Heuristic Ranking and Diversification of Web Documents

Jiyin He Krisztian Balog Katja Hofmann Edgar Meij
Maarten de Rijke Manos Tsagkias Wouter Weerkamp

ISLA, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: We describe the participation of the University of Amsterdam's Intelligent Systems Lab in the web track at TREC 2009. We participated in the adhoc and diversity task. We find that spam is an important issue in the ad hoc task and that Wikipedia-based heuristic optimization approaches help to boost the retrieval performance, which is assumed to potentially reduce spam in the top ranked results. As for the diversity task, we explored different methods. Clustering and a topic model-based approach have a similar performance and both are relatively better than a query log based approach.

1 Introduction

This year's Web track consists of two tasks, the *ad hoc* and *diversity* task. The ad hoc task is traditional ad hoc retrieval in a web setting, where the goal is to return a list of documents from a static document collection, ranked by decreasing relevance. Document relevance is considered independent from the rest of the documents within the list.

The second task, diversity, is new; the goal is to return a ranked list of documents which *together provide complete coverage of a query, while avoiding excessive redundancy in the result list*. Here, in contrast to the ad hoc task, a document's relevance is dependent on the presence of other documents in the same ranked list.

For the ad hoc task we explore two basic approaches: (i) query rewriting using Markov Random Fields to get a better representation of the original query, and (ii) remodeling the query using an external collection as source for adding and reweighing query terms. On top of these approaches we (i) remove Wikipedia pages that are not content pages (e.g., category pages, link-to pages, etc.), and (ii) promote Wikipedia pages in the initial ranked list to the top of the ranking.

For the diversity task we experiment with three types of method for result diversification, an AOL query log based approach, a topic model based approach as well as a clustering based approach. We find that the topic model based approach and clustering based approach outperform our baseline run, i.e., the MRF ad hoc run as measured by diversity metrics.

In the remainder of this paper, we first describe the approaches we applied to both tasks in Section 2. Then we describe our experimental settings in Section 3, the results and a discussion of our submitted runs in Section 4. Section 5 concludes the description of our participation in this year's Web track.

2 Methods

In this section, we describe our proposed approaches to the ad hoc task, in Section 2.1 and to the diversity task (Section 2.2).

2.1 Ad hoc Task

The goal of the ad hoc task can be considered as one of the most basic ones in IR: to rank documents according to their relevance to a given query. Despite its "standard" character, the nature and the size of the new Clueweb collection render the task challenging and interesting.

We do not apply spam filtering on the collection, although insights from preliminary data exploration suggested that retrieval results may benefit substantially for any ad hoc retrieval system on this collection. For now, we try two basic approaches and use two optimization techniques. Below we describe the two approaches and the optimizations.

Markov Random Fields Following the ideas from (Metzler and Croft, 2005), we use Markov Random Fields (MRF) to rewrite our initial query. The goal of applying this technique is to be better able to represent possible phrases in the query. A three term query like "obama white house" would result in all possible phrases (e.g., "obama white," "white house," "obama house," and "obama white house") as well as the single terms (Mishne and de Rijke, 2005). Previous TREC years showed that this technique is very effective.

External expansion Given that we are dealing with a web collection that can be quite noisy, we use *external query expansion*, a technique that proved useful in retrieval in the blogosphere (Arguello et al., 2008; Weerkamp et al., 2009; Weerkamp and de Rijke, 2009)) and that originated from targeting better relevance model estimations (Diaz and Metzler, 2006). The goal of this technique is to use an "external" collection that is less noisy than the target collection to improve the estimation of the query model.

The Clueweb collection offers a natural "external" collection, Wikipedia, as this part of the collection is free of spam and relatively clean (compared to other web documents). Wikipedia would therefore be usable in modeling our query: We run our queries against the Wikipedia collection, select the top 10 terms (using relevance models from (Lavrenko and Croft, 2001) and mix these with the original query terms.

Optimizing our approaches We use two ways of optimizing our runs: (i) Wikipedia filtering, and (ii) Wikipedia promotion. The first technique is used to filter out Wikipedia pages that do not contain real content. These pages are for example the link-to, category, and disambiguation pages that are mainly included for navigational purposes. We feel that these pages can be removed without danger of missing relevant documents, thereby possibly pushing relevant documents higher up the ranking. The second technique, Wikipedia promotion, is based on the observation that Wikipedia pages are pages we can certainly trust, whereas other web documents could very well be spam. We translate this observation into the promotion of all Wikipedia pages in the results to the top of the ranking (maintaining their relative order).

