Result Diversification Based on Query-Specific Cluster Ranking (Abstract) *

Jiyin He, Edgar Meij and Maarten de Rijke
ISLA, University of Amsterdam
The Netherlands
{j.he, edgar.meij, derijke}@uva.nl

ABSTRACT

Result diversification is a retrieval strategy for dealing with ambiguous or multi-faceted queries by providing documents that cover as many potential facets of the query as possible. We propose a result diversification framework based on query-specific clustering and cluster ranking, in which diversification is restricted to documents belonging to a set of clusters that potentially contain a high percentage of relevant documents. Empirical results on the TREC 2009 Web track test collection show that the proposed framework improves the performance of several existing diversification methods, including MMR, IA-select, and FM-LDA. The framework also gives rise to a simple yet effective cluster-based approach to result diversification that selects documents from different clusters to be included in a ranked list in a round robin fashion.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms: Algorithms, Experimentation

Keywords: Result diversification, Query-specific clustering

1. INTRODUCTION

Queries submitted to web search engines are often ambiguous or multi-faceted in the sense that they have multiple interpretations or sub-topics. One retrieval strategy that attempts to cater for multiple interpretations of such a query is to *diversify* the search results. Without explicit or implicit user feedback or history, the retrieval system makes an educated guess as to the possible facets of the query and presents as diverse a result list as possible by including documents pertaining to different facets of the query within the top ranked documents.

Following the Cluster Hypothesis [6], query-specific cluster-based retrieval is the idea of clustering retrieval results for a given query, which was shown to improve retrieval effectiveness if one can place documents from high quality clusters (in which a relatively large fraction of documents is relevant) at the top of the ranked list [5]. In this paper, we consider a ranking approach based on query-specific cluster-based retrieval in the context of result diversification. Specifically, we propose to rank and select a set of high quality clusters and then apply diversification only to the documents within these clusters. We posit that such a strategy should lead to improved results as measured in terms of both relevance and diversity since it only diversifies documents that are likely to be relevant.

*The full version of this paper is accepted for publication by *Journal of American Society for Information Science and Technology*.

DIR'11 February 4, 2011, Amsterdam, the Netherlands
ACM X-XXXXX-XX-X/XX/XX.

2. METHOD

The overall goal of our approach is to rank query-specific clusters with respect to their relevance to the query and to limit the diversification process to documents contained in the top ranked clusters only, in order to improve the effectiveness of diversification as measured in terms of both relevance and diversity.

For a query q and a ranked list of top n documents D_q^n retrieved in response to q, we cluster D_q^n into K clusters. Assume that we have a ranking method $cRanker(\cdot)$ that ranks clusters with respect to their relevance to a query and a diversification method $Div(\cdot)$ that diversifies a given ranked list of documents. We propose the following procedure for diversification. The input of the procedure is the output of $cRanker$, that is, a ranked set of clusters $RC = c_1, \ldots, c_K$, and the documents contained in each cluster, D_q^c. A free parameter T is used to indicate the number of top ranked clusters to be selected for diversification. Furthermore, $dRanker(\cdot)$ is assumed to be a document ranker that ranks documents according to certain criteria, for example, ranking documents in descending order of their retrieval scores. The diversification procedure first applies $Div(\cdot)$ to the documents assigned to the top T ranked clusters; documents assigned to clusters ranked below the top T are ranked by $dRanker(\cdot)$ and appended to the ranked list of documents obtained from the top T clusters.

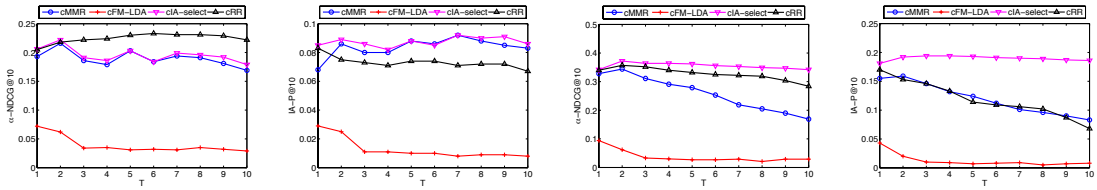
Clustering. We use latent Dirichlet allocation (LDA) [2] to cluster the initial retrieved ranked list. First, we train the topic models over D_q^n with a pre-fixed number of K clusters (or latent topics). A document d is then assigned to a cluster c^* such that

$$c^* = \arg\max_c p(c|d), \tag{1}$$

where $p(c|d)$ is estimated using the LDA model.

Diversification. For $Div(\cdot)$ we consider the following three diversification methods: Maximal Marginal Relevance (MMR) [3], Facet Model with LDA (FM-LDA) [4], and Intent Aware select (IA-select) [1]. In addition, we propose a cluster-based approach referred to as Round-Robin (RR). For this, we first rank the clusters according to their relevance of the clusters to the query. Then, documents within each cluster are ranked in the order of their original retrieved scores and, finally, we select documents belonging to different clusters in a round robin fashion.

Cluster ranking. For simplicity, we only discuss two ways to rank clusters that are necessary for investigating the effectiveness of our proposed framework for result diversification: *query likelihood* and *oracle*. For an input query, the query likelihood ranker ranks the clusters in descending order of the probability $p(c|q)$, which is inferred from the LDA model as described above. In other words, the clusters are ranked according to their likelihood given the query. Presumably, if a cluster has a high probability to generate a query, the documents contained in this cluster are more likely to be relevant to the query. Hence, the cluster is more likely to contain



(a) (b) (c) (d)

Figure 1: Diversification with cluster ranking using query likelihood ranker (1(a), 1(b)) and oracle ranker (1(c), 1(d)) over different numbers of selected top ranked clusters (T). K is set to 10 (30 and 50 show similar trends).

relevant documents. The oracle ranker, on the other hand, ranks the clusters using information from explicit relevance judgments. Here, the probabilities $p(c|q)$ are estimated using the judgments of retrieved documents in D_q^n, computed as

$$p(c|or a_q) = \frac{|D_q^c \cap D_q^R|}{|D_q^c|}. \quad (2)$$

where D_q^R are the documents judged to be relevant. That is, we rank clusters according to the number of relevant documents contained in them, normalized by the size of the cluster.

