Improving Difficult Queries by Leveraging Clusters in Term Graph

Rajul Anand and Alexander Kotov

Department of Computer Science, Wayne State University, Detroit MI 48226, USA
{rajulanand,kotov}@wayne.edu

Abstract. Term graphs, in which the nodes correspond to distinct lexical units (words or phrases) and the weighted edges represent semantic relatedness between those units, have been previously shown to be beneficial for ad-hoc IR. In this paper, we experimentally demonstrate that indiscriminate utilization of term graphs for query expansion limits their retrieval effectiveness. To address this deficiency, we propose to apply graph clustering to identify coherent structures in term graphs and utilize these structures to derive more precise query expansion language models. Experimental evaluation of the proposed methods using term association graphs derived from document collections and popular knowledge bases (ConceptNet and Wikipedia) on TREC datasets indicates that leveraging semantic structure in term graphs allows to improve the results of difficult queries through query expansion.

Keywords: difficult queries, term graphs, graph clustering, knowledge bases

1 Introduction

Vocabulary mismatch between documents and queries, when the searchers and authors of relevant documents use different terms to refer to the same concepts, is one of the major causes of poor initial results for some queries (or difficult queries). Due to the lack of positive relevance signals in the initial retrieval results, improvement of retrieval accuracy for such queries cannot be achieved by employing standard techniques, such as pseudo-relevance feedback, and requires utilization of additional resources, such as term graphs. Term graphs are weighted directed graphs, in which the nodes correspond to the basic lexical units (terms or phrases) and the weighted edges represent the strength of semantic relatedness between a pair of such units. Term graphs are rich sources of terms for query and document expansion and can be constructed either manually or automatically. Automatically constructed term graphs (or statistical term association graphs) are derived from document collections by calculating a term co-occurrence based information-theoretic measure, such as mutual information (MI) [8], for each pair of distinct terms in the vocabulary of a collection. Besides term association graphs, expansion LMs can also be derived from manually curated knowledge repositories, such as ConceptNet [6]) (large semantic network) or DBpedia (Wikipedia infoboxes represented as an RDF graph).

Automatically constructed term association graphs have been recently applied to address the problem of vocabulary mismatch in ad-hoc information retrieval through document and query expansion. In particular, Karimzadehgan and Zhai [2] leveraged the MI-based term association graph to estimate translation model for document expansion, Kotov and Zhai [3] used term association graphs for interactive query disambiguation and Bai et al. [1] experimented with using different number of top-k related terms from statistical term association graphs for query expansion. All these methods expand a given query or document term with either the *top-k or all related terms from the term association graph*. We, however, hypothesize that *unstructured and indiscriminate utilization of term association graphs results in suboptimal retrieval performance*, since statistical term association graphs are usually fairly noisy. To overcome this problem, we propose to capture semantic structure in term association graphs through graph clustering and leverage the identified clusters to derive more precise and robust query expansion language models (LMs) to improve the retrieval results of difficult queries.

We illustrate our approach with an example in Figure 1. This example shows a fragment of a term graph, which includes the query term "greek" from the TREC topic 433 "Greek philosophy, stoicism" and the 8 terms that are most strongly associated with it. Previously proposed methods [1, 2] include all these related terms into the resulting expansion LM. Instead, we propose to apply graph clustering methods to term graphs to first determine a set of clusters (connected components) with an intuition that such components correspond to sets of semantically coherent expansion terms. Given a query, our method would then include only those related terms from a term graph that are in the same clusters with the query terms. We hypothesize that such filtering allows to effectively discard spurious term associations and improve the retrieval effectiveness of resulting expansion LMs. Applying the proposed method to our example, the query term "greek" will contribute the terms "greece", "cyprus", "cypriot" and "athens" to the query expansion LM (shown in gray shade).