Our three final runs for the ad hoc task use: (i) Markov Random Fields and Wikipedia filtering, (ii) Markov Random Fields and Wikipedia filtering and promotion, and (iii) External expansion and Wikipedia filtering. We report on the results of the runs in the next section.

2.2 Diversity Task

For the diversity task, we experimented with 3 types of approach: *Single Pass Clustering* (SPC), a *topic model-based approach*, and *AOL query suggestion*. The first two approaches share common features: they re-rank an initially retrieved list of documents for generating the final result list, and try to model the topical facets contained in the initial retrieved ranking list without using external resources. The difference between the two approaches mainly lies in the methods used for detecting topics and for re-ranking. For topic detection, the first approach, SPC, clusters documents into a number of topics and each document is assigned to one topic, while the topic model-based approach represents each document as a mixture of topics. For re-ranking, the SPC approach selects documents from different clusters so that selected documents are supposedly about different topics, while the topic model-based approach tries to maximize the probability that most if not all topics being present in the selected document list. The third approach, *AOL query suggestion* uses an external resource, i.e., AOL query logs, for modeling the topical facets of a query. It also generates the final result list in a different fashion which will be further described below.

Single Pass Clustering The first method we employed is Single Pass Clustering (SPC) (Hill., 1968), which provides not only an efficient clustering algorithm, but also mimics a reasonable heuristic that a user might employ. That is, start at the top and work down the initial retrieved list of documents, and assign each to a cluster. The process for assignment is performed as follows: The first document is taken and assigned to the first cluster. Then each subsequent document is compared against each cluster with a similarity measure (in our case a standard cosine measure using a TF.IDF weighting scheme). A document is assigned to the most likely cluster, as long as the similarity score is higher than a certain threshold (set to 0.2 for our run); otherwise, the document is assigned to a new cluster.

Once this clustering has been performed on the initial result list, we re-rank documents as follows. First, we output a single document from each cluster, specifically, the ones that were ranked the highest initially. Second, we iterate over the initial list of documents, and output each that has not been returned in the first phase.

Topic Model Approach This approach is inspired by previous work on diversifying a ranked list with Maximal Marginal Relevance (MMR) by Carbonell and Goldstein (1998) and based on a topic modeling approach, i.e., LDA (Blei et al., 2003). It treats the reranking problem as a procedure of selecting a sequence of documents, where a document is selected depending on both its relevance with respect to the query and the documents that have already been selected before it, so as to have a set of documents that (i) are most relevant to the query and (ii) represent most if not all topical aspects.

We proceed as follows. First, we use LDA to extract 10 topics from the top 2,500 documents in the initial retrieved set of results, where the initial results are generated from the ad hoc run *uvamrf* as described above, and each document can be represented as a mixture of 10 topics. On top of that, we start the re-ranking procedure by selecting the top relevant document in the initial list as the first document in the new ranked list. Then, we select a next document that can maximize the expected joint probability of presence of all topics in the selected result set. Since the sum of topic proportions within a document equals 1, the maximum joint probability (i.e., product of the probabilities of presence of each topic) occurs when the topics have equal proportion in

the selected set. On the other hand, we use the retrieval score from the initial run as a prior probability that a document is selected as the next one, so as to take into account the relevance relation between the document and the original query.

Formally, given a query q, a set of candidate documents $Ca = \{d_j\}_{j=1}^{n}$ and a set of latent topics $T = \{t_i\}_{i=1}^{m}$, a document is selected from Ca for inclusion in the ranked list S such that

$$\arg\max_{d \in Ca} \quad P(q|d) \prod_{i=1}^{m} P(t_i \in S \cup \{d\}), \tag{1}$$

where $P(q|d)$ is the query likelihood between the query q and document d calculated as in a standard language modeling framework. The term $P(t_i \in S \cup \{d\})$ denotes the probability of a topic being present in the set $S' = S \cup \{d\}$, which is estimated by

$$P(t_i \in S') = \sum_{d_j \in S'} P(t_i \in d_j) P(d_j). \tag{2}$$

AOL — Diversification using query logs This approach employs queries from a query log to discern and obtain diverse query formulations. The intuition is that terms that are frequently queried in conjunction with a set of given query terms provide a diverse set of *aspects* of those given query terms. We proceed as follows. First, we normalize all the queries in the AOL query logs and remove web addresses and non-alphabet characters. We then look up for each test topic whether it appeared as a phrase in the query logs. If so, we take the top 25 queries with a minimum number of occurrences of 5. An example is given in Table 1.

