An Empirical Comparison of Term Association and Knowledge Graphs for Query Expansion

Saeid Balaneshinkordan and Alexander Kotov

Department of Computer Science, Wayne State University, Detroit, MI 48202, USA
{saeid.balaneshinkordan,kotov}@wayne.edu

Abstract. Term graphs constructed from document collections as well as external resources, such as encyclopedias (DBpedia) and knowledge bases (Freebase and ConceptNet), have been individually shown to be effective sources of semantically related terms for query expansion, particularly in case of difficult queries. However, it is not known how they compare with each other in terms of retrieval effectiveness. In this work, we use standard TREC collections to empirically compare the retrieval effectiveness of these types of term graphs for regular and difficult queries. Our results indicate that the term association graphs constructed from document collections using information theoretic measures are nearly as effective as knowledge graphs for Web collections, while the term graphs derived from DBpedia, Freebase and ConceptNet are more effective than term association graphs for newswire collections. We also found out that the term graphs derived from ConceptNet generally outperformed the term graphs derived from DBpedia and Freebase.

Keywords: Query Expansion, Term Graphs, Knowledge Bases, Difficult Queries

1 Introduction

Vocabulary gap, when searchers and the authors of relevant documents use different terms to refer to the same concepts, is one of the fundamental problems in information retrieval. In the context of language modeling approaches to IR, vocabulary gap is typically addressed by adding semantically related terms to query and document language models (LM), a process known as query or document expansion. Therefore, effective and robust query and document expansion requires information about term relations, which can be conceptualized as a term graph. The nodes in this graph are distinct terms, while the edges are weighed according to the strength of semantic relationship between pairs of terms.

Term association graph is constructed from a given document collection by calculating a co-occurrence based information theoretic measure, such as mutual information [7] or hyperspace analog to language [2], between each pair of terms in the collection vocabulary. Term graphs can also be derived from knowledge bases, such as DBpedia[1], a structured version of Wikipedia, Freebase[2], a pop-

[1] http://wiki.dbpedia.org/
[2] http://freebase.com/

ular graph-structured knowledge base created from different data sources, and ConceptNet[3], a large semantic network constructed via crowdsourcing.

Term association and knowledge graphs have their own advantages and disadvantages. The weights of edges between the terms in automatically constructed term graphs *are specific to each particular document collection*. On the other hand, methods that establish semantic term relatedness based only on co-occurrence require large amounts of data and often produce noisy term graphs. Semantic term associations in external resources (e.g. thesauri, encyclopedias, ontologies, semantic networks) are static and manually curated, but may result in a topic drift. It is also generally unknown which external resource would be the most effective for a particular collection type (e.g. shorter Web document versus longer news articles).

While the methods for retrieval from DBpedia [12] as well as query expansion utilizing ConceptNet [5], Freebase [9] and Wikipedia [10] in the context of pseudo-relevance feedback (PRF) have been examined in detail in previous studies, in this work, we focus on empirical comparison of retrieval effectiveness of term graphs derived from knowledge repositories with automatically constructed terms association graphs on the same standard IR collections of different type. Our work is also the first one to evaluate the effectiveness of DBpedia for query expansion at the level of individual terms without PRF.

2 Methods

2.1 Statistical term association graphs

Statistical term association graphs are constructed by calculating a co-occurrence based information theoretic measure of similarity, such as Mutual information (MI) [7] or Hyperspace Analog to Language (HAL) [2], between each pair of terms in the vocabulary of a given document collection and considering the top-k terms with the highest value of that measure for each given term. The key difference between MI and HAL is in the size of contextual window to calculate co-occurrence. Term co-occurrences within entire documents are considered in MI calculation, whereas a sliding window of small size is used for HAL.

Mutual information measures the strength of association between a pair of terms based on the counts of their individual and joint occurrence. The higher the mutual information between the terms, the more often they tend to co-occur in the same documents, and hence the more semantically related they are.