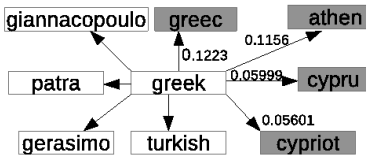

Fig. 1. Constructing more robust query (document) expansion LM by filtering out the terms that are not in the same term graph cluster as the query (document) term.

The key difference between our proposed approach and the previously proposed clustering-based retrieval methods [5, 7] is that our method leverages clusters in a term graph, rather than a document collection. The main contributions of this work are two-fold:

- we propose a method to derive more robust query expansion LMs based on leveraging clusters in term graphs and experimentally demonstrate that the query expansion LMs constructed using the proposed method are more effective in improving the accuracy of difficult queries than the query expansion LMs obtained by including all related terms;
- we compare the retrieval effectiveness of term association graphs with term graphs derived from popular knowledge repositories (DBpedia and ConceptNet). Although ConceptNet [4] and Wikipedia [10] have both been individually utilized for different IR tasks, our approach is the first to leverage the clusters in term graphs derived from these knowledge repositories.

2 Method

2.1 Term graph construction and clustering

The proposed method uses mutual information (MI) [8], a co-occurrence based information-theoretic measure, to captures semantic relatedness between the nodes in the term graph. Infomap [9] is a *state-of-the-art*, *non-parametric* algorithm for finding communities in large networks, which utilizes information-theoretic measures and models stochastic graph flow to obtain the optimal clusters. The algorithm uses hierarchical map equation to measure the per-step average code length necessary to describe a random walker's movements on a graph, given its hierarchical partition, and finds the partition that minimizes the code length. For a term graph with n nodes divided into m modules, the lower bound on the code length is defined by the map equation:

$$L(M) = \sum_{i=1}^{m} Q_i \ \log \sum_{i=1}^{m} Q_i - 2\sum_{i=1}^{m} Q_i \ \log \ Q_i - \sum_{j=1}^{n} p_j \ \log \ p_j + \sum_{i=1}^{m} (Q_i + \sum_{j \in i} p_j) \ \log(Q_i + \sum_{j \in i} p_j)$$

where Q_i is the probability of the random walk to exit the partition i and p_j is the frequency of node j.

2.2 Datasets

Table 1. Statistics of experimental datasets.

Dataset	# docs	size (MB)	# tops	# hard
AQUAINT	10,033,461	3,042	50	17
ROBUST	528,155	1,910	250	75
GOV	1,247,753	18,554	225	147

For all experiments in this work we used AQUAINT, ROBUST and GOV TREC collections, various statistics of which are summarized in Table 1. For each experimental dataset, we constructed a term association graph using MI as a similarity measure. DBpedia term graph was constructed by treating DBpedia 3.9[1] extended abstracts, which contain all the words in the first section

[1] `http://wiki.dbpedia.org/Downloads39`

of Wikipedia articles, as a document collection and using MI as a similarity measure. ConceptNet term graph was constructed by removing all non-English terms and negative associations from the core ConceptNet 5 term graph. We considered two versions of the ConceptNet term graph. The first version uses original weights of edges provided with ConceptNet 5 (**CNET**)[2], while in the second version, the weights between the concepts are calculated for each collection using MI (**CNET-MI**). We further customized Wikipedia and ConceptNet term graphs for each experimental collection by removing all the nodes that do not occur in the index of that collection.

Our proposed methods can be divided into three categories. Methods using collection term association graphs include **COL-ALL**, which uses all related terms to construct query expansion LM and **COL-INFO**, which selects expansion terms based on Infomap clustering. Similarly, **WIKI-ALL**, **CNET-ALL**, **CNET-MI-ALL** use all related terms from the corresponding term graph, while **WIKI-INFO**, **CNET-INFO**, **CNET-MI-INFO** filter expansion terms based on Infomap clustering from Wikipedia and ConceptNet term graphs, respectively.