Topic	AOL query	Frequency
dinosaurs	remote control dinosaurs	30
dinosaurs	jim henson dinosaurs	25
dinosaurs	allosaurus dinosaurs	24
dinosaurs	flying dinosaurs	21
dinosaurs	walking with dinosaurs	16

Table 1: Example of using the AOL query logs for diversification.

For each of these "expanded" queries we generate an MRF query and, on the basis of this, a new ranking. Each of these ranked lists now represents a ranking of documents based on one aspect of the initial topic. In order to arrive at a final ranking, the lists are merged. We do so by first sorting them by aspect occurrence frequency (as found in the query log) and then adding the highest ranked document that has not been selected yet to the final ranking in a round-robin fashion.

3 Experimental Settings

3.1 Data

For both tasks in the Web track, we use the category A set of the Clueweb collection (the full collection). For indexing, we do not use any form of stemming and remove a conservative list of 588 stopwords. We index the headings, titles, and contents as searchable fields and do not remove any HTML tags. Our approaches retrieve against the text content of the web pages and leave out information provided by anchor texts or hyperlinks among web pages.

3.2 Evaluation metrics

For the ad hoc task, we use the traditional relevance oriented measures, i.e., MAP, P@5 and P@10. For the diversity task, the results are evaluated with the α-NDCG measure as proposed by Clarke et al (2008) and IA-Precision as developed in this year's Web Track (Clarke et al., 2009). The latter two metrics allow for measuring both the relevance and the novelty of the result ranked list.

3.3 Significant test

We use the Wilcoxon signed-rank test to test for significant differences between runs. We report on significant increases (or drops) for $p < .01$ using $^{\blacktriangle}$ (and $^{\blacktriangledown}$) and for $p < .05$ using $^{\triangle}$ (and $^{\triangledown}$).

4 Results and Discussion

4.1 Adhoc Task

The results of our adhoc runs are displayed in Table 2. We observe that the run using Wikipedia promotion outperforms the other two runs significantly. The difference with its baseline, MRF with just filtering, is huge, especially on the precision metrics. Comparing the two approaches, external expansion and MRF, in their "baseline" setting, we see a marginal advantage for external expansion, but differences are not significant.

Figure 1 shows the per-topic level run comparison between the MRF run and the Wikipedia promotion run. We see that most topics are helped and only a small portion of the topics are slightly hurt by Wikipedia promotion, which indicates that Wikipedia is a reliable resource for web retrieval, and is probably due to the fact that Wikipedia does not contain any spam.

Approach	MAP	P10	MRR	runID
EE + filter	0.0682	0.1100	0.1627	uvaee
MRF + filter	0.0626	0.0940	0.1255	uvamrf
MRF + filter + prom.	**0.1092**$^{\blacktriangle}$	**0.4100**$^{\blacktriangle}$	**0.5272**$^{\blacktriangle}$	uvamrftop

Table 2: Results of our submitted runs for the ad hoc task.

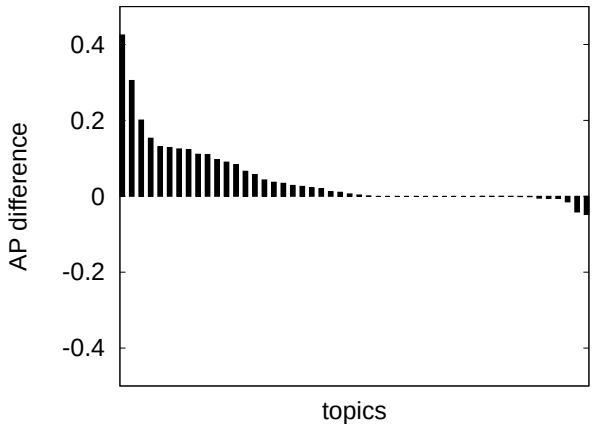

Figure 1: Per-topic comparison of AP between uvamrf and uvamrftop.

Approach	α-ndcg@5	α-ndcg@10	α-ndcg@20
uvamrf	0.042	0.060	0.076
uvamrftop	**0.123**$^{\blacktriangle}$	**0.129**$^{\triangle}$	**0.139**
AOL	0.055	0.074	0.098
SPC	0.068	0.093	**0.127**$^{\triangle}$
TM	**0.090**$^{\triangle}$	**0.097**	0.125$^{\triangle}$

Approach	IA-P@5	IA-P@10	IA-P@20
uvamrf	0.020	0.028	0.035
uvamrftop	**0.090**$^{\blacktriangle}$	**0.089**$^{\triangle}$	**0.079**$^{\triangle}$
AOL	0.023	0.030	0.037
SPC	0.036	**0.043**$^{\triangle}$	**0.051**
TM	**0.047**$^{\triangle}$	0.041	0.043$^{\triangle}$

Table 3: Result of diversity task. The names of approaches correspond to AOL query suggestion (AOL), single pass clustering (SPC) and topic model based approach(TM). Results of diversification runs are compared to the baseline run MRF.