Hyperspace Analog to Language is a representational model of high dimensional concept spaces, which was created based on the studies of human cognition. Previous work [8] has demonstrated that HAL can be effectively utilized in IR. Constructing the HAL space for an n-term vocabulary involves traversing a sliding window of width w over each term in the corpus. All terms within a sliding window are considered as part of the local context for the term, over which the sliding window is centered. Each word in the local

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

context is assigned a weight according to its distance from the center of the sliding window (words that are closer to the center receive higher weight). An $n \times n$ HAL space matrix $\mathbf{H}$, which aggregates the local contexts for all the terms in the vocabulary, is created after traversing an entire corpus. After that, the global co-occurrence matrix is produced by merging the row and column corresponding to each term in the HAL space matrix. Each distinct term w_i in the vocabulary of the collection corresponds to a row in the global co-occurrence matrix $\mathbf{H}_{w_i} = \{(w_{i1}, c_{i1}), \ldots, (w_{in}, c_{in})\}$, where $c_{i1}, \ldots, c_{in}$ are the number of co-occurrences of the term w_i with all other terms in the vocabulary. After the merge, each row $\mathbf{H}_{w_i}$ in the global co-occurrence matrix is normalized to obtain a HAL-based semantic term similarity matrix for the entire collection:

$$\mathbf{S}_{w_i} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}$$

Due to the context window of smaller size, HAL-based term association graphs are typically less noisy than MI-based ones.

2.2 Knowledge repositories

In addition to statistical term association graphs, we also experimented with the term graphs based on DBpedia, Freebase and ConceptNet. The key difference between DBpedia, Freebase and ConceptNet lies in the type of knowledge they provide.

DBpedia is a structured version of Wikipedia infoboxes, which provides descriptions of entities (people, locations, organizations, etc.) as RDF triplets. We used DBpedia 3.9[4] extended abstracts, which usually contain all words in the first section of the Wikipedia article corresponding to an entity, for term graph construction. Treating extended abstracts as documents, we generated two term graphs DB-MI and DB-HAL using MI and HAL as similarity measures, respectively. Those graphs were customized for each document collection by removing the words that are not in the index of a given collection.

Freebase, similar to DBpedia, provides descriptions of entities as RDF triplets, but features a more comprehensive list of concepts than DBpedia. We used the text property of documents (/common/document/text), which contains extended textual descriptions of entities, to generate the FB-MI and FB-HAL term graphs.

ConceptNet [6] codifies commonsense knowledge as subject-predicate-object triplets (e.g. “alarm clock”, UsedFor, “wake up”) and can be viewed as a semantic network, in which the nodes correspond to semi-structured natural language fragments (e.g., “food”, “grocery store”, “buy food”, “at home”) representing real or abstract concepts and the edges represent semantic relationships between the concepts. For experiments in this work, we used the weights between the concepts provided by ConceptNet 5 (CNET)[5], as well as the ones calculated for each

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

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

collection using MI (CNET-MI) and HAL (CNET-HAL). As in the case of DBpedia, we customized the term graph by removing the words that are not in the index of a given collection.

2.3 Retrieval model and query expansion

We used the KL-divergence retrieval model with Dirichlet prior smoothing [11], according to which each document D in the collection is scored and ranked based on the Kullback-Leibler divergence between the query LM Θ_Q and document LM Θ_D. In language modeling approaches to IR, query expansion is typically performed via linear interpolation of the original query LM $p(w|Q)$ and query expansion LM $p(w|\hat{Q})$ with the parameter α:

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

Query expansion using a term graph involves finding a set of semantically related terms for each query term q_i (i.e. all direct neighbors of query terms in the term graph) and estimating $p(w|\hat{Q})$ according to the following formula:

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

where $p(w|q_i)$ is the strength of semantic association between w and q_i according to a particular term graph.

3 Experiments

3.1 Datasets

For all experiments in this work we used AQUAINT, ROBUST and GOV datasets from TREC, which were pre-processed by removing stopwords and applying the Porter stemmer. To construct the term association graphs, all rare terms (that occur in less than 5 documents) and all frequent terms (that occur in more than 10% of all documents in the collection) have been removed [4, 3]. Term association graphs were constructed using either the top 100 most related terms or the terms with similarity metric greater than 0.001 for each distinct term in the vocabulary of a given collection. HAL term association graphs were constructed using the sliding window of size 20 [4]. The reported results are based on the optimal settings of the Dirichlet prior μ and interpolation parameter α empirically determined for all the methods and the baselines. Top 85 terms most similar to each query term were used for query expansion [1]. KL-divergence retrieval model with Dirichlet prior smoothing (**KL-DIR**) and document expansion based on translation model [3] (**TM**) were used as the baselines.