Determining the cut-off T. Automatically determining the optimal cut-off T is non-trivial. We typically do not have sufficiently many test queries to learn the optimal value of T, hence we apply leave-one-out cross-validation to find the optimal value of T for each query. Specifically, we optimize T over a set of training queries for a given K and a given diversification method for a given evaluation metric by exhaustive search, i.e., over all possible values of $T = 1, ..., K$. Then we apply the learned T on the test query.

3. RESULTS AND DISCUSSION

We apply our proposed diversification framework on the TREC 2009 Web track catB test collection. We use the Markov Random Field model (MRF) [7] to generate the initial ranked list and set $n = 1000$. We then conduct the LDA clustering on the initial ranked list, setting $K = 10, 30,$ and 50.

Figure 1 shows the trends of the performance of each diversification method with cluster ranking (cMMR, cFM-LDA, cIA-select and cRR) across values of T, the number of top-ranked clusters whose documents are used for diversification. For each method, when $T = K$, diversification with cluster ranking is equivalent to diversifying the complete list of initially retrieved documents. Here, we only show the results measured using α-NDCG@10 and IA-P@10; a similar trend can be observed for α-NDCG@X and IA-P@X, for X =5 and 20. We observe that with both the query likelihood and the oracle cluster ranker, diversification performance is hardly influenced by selecting all clusters, i.e., by diversifying the complete ranked list of documents. Also, for each method there is an optimal value of T that maximizes the performance of the method, the value of which is smaller than the total number of clusters, i.e., for which the optimal value of T satisfies $T < K$.

If we compare the query likelihood ranker to the oracle cluster ranker, we see that the retrieval performance fluctuates a lot as T increases in Fig. 1(a) and 1(b), that is, with many local maximums, while in Fig. 1(c) and 1(d), the performance curves are relatively smooth: they remain the same or decrease once an initial maximum has been reached. This implies that, with a near perfect ranking of clusters, we can find the global optimal T by simply adding documents belonging to a cluster ranked next, until the performance starts to decrease. On top of that, we clearly see that optimal results are achieved by selecting a small number of top ranked clusters.

Table 1 compares diversification with cluster ranking against diversifying the complete list of retrieved documents. cX indicates the runs with cluster ranking and selection, where X is the name of a diversification method. K is the total number of clusters. Here we only list the results from K that result in best performance for the original diversification method (i.e., without cluster ranking). We also list the average predicted value of T. On top of that, we include the performance achieved by each method when T is optimal, indicated by T^*. These values correspond to the peaks in Figure 1.

K	Method	α-NDCG@5	score avg. T	α-NDCG@10	score avg. T	IA-P@5	score avg. T	IA-P@10	score avg. T
10	MMR	0.122	–	0.169	–	0.066	–	0.083	–
	cMMR	**0.191**$^{\Delta}$	1.98	**0.216**	2.00	0.070	2.44	0.069	6.82
	cMMR$^{T^*}$	**0.191**$^{\Delta}$	2	**0.216**	2	**0.090**	2	**0.092**	7
10	FM-LDA	0.027	–	0.029	–	0.011	–	0.008	–
	cFM-LDA	**0.058**	1.00	**0.072**$^{\Delta}$	1.00	**0.031**$^{\Delta}$	1.00	**0.029**$^{\Delta}$	1.00
	cFM-LDA$^{T^*}$	**0.058**	1	**0.072**$^{\Delta}$	1	**0.031**$^{\Delta}$	1	**0.029**$^{\Delta}$	1
50	IA-select	0.146	–	0.193	–	0.078	–	0.092	–
	cIA-select	0.181$^{\Delta}$	15.06	0.208	27.14	0.100	31.36	0.092	23.54
	cIA-select$^{T^*}$	**0.199**$^{\Delta}$	9	**0.226**$^{\Delta}$	27	**0.105**$^{\Delta}$	32	**0.096**	23
10	RR	0.198	–	0.222	–	0.079	–	0.067	–
	cRR	0.199	2.68	**0.233**$^{\Delta}$	6.00	0.085	2.00	**0.083**	1.00
	cRR$^{T^*}$	**0.204**	2	**0.233**$^{\Delta}$	6	**0.091**	2	**0.083**	1

Table 1: Results of proposed diversification framework. $^{\Delta}$indicates a significant difference given by a paired t-test with p-value<0.05.

We observe that diversification with cluster ranking outperforms the original algorithms in nearly all cases, even though query likelihood is not a perfect ranker for ranking clusters and T has not been fully optimized. If we take the optimal T with respect to the average performance over all queries, i.e., T^*, we see further improvements, and more improvements are statistically significant compared to that of the predicted T.

Acknowledgements This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Center for Creation, Content and Technology (CCCT), and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.

4. REFERENCES

[1] R. Agrawal and S. Gollapudi and A. Halverson and S. Ieong. Diversifying search results. In *WSDM'09*, 2009.
[2] D. M. Blei and A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, 2003.
[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries In *SIGIR'98*, 1998.
[4] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM'09*, 2009.
[5] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *SIGIR'96*, 1996.
[6] N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. In *Information Storage and Retrieval*, 1971.
[7] D. Metzler and W. B. Croft. A Markov Random Field Model for Term dependencies. In *SIGIR'05*, 2005.