank 5, less than 69% of the results of the *Anchor* run are judged and at rank 30, less than 29% is judged, while for the *Text* run, over 89% of the results at rank 5 are judged and over 68% at rank 30. This is a strong indication that the traditional P@30 score of the *Anchor* run is underestimated. Together with the much higher MRR of the *Anchor* run and the lower number of judged results in the top ranks, this supports the high statMPC(30).

The *Web only* index gives much better results than the *Wikipedia* index. This is to be expected, as the *Web only* index has many more documents and also arguably more re-

Table 1: Results for the 2009 Adhoc Task. Best scores are in bold-face.

| | statMAP | | statMPC@30 | |
Run	$\beta = 0$	$\beta = 1$	$\beta = 0$	$\beta = 1$
Text	0.0991	0.1442	0.2208	0.3079
Anchor	0.0676	0.0567	0.2010	**0.5558**
0.7 Text + 0.3 Anchor	0.1244	**0.1687**	0.2952	0.4812
Web only	0.0880	**0.1044**	0.2181	**0.2528**
Wikipedia	0.0483	0.0748	0.1946	0.2433

Table 2: Ad hoc results using traditional measures and binary judgements

| | | | % Judged in | |
Run	P@30	MRR	Top 5	Top 30
Text $\beta = 1$	0.2827	0.3061	89.20	68.33
Anchor $\beta = 1$	0.1607	0.6335	68.80	28.20

Table 3: Impact of length prior on Diversity performance of baseline runs. Best scores are in bold-face.

| | α-nDCG@10 | | IA-P@10 | |
Run	$\beta = 0$	$\beta = 1$	$\beta = 0$	$\beta = 1$
Text	0.094	0.120	0.038	0.054
Anchor	0.178	**0.257**	0.054	0.082
0.7 Text + 0.3 Anchor	0.156	0.223	0.066	**0.083**
Web only	0.081	0.094	0.032	0.040
Wikipedia	0.065	**0.124**	0.037	**0.071**

dundant information. But as both sub-collections have relevant documents, the combined index contains more relevant documents and is therefore even more effective.

For the Diversity Tasks we report the official α-nDCG@10 and IA-P@10 measures in Table 3. Again, we see that the length prior has a big positive impact on the diversity scores of the baseline runs. Give their impact on the Adhoc scores, this is not surprising. The runs with the length prior have more relevant documents in the top ranks and thus have more documents that receive score for the diversity measures as well. The anchor text run scores much higher for the diversity measures than the full-text run, in line with the Adhoc results. Although we explained why the observed high early precision score for the Adhoc Task might be an over-estimation, these Diversity results, which are based on different pools and different relevance judgements, indicate that the anchor text run really has more relevance in the top ranks.

When we look at the performance of the *Web only* and *Wikipedia* runs, we see that the length prior again improves the ranking. Recall that on the Adhoc measures, the *Wikipedia* run was less effective than the *Web only* run, with and without length prior. However, for the Diversity Task,

the *Wikipedia* run scores higher on both reported measures. This could mean that the Wikipedia results are more precise, or that it is easier to find relevant pages in the relatively homogeneous and spam-free Wikipedia than in the much larger Web. This will be discussed further in Section 3.3.2.

3.3.2 Diversifying methods

Finally, we show the impact of the diversity specific methods in Table 4. Runs filtered on distinct terms are denoted with $TF(n)$ wherer n is the size of the sliding window. Runs filtered on distinct links are denoted with $LF(d, n)$ where d is the direction of the links (incoming or outgoing) and n is the size of the sliding window. We use $RF(10)$ to denote a run merged from the 10 relevance feedback runs.

If method A scores better on a diversity measure than method B, it does not necessarily mean it has a more *diverse* ranking. The higher score could simply be the result of a better relevance ranking. To see if differences observed in the scores of the diversity measures are caused by a better relevance ranking or a more diverse ranking, we present standard document ranking measures as well. We compare α-nDCG@10 with standard nDCG@10 and IA-P@10 with P@10. For this, we mapped the Diversity *qrels* to standard TREC Adhoc *qrels* by assuming a document is relevant for a topic if it is relevant for at least one sub-topic.