3.2 Results

Retrieval performance of query expansion using different types of term graphs and the baselines on different collections and query types is summarized in Tables 1, 2 and 3. The best and the second best values for each metric are highlighted in boldface and italic, while † and ‡ indicate statistical significance in terms of MAP ($p < 0.05$) using Wilcoxon signed rank test over the **KL-DIR** and **TM** baselines, respectively.

(a)

Method	MAP	P@20	GMAP
KL-DIR	0.1943	0.3940	0.1305
TM	0.2033	0.3980	0.1339
NEIGH-MI	0.2031†	0.3970	0.1326
NEIGH-HAL	0.1989†	0.3900	0.1319
DB-MI	**0.2073**†‡	**0.4160**	**0.1468**
DB-HAL	*0.2059*†‡	*0.4080*	*0.1411*
FB-MI	0.2055†‡	0.3990	0.1336
FB-HAL	0.2056†‡	0.3960	0.1384
CNET	0.2051†‡	0.3900	0.1388
CNET-MI	0.2042†	0.3920	0.1371
CNET-HAL	0.2058†‡	0.3920	0.1388

(b)

Method	MAP	P@20	GMAP
KL-DIR	0.0474	0.1250	0.0386
TM	0.0478	0.1250	0.0386
NEIGH-MI	0.0476	0.1375	0.0393
NEIGH-HAL	0.0474	0.1500	0.0378
DB-MI	0.0528†‡	**0.1906**	0.0452
DB-HAL	*0.0544*†‡	*0.1538*	*0.0455*
FB-MI	0.0534†‡	0.1333	0.0437
FB-HAL	**0.0564**†‡	0.1444	**0.0471**
CNET	0.0504†‡	0.1219	0.044
CNET-MI	0.0496†	0.1156	0.0422
CNET-HAL	0.0502†	0.1219	0.0436

Table 1. Retrieval accuracy for (a) all queries and (b) difficult queries on AQUAINT dataset.

(a)

Method	MAP	P@20	GMAP
KL-DIR	0.2413	0.3460	0.1349
TM	0.2426	0.3488	0.1360
NEIGH-MI	0.2432	0.3460	0.1360
NEIGH-HAL	0.2431	0.3454	0.1333
DB-MI	0.2482†‡	0.3524	0.1397
DB-HAL	0.2426	0.3444	0.1349
FB-MI	0.2452†‡	0.3526	0.1232
FB-HAL	0.2476†‡	0.**3540**	0.1261
CNET	0.2452†	0.3472	0.1407
CNET-MI	*0.2495*†‡	*0.3530*	*0.1459*
CNET-HAL	**0.2503**†‡	0.3528	**0.1463**

(b)

Method	MAP	P@20	GMAP
KL-DIR	0.0410	0.1290	0.0261
TM	0.0458	0.1290	0.0267
NEIGH-MI	0.0429†	0.1323	0.0273
NEIGH-HAL	0.0419	0.1260	0.0265
DB-MI	0.0503†‡	0.1449	0.0301
DB-HAL	0.0474†	0.1437	0.0273
FB-MI	0.0381	0.1222	0.0200
FB-HAL	0.0393	0.1272	0.0211
CNET	**0.0559**†‡	**0.1487**	**0.0334**
CNET-MI	*0.0560*†‡	**0.1487**	*0.0326*
CNET-HAL	0.0558†‡	*0.1475*	0.0323

Table 2. Retrieval accuracy for (a) all queries and (b) difficult queries on ROBUST dataset.