4.2 Diversity Task

Table 3 shows the results of our submitted runs for the diversity task. The results can only be considered indicative considering the heuristic selection of parameter values. Nevertheless, we observe that SPC and the topic model based methods display a similar performance. Intuitively, this is likely due to the common features shared during the topic detection process: given that LDA can also be seen as a method for clustering, the resulting clustering/topic structure may be similar. However, in order to gain insight into the similarities and differences in behavior of the two approaches, further comparison and analysis are needed.

On the other hand, when we the results compare to the initial ranked list, i.e., the MRF run, we see that all methods outperform the baseline, where the improvements of SPC and the topic modeling-based method are significant. More interestingly, we see that if we evaluate the MRF run with Wikipedia promotion with the diversity metrics, the performance is better than all three diversification methods. The Wikipedia promotion run retrieves more relevant documents at the top of the ranked list than the MRF run on which our submitted diversification runs are based.

5 Conclusion

In this year's web track, we submitted runs for the ad hoc task and diversity task. For the ad hoc task, we explored a basic retrieval approache, viz. Markov Random Fields for modeling query term proximity and external query expansion. On top of that, we applied two types of heuristic optimization, i.e., Wikipedia filtering and Wikipedia promotion. Combining the basic approaches with the optimization methods, we submitted three runs: (i) Markov Random Fields with Wikipedia filtering, (ii) Markov Random Fields with Wikipedia filtering and promotion, and (iii) External Expansion with Wikipedia filtering. Although we did not explicitly apply any spam filtering techniques, the preliminary results suggest that spam is an important issue for experiments based on the ClueWeb collection. For the diversity task, we explored three types of approach: (i) Single Pass Clustering, (ii) topic modeling, and (iii) diversification using a query log. All three methods outperform the baseline approach, i.e., the MRF run without diversification.

Although the results are not exactly comparable across methods, we were able to identify issues shared by all three methods. For example, the heuristic method for choosing parameters calls for systematic experiments that will allow us to gain further insights in to the algorithms' performance under different parameter settings. On the other hand, intuitively, the performance of the clustering and topic model-based methods depends heavily on the initial retrieval run used for re-ranking, which is an interesting issue for further analysis.

In addition, we found that our heuristic Wikipedia-based promotion technique results in high scores in terms of diversity metrics. The Wikipedia promotion retrieves more relevant documents at the top of the ranked list, while our other diversification runs our MRF-based baseline run based diversification runs in general have very few relevant docu-

ments being retrieved.

6 Acknowledgments

This research was supported by the DAESO and DuOMAn project carried out within the STEVIN program which is funded by the Dutch and Flemish Governments under project number STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802, and by the Virtual Laboratory for e-Science project, which is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science and is part of the ICT innovation program of the Ministry of Economic Affairs.

7 References

Arguello, J., Elsas, J., Callan, J., and Carbonell, J. (2008). Document representation and query expansion models for blog recommendation. In *ICWSM 2008*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, New York, NY, USA. ACM.

Clarke, C., Craswell, N., and Soboroff, I. (2009). Preliminary report on the TREC 2009 Web Track. In *TREC 2009 Notebook*.

Clarke et al, C. (2008). Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666. ACM.

Diaz, F. and Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, pages 154–161, New York, NY, USA. ACM.

Hill., D. R. (1968). A vector clustering technique. In Samuelson, editor, *Mechanised Information Storage, Retrieval and Dissemination*, North-Holland, Amsterdam.

Lavrenko, V. and Croft, B. W. (2001). Relevance based language models. In *SIGIR '01*. ACM.

Metzler, D. and Croft, W. B. (2005). A Markov random field model for term dependencies. In *SIGIR '05*, pages 472–479, New York, NY, USA. ACM.

Mishne, G. and de Rijke, M. (2005). Boosting Web Retrieval through Query Operations. In Losada, D. and Fernández-Luna, J., editors, *ECIR 2005*, pages 502–516.

Weerkamp, W., Balog, K., and de Rijke, M. (2009). A generative blog post retrieval model that uses query expansion based on external collections. In *ACL-ICNLP 2009*.

Weerkamp, W. and de Rijke, M. (2009). External query expansion in the blogosphere. In *Seventeenth Text REtrieval Conference (TREC 2008)*. NIST.