2.3 Retrieval model and query expansion

We used the KL-divergence retrieval model with Dirichlet prior smoothing [11] (**KL-DIR**), according to which the retrieval task involves estimating Θ_Q, a query language model (LM) for a given keyword query $Q = \{q_1, q_2, \ldots, q_k\}$, and document language models Θ_{D_i} for each document D_i in the document collection $\mathcal{C} = \{D_1, \ldots, D_m\}$. We define a query as difficult (or hard), if the average precision of results retrieved with the **KL-DIR** retrieval model is less than 0.1.

In language modeling approach to IR, query expansion is typically performed via linear interpolation of the original query model $P(w|Q)$ and query expansion LM $P(w|\hat{Q})$ with parameter λ:

$$P(w|\tilde{Q}) = \lambda P(w|Q) + (1-\lambda)P(w|\hat{Q}) \tag{1}$$

Estimating query expansion LM $P(w|\hat{Q})$ using clusters in a term graph involves finding a set of semantically related terms $\mathcal{E}_{q_i}$ for each query term q_i (i.e. all direct neighbors of the query term q_i in the term graph that are in the same term graph cluster C_{q_i} as q_i) and normalizing the probabilities using the following formula:

$$p(w|\hat{Q}) = \frac{\sum_{i=1}^{k} p(w|q_i)}{\sum_{i=1}^{k} \sum_{w \in C_{q_i}} p(w|q_i)} \tag{2}$$

3 Experiments

We pre-processed each dataset by removing stopwords and stemming (using Porter stemmer). To construct collection term association graphs, we removed

[2] `http://conceptnet5.media.mit.edu/downloads/20130917/associations.txt.gz`

all the terms that either occur in less than five documents or in more than 10% of all documents in a given collection. We used the following settings of Infomap parameters: self link teleportation probability was set to 0.1, node teleportation probability was to 0.01 and random seed to 111222333. The optimal value for self link teleportation probability was determined empirically to reduce the number of very small clusters (which include less than 5 terms). We used the **KL-DIR** retrieval model and document expansion using translation model based on MI term graph (**TM**) [2] as the baselines. The reported results are based on the optimal settings of the Dirichlet prior μ, interpolation parameter λ that were empirically determined for all methods and the baselines. Summary of retrieval performance of the proposed methods and the baselines on each experimental dataset is provided in Tables 2, 3 and 4. The entries corresponding to the highest and second highest values were highlighted in boldface and italics. We performed statistical significance testing of MAP values using Wilcoxon signed rank test (▲ and • represent statistically significance difference ($p < 0.05$) relative to **KL-DIR** and **TM** baselines, respectively).

Table 2. Summary of retrieval performance on AQUAINT collection for difficult topics.

Method	MAP	P@5	GMAP
KL-DIR	0.0474	0.1250	0.0386
TM	0.0478	0.1250	0.0386
COL-ALL	0.0476	0.1375	0.0393
COL-INFO	0.0482	0.1375	0.0397
WIKI-ALL	*0.0528*$^{▲•}$	0.1850	*0.0452*
WIKI-INFO	0.0501$^{▲}$	0.1750	0.0405
CNET-ALL	0.0504$^{▲}$	*0.1875*	0.0440
CNET-INFO	**0.0531**$^{▲•}$	**0.1950**	**0.0471**
CNET-MI-ALL	0.0496$^{▲•}$	*0.1875*	0.0422
CNET-MI-INFO	0.0527$^{▲•}$	**0.1950**	0.0416

Table 3. Summary of retrieval performance on ROBUST collection for difficult topics.