We see that the term filter leads to a drop in performance for all baseline runs on all measures. The number of unseen terms seems ineffective as a feature to diversify search results. The link filter leads to better scores on both the traditional Adhoc measures as on the Diversity measures. Overall, the incoming links are more effective than the outgoing links, although in combination with the merged $RM(10)$ run, the outgoing links are slightly more effective for P@10 and IA-P@10. The feedback run $RF(10)$ also improves the document ranking and diversity of the baseline run. On the *Anchor* run, the filters are not effective. Of course, the *Anchor* run already uses the number of incoming links implicitly through the length prior. Further boosting documents with many new incoming or outgoing links only hurts performance. By combining the anchor text and full-text runs, we get a slight improvement on IA-P@10. If we then apply the link filters, the P@10 and IA-P@10 scores go up further. The incoming links are more effective than the outgoing links.

It is hard to judge whether the diversity methods actually affect the diversity of the baseline runs. If we compare the scores for the ad hoc measures nDCG@10 and P@10 with the diversity measures α-nDCG@10 and IA-P@10, we see similar patterns. Runs that score higher on nDCG@10 also score higher on α-nDCG@10 and runs that score higher on IA-P@10 also score higher on P@10. This suggest that the changes on the diversity scores do not reflect changes in actual diversity. The link filters seem to merely work as indegree priors and push up important documents. Ad hoc preci-

sion goes up a lot but diversity goes up only a little bit. The run is not more diverse but simply has more relevance in the top ranks.

To shed some more light on how our methods affect the diversity of the results, we look at the percentage of sub-topics for which relevant documents are found. In Table 5 we show the percentage (macro average) of sub-topics covered by the retrieved results at various rank cut-offs. In the relevance judgements we find relevant documents for 199 different sub-topics for 49 topics. This means that for one of the 50 topics, not a single document in the pool was judged relevant for one of the chosen sub-topics. We see that the top 10 documents of the $Text$ run contain relevant documents for only 16.3% out of the 199 sub-topics while the top 10 of the $Anchor$ run covers 28.5%. The anchor text run is thus not only more precise, but also more diverse. The term filter has a small negative impact on the number of sub-topics found, while the link filters have a positive impact, except for the $Anchor$ run. The outlink filter is boost more diverse sub-topics than the inlink filter. The merged query expansion runs make the top ranked results more diverse, showing that the improvements for the diversity measures in Table 4 are not only based on higher precision. Combining the $Text$ and the $Anchor$ runs has almost no impact on the number of sub-topics covered in the top ranks of the baseline run. For this run, the inlink filter is more effective than the outlink filter. If we look further down the ranking, we see that relevant documents for much more sub-topics are retrieved. The impact of the diversity methods is almost negligible at rank 100 and lower. The combination of $Text$ and $Anchor$ runs does increase the number of topics found later in the ranking. The *Wikipedia* run is far less diverse than the *Web only* run. The higher diversity score must come from a better relevance ranking of the top results.

Note that the sliding window filter allow documents to move up $n - 1$ at the most. Thus, for the top 10 documents, a sliding window of $n = 10$ documents can select documents from the top 19 results of the original ranking. The number of sub-topics found in the top 20 of the original ranking provides an upper bound of the number of topics that we can possibly have in the top 10 of the filtered runs. The small impact of the filters is due to the low diversity in the initial text-based relevance ranking. With only 26.1% of the sub-topics covered in the top 20 results for 49 topics (1.06 sub-topics per topic), there is not much to diversify. For the filters to have more impact, the windows size needs to be increased to move up documents from further down the ranking. As mentioned before, the danger is that this leads to infiltration of off-topic documents that have many links or are very long. The sliding window size is kept low to broadly respect the initial text-based ranking. With larger window sizes, the impact of the initial ranking decreases.

