Integrating Multiple Resources for Diversified Query Expansion

Arbi Bouchoucha, Xiaohua Liu, and Jian-Yun Nie

Dept. of Computer Science and Operations Research
University of Montreal
Montreal (Quebec), Canada
{bouchoar, liuxiao, nie}@iro.umontreal.ca

Abstract. Diversified query expansion aims to cover different possible intents of a short and ambiguous query. Most standard approaches use a single source of information, e.g., the initial retrieval list or some external resource like ConceptNet. The coverage is thus limited by that of the resource. To alleviate this issue, we propose the use of multiple resources. More specifically, our framework first automatically generates a list of diversified queries for each resource, and then combines the retrieved documents for all the expanded queries following the Maximal Marginal Relevance principle. We evaluate our framework on several TREC data sets, and demonstrate that our framework outperforms the state-of-the-art approaches, suggesting the effectiveness of incorporating different resources for diversified query expansion.

Keywords: Diversified Query Expansion, Resource Integration.

1 Introduction

Queries in Web search are usually short, ambiguous (e.g., "Java") and under-specified [1]. To address this issue, various search result diversification (SRD) technologies have been proposed (see for example, [1, 4, 7]). Traditional SRD approaches are based on query expansion (QE) and pseudo-relevance feedback (PRF). One weakness of such approaches is that their performance much depends on the initial retrieval results.

Diversified query expansion (DQE) represents the most recent approach to SRD. One distinguished feature of DQE is the utilization of external resources, e.g., ConceptNet [8], Wikipedia, or query logs, to generate a set of diversified queries, whose retrieval results are then combined into a new list. One representative work of DQE is conducted by Bouchoucha et al. [3], which expands queries using ConceptNet and uses the Maximal Marginal Relevance (MMR) strategy [4] to select diversified terms. Their approach outperforms the state-of-the-art existing SRD methods on TREC data.

Following this work, we propose to combine multiple resources to diversify queries. Our approach is largely motivated by the following observation: there are a large number of queries for which ConceptNet cannot yield good performance

but some other resources can suggest good terms. For example, "defender", the #20 query from the TREC 2009 Web track, has six different subtopics[1], but traditional IR models for this query return no relevant documents. Therefore PRF is unable to suggest useful terms to expand the query. ConceptNet returns results covering subtopic 2, 3 and 6, while Wikipedia and query logs provide documents covering subtopic 1, 2, 3, 4 and 1, 2, 4, 5, respectively. By integrating all these sources, we obtain a list of documents covering all the subtopics.

We further propose a unified framework to integrate multiple resources for DQE. For a given resource (e.g., ConceptNet), our framework first generates expansion candidates, from which a set of diversified terms are selected following the MMR principle. Then the retrieved documents for any expansion query of any resource are combined and again the MMR principle is used to output diversified results. In this work, we integrate four typical resources, *i.e.*, ConceptNet, Wikipedia, query logs, and initial search results.

It is worth noting that the idea of combining multiple resources has been successfully exploited for other IR tasks. For example, Deveaud et al. [5] combine several external resources to generalize the topical context, and suggest that increasing the number of resources tends to improve the topical representation of the user information need. Bendersky et al. [2] integrate multiple information sources to compute the importance features associated with the explicit query concepts, and to perform PRF. Compared with these studies, our work has two significant differences: 1) these resources are used to directly generate diversified queries; and 2) MMR is used to cover as many aspects of the query as possible.

We evaluate our approach using TREC 2009, 2010 and 2011 Web tracks. The experimental results show that multiple resources do complete each other, and integrating multiple resources can often yield substantial improvements compared to using one single resource.

Our contributions are twofold: 1) we propose the integration of multiple resources for DQE, and a general framework based on MMR for the implementation; 2) we show the effectiveness of our method on several standard datasets.

The remainder of this paper is organized as follows. Section 2 presents details of our method. Section 3 describes our experiments. Finally in Section 4, we conclude with future work.

