Experiments with Result Diversity and Entity Ranking: Text, Anchors, Links, 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. Anchor text is also very effective for entity ranking. Using Wikipedia as a pivot results in a gain of precision, but at the cost of a loss of recall.

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). 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 standard retrieval methods using either a full-text index, or a propagated anchor-text index.

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 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 [3] 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 [4, 5].

For ad hoc search, pages with more text have a higher prior probability of being relevant [6]. 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 [1, 6] but for ad hoc search it seems less useful [2, 5]. 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

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: 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**

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 [4, 5] 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 mean P@30 might be an over-estimation. We removed the inclusion probability column from the official *prels* and used standard *trec_eval* to see if the traditional P@30 measure gives similar results and found that the anchor text run has a much lower P@30 than the full-text run. Of course, with the pooling approach tailored for the statistical measures, these scores are also not a reliable, but give a lower bound to the actual score. 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 redundant 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 2. 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 3. 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 document ranking. To see if differences observed in the scores of the diversity measures are caused by a better document 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. Over-

Table 3: 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**
$AnchorTF(10)$	0.2665	0.250	0.2380	0.079
$AnchorLF(in, 10)$	0.2442	0.233	0.2060	0.066
$AnchorLF(out, 10)$	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

all, 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 precision 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 subtopics for which relevant documents are found. In Table 4 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 3 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

Table 4: 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

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.

4 Entity Ranking

Due to the last-minute availability of the results, we can only provide an initial discussion, and refer to the final proceeding for more details and further experiments. For the entity ranking track, we have experimented with different approaches which are discussed in this section: using *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 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).

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. 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.

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 $sim(t, e)$. Note that this run uses only the example entity and the ClueWeb link graph. No content-based feature is used.

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 both runs excluding (Wiki Base) and including (Wiki Cats) the second step. More detail on the category mappings used in the second steps follow below.

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 two different methods. In our first method 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'

The second method exploits Wikipedia category hierarchy. We map the entity types to the most general matching Wikipedia categories. All subcategories down the hierarchy of the chosen Wikipedia categories are also considered relevant. The degree of relevance is expressed in the distance to the target category, i.e. how many levels in the hierarchy separate the document category from the target category. For each document we take the minimum distance of the distances to all its categories. Entity type 'Persons' can be mapped directly to Wikipedia category 'People'. Similarly, entity type 'Organizations' can be mapped directly to the category 'Organizations'. The 'Products' entity type cannot be mapped directly to one Wikipedia category, instead we map it to the categories 'Product by Company' and 'Introductions by year'.

4.4 Results

We focus our discussion of results on our official runs. More variations and runs using the Wikipedia category hierarchy will be discussed in the final proceedings paper. We report the results of our official runs in NDCG@R and P@10 in Table 5. The score based on the 'Primary Home Page evaluation' counts only primary, non-Wikipedia pages as relevant, whereas the score based on the 'Wikipedia evaluation' also gives credit for retrieved Wikipedia pages.

The first thing to notice is that the performance of the runs is rated very differently by the two evaluation measures: NDCG@R and P@10. The best run considering P@10 'Wiki Cats', is the worst run considering NDCG@R, and vice versa, the best run considering NDCG@R 'Anchor Text' is almost the worst run considering P@10.

As was to be expected the runs using Wikipedia benefit most when the evaluation also gives credit to retrieved Wikipedia pages, e.g. P@10 for the 'Wiki Cats' run triples from 0.0550 to 0.1650 when the Wikipedia evaluation is used. Looking at NDCG@R however, the 'Anchor Text' run still considerably outperforms all other methods.

The Wikipedia runs have a lower recall, because a considerable part of the correct related entity homepages will not be linked to from Wikipedia, which gives these runs a lower NDCG@R. For early precision P@10 however, the Wikipedia runs are much better. By considering the entity types, which is done in the 'Wiki Cats' run, early precision improves even further, but the cost is a lower NDCG@R.

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

Table 5: Entity Ranking Results Official Runs

Evaluation	Measure	Anchor Text	Co-citations	Wiki Base	Wiki Cats
Primary HP	P@10	0.0450	0.0400	0.0500	**0.0550**
	NDCG@R	**0.1773**	0.1265	0.1043	0.0805
WP	P@10	0.0700	0.0600	0.1200	**0.1650**
	NDCG@R	**0.1823**	0.1401	0.1324	0.1208

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).

Anchor text works well on finding primary Web pages. Although early precision is low, it has much better recall than the runs using co-citations information or Wikipedia categories. Co-citations are less effective For precision, the Wikipedia category mappings are more effective than the co-citations and the anchor text. Although the difference is small when considering the ranking of the primary home pages, if we look at the ranking of the Wikipedia pages, the Wikipedia categories give much more precise results. Since the anchor text, co-citation and Wikipedia category runs are very different, each with different strengths, we will look into ways of combining these sources of evidence to create a ranking with the high precision of the Wikipedia category run and the high recall of the anchor text run.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513, 639.072.601, and 640.001.501).

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.