Table 4: Results for runs using the sliding window filters and merge of multiple query expansions on the 2009 Adhoc topics. Best scores are in bold-face.

| | | Diversity | | |
Run	nDCG@10	α-nDCG@10	P@10	IA-P@10
$Text$	0.1564	0.120	0.1700	0.054
$Text\,TF(10)$	0.1450	0.122	0.1560	0.048
$Text\,LF(in,10)$	**0.1924**	**0.154**	**0.2020**	**0.068**
$Text\,LF(out,10)$	0.1873	0.145	0.2000	0.063
$Text\,RF(10)$	0.1888	0.150	0.2080	0.067
$Text\,RF(10)\,TF(10)$	0.1536	0.123	0.1700	0.049
$Text\,RF(10)\,LF(in,10)$	**0.2098**	**0.170**	0.2200	0.068
$Text\,RF(10)\,LF(out,10)$	0.2053	0.168	**0.2260**	**0.069**
$Anchor$	**0.2780**	**0.257**	**0.2460**	**0.082**
$Anchor\,TF(10)$	0.2665	0.250	0.2380	0.079
$Anchor\,LF(in,10)$	0.2442	0.233	0.2060	0.066
$Anchor\,LF(out,10)$	0.2373	0.236	0.2080	0.071
$0.7\,Text + 0.3\,Anchor$	0.2459	0.223	0.2420	0.083
$0.7\,Text + 0.3\,Anchor\,TF(10)$	0.2363	0.209	0.2280	0.075
$0.7\,Text + 0.3\,Anchor\,LF(in,10)$	**0.2719**	**0.244**	**0.2640**	**0.090**
$0.7\,Text + 0.3\,Anchor\,LF(out,10)$	0.2593	0.229	0.2540	0.086

Table 5: Percentage of sub-topics (macro average) for which at least one relevanat document is found at different rank cut-offs.

| | Top | | | |
Run	10	20	100	1000
$Text$	16.3	26.1	41.0	51.4
$Text\,TF(10)$	16.8	23.5	40.6	51.4
$Text\,LF(in,10)$	19.4	26.6	40.6	51.4
$Text\,LF(out,10)$	20.3	29.2	40.7	51.4
$Text\,RF(10)$	21.4	27.4	41.3	51.3
$Text\,RF(10)\,TF(10)$	18.4	27.2	41.4	51.3
$Text\,RF(10)\,LF(in,10)$	22.0	33.0	40.9	51.3
$Text\,RF(10)\,LF(out,10)$	23.3	33.3	41.4	51.3
$Anchor$	28.5	34.2	44.7	52.0
$Anchor\,TF(10)$	27.2	33.7	43.9	52.0
$Anchor\,LF(in,10)$	25.9	32.6	45.2	52.0
$Anchor\,LF(out,10)$	28.2	32.2	44.7	52.0
$Text + Anchor$	27.2	34.8	50.2	59.3
$Text + Anchor\,TF(10)$	25.3	32.7	50.5	59.3
$Text + Anchor\,LF(in,10)$	29.4	37.1	50.5	59.6
$Text + Anchor\,LF(out,10)$	27.8	35.4	50.1	59.6
$Web\,only$	15.1	24.8	40.9	50.4
$Wikipedia$	8.7	8.7	11.1	12.6

4 Entity Ranking

For the entity ranking track, we have experimented with different approaches, which are discussed in this section. We use *anchor text* representations (assuming the entity's name will be frequent in incoming anchors); *co-citations* (assuming similar entities will receive similar incoming links); and use *Wikipedia as a pivot* (assuming entities have unique Wikipedia pages, which are neatly organized and may contain external links toward the most suitable homepage).

4.1 Anchor Text

Our first approach tries to apply an ad hoc retrieval method to the task of related entity finding. We use the ClueWeb Anchor text index that is described in Section 2. Queries consist of the concatenation of the entity name and the narrative. The initial result ranking is in the ad hoc format. To convert the results to the entity ranking format, we use a very naive approach. In our official run the first 300 results of the initial ranking are grouped into groups of three. Each result entity consist of a group of three pages, where each page is an entity homepage. If Wikipedia results occur in the initial ranking, they are added to the result entities ordered by score. We show additional results, where each result entity consists of only one web and one Wikipedia page

4.2 Co-citations

For the Entity Ranking topics an example relevant entity is provided. Given the large link graph of the ClueWeb collection, we want to exploit co-citation information to find

entities similar to the example entity. For this, we first find the set S of all pages s that link to the example entity e. For each page s, we consider all outgoing links as pointers to pages t about possibly similar entities. The number of pages in S that link to a target page t is the co-citation frequency of t and e. The more t and e are co-cited, the more similar they are. We consider the links from pages with a small number of outgoing links to be more valuable than links from pages with a high outgoing link degree. Thus, we weight each link from a page s to page t by the outgoing link degree of s. More formally, the similarity score between a target entity t and example entity e is given by:

$$sim(t, e) = \sum_s l(e \leftarrow s) \sum_t \frac{l(s \rightarrow t)}{outdegree(s)} \qquad (3)$$

where $l(s \rightarrow t)$ is 1 if there is a link from s to t and 0 otherwise. The entities are then ranked by their similarity score $sim(t, e)$.