2 Proposed Framework

Our proposed framework consists of two layers. The first layer integrates a set of resources, denoted by R, to generate diversified queries. Given a query Q, it iteratively generates a good expansion term c^* for each resource $r \in R$, which is both similar to the initial query Q and dissimilar to the expansion terms already selected:

$$c^* = argmax_{c \in C_{r,Q}} (\lambda_r \cdot sim(c, Q) - (1 - \lambda_r) \cdot max_{c_i \in S_{r,Q}} sim_r(c, c_i)) \quad (1)$$

[1] `http://trec.nist.gov/data/web/09/wt09.topics.full.xml`

Here, $C_{r,Q}$ and $S_{r,Q}$ represent the set of candidate terms and the set of selected terms for r, respectively; the parameter λ_r (in [0,1]) controls the trade-off between relevance and redundancy of the selected term; $sim_r(c, c_i)$ returns the similarity score of two terms for resource r; $sim_r(c, Q)$ is the similarity score between term c and the query Q, which is computed using Formula 2, where q is a subset of Q and $|q|$ denotes the number of words of q.

$$sim_r(c, Q) = max_{q \in Q}(sim_r(c, q) \cdot \frac{|q|}{|Q|}) \tag{2}$$

Once c^* is selected, it is removed from $C_{r,Q}$ and appended to $S_{r,Q}$. With the parameter λ_r, initial term candidates $C_{r,Q}$, and the term pair similarity function $sim_r(c, c_i)$, which depend on the particular resource, Formula 1 becomes a generalized version of Maximal Marginal Relevance-based Expansion (MMRE) proposed by Bouchoucha et al. [3]. And by instantiating λ_r, $C_{r,Q}$ and $sim_r(c, c_i)$, our framework can integrate any resource.

We investigate four typical resources in this work: ConceptNet, Wikipedia, query logs, and initial search results, hereafter denoted by C, W, QL and D, respectively. For ConceptNet, we use the same approach introduced in [3] to compute $C_{C,Q}$ and $sim_C(c, c_i)$, which is based on how similar are the terms related to c and c_i in ConceptNet. For Wikipedia, we define $C_{W,Q}$ as the outlinks, categories, and the set of terms that co-occur with Q or a part of Q [2]; $sim_W(c, c_i)$ is defined by Formula 3, where W (W_i) is the set of vectors containing term c (c_i) obtained by ESA, and $sim(w, w_i)$ is simply the cosine similarity of vector w and w_i.

$$sim_W(c, c_i) = \frac{1}{|W||W_i|} \sum_{w \in W, w_i \in W_i} sim(w, w_i) \tag{3}$$

For query logs, $C_{QL,Q}$ includes the queries that share the same click-through data with Q, as well as the reformulated queries of Q that appear in a user session within a 30 minutes-time window; $sim_{QL}(c, c_i)$ is defined by Formula 4 based on the co-occurences of c and c_i within these queries.

$$sim_{QL}(c, c_i) = \frac{2 \cdot |\{Q^{'} | Q^{'} \in C_{QL,Q}, c \in Q^{'}, c_i \in Q^{'}\}|}{|\{Q^{'} | Q^{'} \in C_{QL,Q}, c \in Q^{'}\}| + |\{Q^{'} | Q^{'} \in C_{QL,Q}, c_i \in Q^{'}\}|} \tag{4}$$

For initial search results, we consider top K returned results as relevant documents (K is experimentally set to 50 in our experiments), and use PRF to generate $C_{D,Q}$; $sim_D(c, c_i)$ is computed using Formula 5, where $freq(c, c_i)$ refers to the co-occurrence of term c and c_i within a fixed window of size 15.

$$sim_D(c, c_i) = \frac{2 \cdot freq(c, c_i)}{\sum_{c^{'}} freq(c, c^{'}) + \sum_{c^{'}} freq(c_i, c^{'})} \tag{5}$$

[2] In cases where no Wikipedia pages match Q or a part of Q, we use *Explicit Semantic Analysis* (*ESA*)[6] to get semantically related Wikipedia pages, from which to extract the outlinks, categories and representative terms to obtain $C_{W,Q}$.

The second layer of our framework generates diversified search results in three steps. First, for each resource r, it generates a set of ranked documents $D_{r,Q}$ using the expansion terms $S_{r,Q}$, which are then combined into one unique set D_Q. Finally, it uses again the MMR principle [4] to iteratively select d^* from the current document candidates. Formula 6 defines this process, where DC_Q denotes the document candidates, which is initialized as D_Q; DS_Q denotes the set of selected documents, which is empty at the very beginning; λ is the parameter that controls the tradeoff between relevance and diversity; $rel(d, Q)$ measures the similarity between document d and query Q; $sim(d, d_i)$ denotes the similarity between two documents (we use the cosine similarity in our experiments).

$$d^* = argmax_{d \in DC_Q}(\lambda \cdot rel(d, Q) - (1 - \lambda) \cdot max_{d_i \in DS_Q} sim(d, d_i)) \quad (6)$$