Examination of experimental results in Tables 1-3 leads to the following major conclusions. First, relative retrieval performance of different types of term graphs varies by the collection. In particular, term graphs derived from external repositories are significantly more effective than term association graphs for newswire datasets (AQUAINT and ROBUST) on both regular and difficult queries, with the HAL-based term association graph (NEIGH-HAL) outperforming the term graphs derived from DBpedia and Freebase (DB-HAL and FB-HAL) for all queries on the GOV collection. For difficult queries on the same dataset, NEIGH-HAL outperforms Freebase- and DBpedia-based terms graphs and has comparable performance with the term graphs derived from ConceptNet. We

(a)

Method	MAP	P@20	GMAP
KL-DIR	0.2333	0.0464	0.0539
TM	0.2399	0.0476	0.0551
NEIGH-MI	0.2415†‡	0.0489	0.0518
NEIGH-HAL	0.2419†‡	0.0456	0.0476
DB-MI	0.2346	0.0467	0.0529
DB-HAL	0.2404†	0.0467	0.053
FB-MI	*0.2420*†‡	0.0484	0.0573
FB-HAL	0.2404†	0.0476	0.0565
CNET	0.2407†	0.0489	0.0584
CNET-MI	0.2416†‡	*0.0504*	**0.0587**
CNET-HAL	**0.2428**†‡	**0.0516**	*0.0586*

(b)

Method	MAP	P@5	GMAP
KL-DIR	0.0311	0.0281	0.014
TM	0.0343	0.0304	0.0146
NEIGH-MI	0.0333†	0.0307	0.013
NEIGH-HAL	*0.0425*†‡	0.0293	0.0122
DB-MI	0.0312	0.0285	0.0136
DB-HAL	0.0306	0.0274	0.0134
FB-MI	0.0350†‡	*0.0319*	0.0154
FB-HAL	0.0339†	0.0293	*0.0152*
CNET	0.0407 †‡	0.0333	0.0172
CNET-MI	0.0427 †‡	0.0367	0.0176
CNET-HAL	**0.0453**†‡	**0.0385**	**0.0181**

Table 3. GOV dataset results on (a) all queries and (b) difficult queries.

attribute this to the fact that the term graph for GOV is larger in size and less dense than the term graphs for AQUAINT and ROBUST, which results in less noisy term associations. Second, using MI and HAL-based weighs of edges in ConceptNet graph (CNET-MI and CNET-HAL) results in better retrieval accuracy that the original ConceptNet weights (CNET) in the majority of cases. This indicates the utility of tuning the weights in term graphs derived from external resources to particular collections. Finally, ConceptNet-based term graphs outperformed Freebase- and DBpedia-based ones on 2 out of 3 collections used in evaluation, which indicates the importance of commonsense knowledge in addition to information about entities.

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 the 14th ACM CIKM*, pages 688–695, 2005.
2. C. Burgess, K. Livesay, and K. Lund. Explorations in Context Space: Words, Sentences and Discourse. *Discourse Processes*, 25:211–257, 1998.
3. M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd ACM SIGIR*, pages 323–330, 2010.
4. A. Kotov and C. Zhai. Interactive Sense Feedback for Difficult Queries. In *Proceedings of the 20th ACM CIKM*, pages 163–172, 2011.
5. A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the 5th ACM WSDM*, pages 403–412, 2012.
6. H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit *BT Technology Journal*, 22(4), pages 211–226, 2004.
7. C. Manning and H. Schütze. Foundations of Statistical Natural Language Processing. *The MIT Press*, 1999.
8. D. Song and P. Bruza. Towards Context Sensitive Information Inference. *JASIST*, 54(4):321–334, 2003.
9. C. Xiong, and J. Callan. Query Expansion with Freebase. In *Proceedings of the 5th ACM ICTIR*, pages 111–120, 2015.
10. Y. Xu, G.J.F. Jones and B. Wang. Query Dependent Pseudo-Relevance Feedback based on Wikipedia. In *Proceedings of the 32nd ACM SIGIR*, pages 59–66, 2009.
11. C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th ACM SIGIR*, pages 111–119, 2001.
12. N. Zhiltsov, A. Kotov and F. Nikolaev. Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data. In *Proceedings of the 38th ACM SIGIR*, pages 253–262, 2015.