4.3 Wikipedia

Our last approach exploits the information in Wikipedia. To complete the task of related entity finding, we take a number of steps.

1. Rank all Wikipedia pages according to their match to the entity name and narrative.

2. Scores of Wikipedia pages which belong to the correct target category (i.e. Persons, Products or Organizations) are boosted.

3. To find primary result pages, we follow the external links on the Wikipedia page to find matches with the Clueweb Category B URLs.

The second step is optional. We have made two official runs: excluding (Wiki Base) and including (Wiki Cats) the second step. More detail on the category mappings used in the second steps follow below. In our official runs, in the third step all Wikipedia pages without matches to the Category B URL's are dropped from the ranking. We made an additional run where Wikipedia pages without Clueweb links are retained in the ranking and a dummy Clueweb page is inserted in the result to make them the right format.

4.3.1 Category Mapping

In the Wikipedia context we consider each Wikipedia page as an entity. The Wikipedia page title is the label or name of the entity. Currently in the English part of Wikipedia there are over 3 million pages. Wikipedia employs a fine grained categorisation system, consisting of more than 70.000 categories. Each page is categorised into at least one category. The categories form a hierarchical structure, but because subcategories can have more than one parent, the structures as a whole is not a tree, but rather a directed acyclic graph.

In the entity ranking track only three high level types of entities are used: persons, products and organisations. 'Persons' is a clearly defined concept. Organisations and products on the other hand are less clearly defined. In the training topics certain groups of people, i.e. a band, or more abstract concepts like 'Motorsport series that Bridgestone officially supports with tyres' are included as organisations. A problem with the 'Products' entity type is the granularity, different versions of a product might have their own homepage, which makes them undesirable eligible as an entity.

To map the entity types to Wikipedia categories, we experiment with the following approach. We manually map a number of lower level Wikipedia categories to each entity type. Each document gets a binary score, either the document categories include one of the target categories or not. All documents including one of the target categories are ranked above all documents not including one of the target categories. The entity types are mapped to the following categories:

- Persons
    - 'Living People'
    - Ending with 'births'
    - Ending with 'deaths'
    - Starting with 'People'

- Organizations
    - Starting with 'Organizations'
    - Starting with 'Companies'

- Products
    - Starting with 'Products'
    - Ending with 'introductions'

4.4 Results

In this section we discuss the results of our official runs and some additional runs. We report the results in the official measures of the track: NDCG@R and P@10. The official NDCG@R score also credits relevant pages, i.e. pages that are related to the query topics without being actual homepages for the entities [1]. We have calculated another NDCG@R score that credits only the primary pages, i.e. the (authorative) entity homepages. Each evaluation measure is calculated for Wikipedia pages only, homepages only, and for their combination where both Wikipedia pages and other homepages are considered. In the assessments we have substituted redirected Wikipedia pages with the single non-redirected page. All results are presented in Table 6.

The Wikipedia based runs retrieve the most primary pages, homepages as well as Wikipedia pages. The Wikipedia

Table 6: Entity Ranking Results

Evaluation	Measure	Pages	Anchor Text		Co-citations	Wiki Base		Wiki Cats
			Groups of 3	Groups of 1		Only links	Dummy pages	Only links
Primary	# Pages	WP	22	22	43	56	**81**	57
		HP	19	18	23	**40**	22	**40**
		All	41	40	65	96	**101**	97
Primary	P@10	WP	0.0300	0.0300	0.0200	0.1000	0.1200	**0.1550**
		HP	0.0450	0.0100	0.0400	0.0500	0.0300	**0.0550**
		All	0.0700	0.0350	0.0600	0.1200	0.1250	**0.1650**
Primary	NDCG@R	WP	0.0427	0.0427	0.0246	0.0896	0.1090	**0.1091**
		HP	0.0495	0.0211	0.0515	**0.0746**	0.0319	0.0465
		All	0.0685	0.0436	0.0611	0.1059	0.1125	**0.1138**
All	NDCG@R	WP	0.1646	0.1653	0.0504	0.1762	**0.1977**	0.1665
		HP	**0.1773**	0.1625	0.1265	0.1043	0.0880	0.0805
		All	0.1820	**0.1828**	0.1397	0.1328	0.1425	0.1187