One core element of the second layer is $rel(d, Q)$, which is defined using Formula 7, where $rel(D_{r,Q}, d)$ and $rank(D_{r,Q}, d)$ are the normalized relevance score[3] and the rank of document d in $D_{r,Q}$[4], respectively. This formula captures our intuition that the more a document is ranked on top and with higher relevance score, the more relevant it is to the query.

$$rel(d, Q) = \sum_{r \in R} \frac{rel(D_{r,Q}, d)}{rank(D_{r,Q}, d)} \quad (7)$$

3 Experiments

We conduct experiments on the ClueWeb09 (category B) dataset, which contains 50,220,423 documents (1.5 TB), and use the test queries from TREC 2009, 2010 and 2011 Web tracks. Indri is used as the basic retrieval system. Our baseline is a query generative language model with Dirichlet smoothing (μ=2000), Krovetz stemmer, and stopword removal using the standard INQUERY stopword list.

We consider four typical resources: the last version of ConceptNet[5], the English Wikipedia dumps of July 8th, 2013, the log data of Microsoft Live Search 2006, which spans over one month (starting from May 1st) consisting of 14.9M queries shared between around 5.4M user sessions, and the top 50 results returned for the original query.

The evaluation results in the diversity task of the TREC 2009, 2010 and 2011 Web tracks are reported based on five official measures: MAP and nDCG for adhoc performance, α-nDCG (in our experiments, $\alpha = 0.5$) and ERR-IA for diversity measure. We also use S-recall to measure the ratio of covered subtopics for a given query. Using greedy search on each resource (step=0.1), we empirically set $\lambda_C = 0.6$, $\lambda_W = 0.5$, $\lambda_{QL} = 0.4$, $\lambda_D = 0.6$, and $\lambda = 0.3$. For each test query, we generate 10 expansion terms using MMRE, with respect to each resource.

[3] exp function is used for normalization, *i.e.*, $x \leftarrow \frac{\exp x}{\sum_{x'} \exp x'}$.

[4] For $d \notin D_{r,Q}$, we set $\frac{1}{rank(D_{r,Q},d)} = 0$.

[5] `http://conceptnet5.media.mit.edu`

Table 1 reports our evaluation results, from which we make four main observations. Firstly, when each resource is considered separately, using query logs often yields significantly better adhoc retrieval performance and diversity. A possible explanation is that the candidate expansion terms generated from query logs are those suggested by users (through their query reformulations), which reflect well the user intent. This suggests the important role of query logs for the diversity task.

Secondly, Wikipedia outperforms ConceptNet for TREC 2009 and TREC 2010, but not significantly in general. However, ConceptNet significantly outperforms Wikipedia for TREC 2011 in all the measures. To understand the reason, we manually assess the different queries to see whether they have an exact matching page from Wikipedia. We find that 36/50, 34/48 and 19/50 queries from TREC 2009, TREC 2010 and TREC 2011 respectively, have exact matching pages from Wikipedia (including the disambiguation and redirection pages), and that only when the query corresponds to a known concept (*i.e.* page) from Wikipedia, the candidate expansion terms suggested by Wikipedia tend to be relevant. This means that Wikipedia helps promoting the diversity of the query results, if the query corresponds to a known concept.

Table 1. Experimental results of different models on TREC Web tracks query sets. BL denotes the baseline model; MMR is the model based on results re-ranking (with λ=0.6) [4]; MMRE_C, MMRE_W,MMRE_{QL}, and MMRE_D refer to the MMRE model based on ConceptNet, Wikipedia, query logs, and search results, respectively; COMB denotes the model combining all the four resources. *, -, +, §,♯, and ‡ indicate significant improvement ($p < 0.05$ in T-test) over BL, MMR, MMRE_C, MMRE_W, MMRE_{QL}, and MMRE_D, respectively.