Method	MAP	P@5	GMAP
KL-DIR	0.0410	0.1544	0.0261
TM	0.0458	0.1646	0.0267
COL-ALL	0.0429$^{▲}$	0.1594	0.0273
COL-INFO	0.0463$^{▲}$	0.1949	0.0279
WIKI-ALL	0.0503$^{▲•}$	0.1848	0.0301
WIKI-INFO	0.0535$^{▲•}$	0.1870	0.0271
CNET-ALL	0.0559$^{▲•}$	0.1899	*0.0334*
CNET-INFO	*0.0580*$^{▲•}$	*0.1924*	**0.0344**
CNET-MI-ALL	0.0560$^{▲•}$	**0.1949**	0.0326
CNET-MI-INFO	**0.0582**$^{▲•}$	0.1899	0.0301

Several conclusions can be drawn from experimental results. First, cluster-based filtering of query expansion LMs derived from both statistical term association graphs and knowledge base term graphs improves retrieval performance in majority of cases on all 3 experimental collections. This indicates that graph clustering is effective at capturing semantically strong associations in the context of a given collection, while discarding the spurious ones. Secondly, query expansion LMs based on filtered term graphs derived from Wikipedia and ConceptNet (**WIKI-INFO**, **CNET-INFO**, **CNET-MI-INFO**) generally outper-

Table 4. Summary of retrieval performance on GOV collection for difficult topics.

Method	MAP	P@5	GMAP
KL-DIR	0.0114	0.0233	0.0103
TM	0.0128	0.0248	0.0107
COL-ALL	0.0120	0.0243	0.0105
COL-INFO	0.0125	0.0245	0.0112
WIKI-ALL	0.0123	0.0242	0.0112
WIKI-INFO	0.0121	0.0236	0.0104
CNET-ALL	0.0128$^{\blacktriangle}$	*0.0258*	0.0121
CNET-INFO	**0.0196**$^{\blacktriangle\bullet}$	**0.0290**	0.0124
CNET-MI-ALL	0.0176$^{\blacktriangle\bullet}$	0.0242	*0.0129*
CNET-MI-INFO	*0.0195*$^{\blacktriangle\bullet}$	0.0255	**0.0131**

formed query expansion LMs derived from association term graphs (**COL-ALL** and **COL-INFO**) as well as document expansion based on translation model (**TM**) according to all metrics on all datasets. Finally, term graphs derived from ConceptNet generally outperformed the ones derived from Wikipedia on all 3 collections, which highlights the importance of commonsense knowledge in IR besides entity information in DBpedia.

4 Conclusion

In this paper, we proposed a method to derive more accurate query expansion LMs by clustering term graphs and experimentally demonstrated that applying this method to statistical term association graphs and term graphs derived from knowledge bases translates into more accurate retrieval results for difficult queries.

References

1. J. Bai, D. Song, P. Bruza, J.-Y. Nie and G. Cao. Query expansion using term relationships in language models for information retrieval. In *Proceedings of CIKM'05*, pp. 688–695, 2005.
2. M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of SIGIR'10*, pp. 323–330, 2010.
3. A. Kotov and C. Zhai. Interactive Sense Feedback for Difficult Queries. In *Proceedings of CIKM'11*, pp. 59–66, 2009.
4. A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *Proceedings of WSDM'12,* pp. 403–412, 2012.
5. O. Kurland. The opposite of smoothing: a language model approach to ranking query-specific document clusters. In *Proceedings of SIGIR'08*, pp. 171–178, 2008.
6. H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit *BT Technology Journal*, 22(4), pp. 211–226, 2004.
7. X. Liu and B. Croft. Cluster-based retrieval using language models. In *Proceedings of SIGIR'04*, pp. 186–193, 2004.
8. C. Manning and Hinrich Schütze. Foundations of statistical natural language processing. *The MIT Press*, 1999.
9. M. Rosvall and C. Bergstrom. Maps of information flow reveal community structure in complex network. In *PNAS 105*: pp. 1118–1123, 2008.
10. Y. Xu, G. Jones and B. Wang. Query dependent pseudo-relevance feedback based on Wikipedia. In *Proceedings of SIGIR'09*, pp. 59–66, 2009.
11. C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pp. 111–119, 2001.