base run with only links throws out a number of primary Wikipedia pages that do not have a link to a Clueweb page. The run with the dummy Clueweb pages finds more primary Wikipedia pages, but less primary pages are found, because of the insertion of dummy pages. Also, since the run with the dummy pages is not an official run, more pages are un-judged.

The difference between the anchor text runs with groups of 3 pages in one result, and one page in each result, is very small. The run with groups of 3 has a higher P@10, more relevant results are found among the top ranked documents. However, since a result with more than one relevant page is rewarded the same as a result with just one relevant page, some relevant pages in the grouped run will be redundant and not rewarded.

The Co-citations run is the only run that makes use of the given entity in the topic to which the result entities should be related. By calculating similarity scores to the given entity we are able to find a reasonable number of primary Wikipedia and homepages. The link information does provide useful information on the relationships between enti-ties. In future work we would like to investigate if this link information can be combined with the using Wikipedia as a pivot approach and achieve additional improvement.

Looking at P10, the Wikipedia run that reranks pages according to their categories scores best. When we compare the base Wikipedia run and the category run, 13 topics get the same score, on 3 topics the base run is best, and on 4 topics the category run scores best. With the small number of topics in the test set, the average score can be influenced by just a few topics. Another issue is that 13 out of the 20 topics have less than 10 primary Wikipedia pages, and 14 out of the 20 topics have less than 10 primary homepages. So there is already a certain upper limit which is less than 1 for P10.

Looking at the official measure of the task, NDCG@R evaluated on web pages, our anchor text run outperforms all other runs, even though it finds the smallest number of primary pages. The anchor text run finds a large number of relevant pages, which is why it performs so well on this measure. We are more interested however in the retrieval of primary pages only, looking at the NDCG@R scores for primary home pages, none of our runs score very well. The baseline Wikipedia run with only links scores best with an NDCG@R of 0.0746. When we also credit Wikipedia pages, the scores improve somewhat.

Besides evaluating the runs on the official test data, we evaluate some separate steps in our approach to analyse where our approach could be improved. For our runs we discarded the given entity information. The given entities were identified by a non-Wikipedia homepages. Since we want to exploit the structured information in Wikipedia this was not very helpful. So, instead we try to match the given entity to a Wikipedia page. We try two approaches: we use the entity name as a query to the full text Wikipedia index and retrieve the top results, and we do exact string matching, where all characters are lower cased and special characters are removed. Results can be found in Table 7.

Table 7: Entity Finding Results

Results	Index	String Matching
None	0	7
Irrelevant	5	1
Relevant	13	0
Primary	2	12

Besides the 12 exact string matches, 3 more primary Wikipedia pages can be found with a less strict matching algorithm, so for 15 out of the 20 entities we can find a pri-mary page in Wikipedia. The top retrieved result from the index returns a relevant result for the majority of topics, but only few primary pages are identified. Using exact string

matching, we can find primary Wikipedia pages with high precision for a majority of the topics. Using the links and categories in Wikipedia from and to the primary homepages, can provide additional information that can help solve the entity relationship search task.

Finally, we want to evaluate the last step in our Wikipedia based approach, i.e. finding a primary homepage on the web for a primary Wikipedia page. To measure how well we do here, we use the 15 primary Wikipedia pages in combination with the given primary homepages in the topic specificiation as our testset. To find primary homepages we look at two specific parts of the Wikipedia page, the website specified in the 'Infobox', and the links specified in the 'External Links' section. For 5 entities both the 'Infobox' website and the first 'External link' point to the given entity site. For 2 entities the 'Infobox' website is correct, and for 5 more entities the 'External links' point to the correct site. In all but one case the first external link is the correct link. Only 2 entities do no have any link to the given entity site. Although this test set is small, our performance is promising, for the large majority of the pages we can find a correct link to a homepage on the Web. Our Wikipedia runs on the entity ranking task, find much more Wikipedia primary pages than primary pages on the Web. This has two reasons. First of all, the Clueweb Category B collection is of a limited size, and does not include a large part of the pages linked to from Wikipedia. When the complete Clueweb collection will be considered, this problem could become significantly smaller. Secondly, a lot of the Web homepages are not judged, especially the pages in our unofficial runs. Judging more pages would make this test set more reusable.