Queries	Model	MAP	nDCG@20	α-nDCG@20	ERR-IA@20	S-recall@20
TREC 2009	BL	0.161	0.240	0.188	0.097	0.367
	MMR	0.166	0.246	0.191*	0.103	0.377
	MMRE_C	0.195*-‡	0.293*-‡	0.269*-‡	0.140*-‡	0.482*-
	MMRE_W	0.208*-+‡	0.319*-+‡	0.274*-‡	0.146*-‡	0.510*-‡
	MMRE_{QL}	0.221*-+§‡	0.340*-+§‡	0.295*-+§‡	0.153*-§‡	0.599*-+§‡
	MMRE_D	0.188*	0.276*	0.224*	0.115*	0.435*
	COMB	**0.258***-+§♮‡	**0.379***-+§♮‡	**0.328***-+§♮‡	**0.181***-+§♮‡	**0.672***-+§♮‡
TREC 2010	BL	0.103	0.115	0.198	0.110	0.442
	MMR	0.106	0.119	0.209*	0.111	0.459
	MMRE_C	0.146*-‡	0.196*-‡	0.293*-‡	0.165*‡	0.664*-‡
	MMRE_W	0.149*-‡	0.203*-‡	0.317*-+‡	0.174*-‡	0.683*-‡
	MMRE_{QL}	0.158*-+§‡	0.221*-+‡	0.341*-+§‡	0.182*-+§‡	0.694*-‡
	MMRE_D	0.117*	0.142*	0.225*	0.148*	0.508*
	COMB	**0.173***-+§♮‡	**0.239***-+§♮‡	**0.352***-+§♮‡	**0.195***-+§♮‡	**0.703***-+§‡
TREC 2011	BL	0.093	0.155	0.380	0.272	0.700
	MMR	0.096	0.159	0.382	0.269	0.714
	MMRE_C	0.155*-§‡	0.320*-§‡	0.552*-§‡	0.397*-§‡	0.975*-§‡
	MMRE_W	0.124*-‡	0.255*-‡	0.449*-‡	0.313*-‡	0.798*-‡
	MMRE_{QL}	0.160*-+§‡	0.342*-+§‡	0.578*-+§‡	0.411*-§‡	0.982*-§‡
	MMRE_D	0.104*	0.163*	0.397	0.279	0.733
	COMB	**0.167***-+§‡	**0.359***-+§♮‡	**0.586***-+§‡	**0.422***-+§‡	**0.990***-+§‡

Thirdly, the set of feedback documents has the poorest performance among all resources under consideration. Its performance drastically decreases from TREC 2009 to TREC 2010 to TREC 2011 in terms of adhoc retrieval and diversity.

This may be due to the fact that the topics are harder and harder from TREC 2009 to TREC 2010 and TREC 2011 (based on the MAP values). The more the collection contains difficult queries, the more likely the set of top returned documents are irrelevant. Hence, the candidate expansion terms generated from these documents tend to include a lot of noise.

Finally, combing all these resources gives better performance, and in most cases the improvement is significant for almost all the measures. In particular, the diversity scores obtained (for α-nDCG@20, ERR-IA@20, and S-recall@20), are the highest scores. This means that the considered resources are complementary in term of coverage of query subtopics: the subtopics missed by some resources can be recovered by other ones, as demonstrated by "defender" in the introduction. Moreover, our combination strategy promotes the selection of the most relevant documents in the final results set, which explains why higher scores for MAP and nDCG@20 are obtained.

4 Conclusions and Future Work

This paper presents a unified framework to integrate multiple resources for DQE. By implementing two functions, one to generate expansion term candidates and the other to compute the similarity of two terms, any resource can be plugged into this framework. Experimental results on TREC 2009, 2010 and 2011 Web tracks show that combining several complementary resources performs better than using one single resource. We have observed that the degree of the contribution of a resource to SRD depends on the query. In future, we are interested in other approaches to resource integration for DQE, e.g., assigning different resources with different weights that are sensitive to the query.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5-14, 2009.
2. M. Bendersky, D. Metzler, and W.B. Croft. Effective query formulation with multiple information sources. In *Proceedings of WSDM*, pages 443-452, 2012.
3. A. Bouchoucha, J. He, and J.-Y. Nie. Diversified query expansion using conceptnet. In *Proceedings of CIKM*, pages 1861-1864, 2013.
4. J. Carbonell, and J. Goldstein. The use of mmr diversity-based reranking for documents and producing summaries. In *Proceedings of SIGIR*, pages 335-336, 1998.
5. R. Deveaud, E. SanJuan, and P. Bellot. Estimating topical context by diverging from external resources. In *Proceedings of SIGIR*, pages 1001-1004, 2013.
6. E. Gabrilovich, and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606-1611, 2007.
7. R.L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulation for web search result diversification. In *Proceedings of WWW*, pages 881-890, 2010.
8. R. Speer, and C. Havasi. Representing general relational knowledge in conceptnet 5. In *Proceedings of LREC*, pages 3679-3686, 2012.