5 Conclusions

In this paper, we detailed our official runs for the TREC 2009 Web Track and Entity Ranking Track and performed an initial analysis of the results. We now summarize our preliminary findings.

We experimented with indexes of different document representations and a sliding window filter to combine text-based ranking with diversity features. Assuming a user starts reading the results list from the top and has seen the first m documents, we choose from documents $m + 1$ to $m + n$ in the text-based ranking the one that has the highest diversity score using some feature, add it to the final results list at rank $m + 1$ and slide down the window to ranks $m + 2$ to $m + n + 1$. As diversity features we consider the number of incoming links not seen in higher ranked results and the number of distinct terms not seen in higher ranked results.

For the initial text-based run, anchor text is very effective as it has more relevant documents in the top 20 ranks than standard full-text runs, which cover more diverse aspects of the search topic. The sliding window filter shows that link information is more effective than the number of unseen words to diversify retrieval results. The expection is

the anchor text run, which already implicitly uses link information through the length prior. For runs using the document text, or a combination of document text and anchor text, the incoming link filter increases the number of sub-topics covered by the top ranked results.

The initial document text-based run covers 0.84 sub-topics in the top 10 and 1.34 sub-topics in the top 20, on average. With a sliding window of size 10, which allows results to move up 9 ranks at the most, the lack of diversity in the top 20 limits the impact the sliding window filter can have on the diversity. To have more impact, the size of the window could be increased, but with such low precision scores, this also increases the chances of infiltration of very long or highly connected but off-topic pages. As the size of the window increases, the impact of the initial text-based ranking decreases. The impact of window size will be addressed in future research.

For entity ranking we experimented with three approaches: using anchor text representations (assuming the entity's name will be frequent in incoming anchors); using co-citations (assuming similar entities will receive similar incoming links); and using Wikipedia as a pivot (assuming entities have unique Wikipedia pages, which are neatly organized and may contain external links toward the most suitable homepage).

From our experiments we can draw the following conclusions. Anchor text works well for finding relevant Web pages, but not so well for finding primary Web pages. The link information used to make the co-citations run, does provide clues to find primary homepages, but just estimating similarity scores for the given entities is not sufficient. Using Wikipedia as a pivot works well for finding primary Wikipedia pages. Additionally, the links to the Clueweb collection from the 'Infobox' and 'External links' section of Wikipedia may be sparse, but the precision of the linked Clueweb pages is very high. Using the high level category information leads to improvements mainly in early precision.

From this first year of the Entity Ranking track we learn that link information is very important: anchor text can be used to find relevant pages, co-citations can be used to find similar entities, and links from Wikipedia to the Web can be used to find primary homepages. Secondly, Wikipedia is an excellent entity repository for this task. It covers the a large range of possible entity ranking topics, and its structure can be used to effectively rank entities.

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.

[2] N. Craswell, D. Hawking, and S. E. Robertson. Effective site finding using link anchor information. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 250–257. ACM, 2001. ISBN 1-58113-331-6.

[3] 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: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, pages 245–251, 2008.

[4] G. Demartini, A. P. Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 entity ranking track. In *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008*, pages 243–252, 2009.

[5] D. Hawking. Overview of the TREC-9 web track. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 87–102. National Institute for Standards and Technology. NIST Special Publication 500-249, 2001.

[6] Indri. Language modeling meets inference networks, 2009. http://www.lemurproject.org/indri/.

[7] J. Kamps. Effective smoothing for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication 500-266, 2006.

[8] J. Kamps. Experiments with document and query representations for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *The Fifteenth Text REtrieval Conference (TREC 2006)*. National Institute of Standards and Technology. NIST Special Publication 500-272, 2007.

[9] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, New York NY, USA, 2002.