Query Expansion based on Local Distributional Thesauri

Fabiano Tavares da Silva[1], **José Everardo Bessa Maia**[1]

[1]Universidade Estadual do Ceará - Fortaleza - Ce - Brasil

Email: fabiano.tavares@aluno.uece.br,jose.maia@uece.br

Abstract. *This work proposes and evaluates an approach to query expansion in Information Retrieval based on Local Context Analysis using a Distributional Semantic Representation. In general, the approach performed better compared to that of query expansion using non-distributional, local or global techniques, running over datasets of different application domains.*

1. Introduction

A Text Information Retrieval System (IRS) is a process capable of storing, retrieving and maintaining information of collections of documents containing unstructured text [Manning et al. 2008]. The word unstructured implies that such documents have little structure that could serve as a guide locating specific content.

To interact with the IRS the user issues a query. A query is the formulation of a user information need. Keyword based queries are popular, since they are easy to express and intuitive. However, formulating appropriate queries to submit to the IRS is one of the key difficulties for users in information retrieval of the their interest. This is because the keywords posed by users only vaguely describe their information needs and may imply different information needs of different users, and this causes ambiguity during query processing [Carpineto and Romano 2012].

A method for improving retrieval performance is supplementing an original query with additional terms. This Query Expansion (QE) can be performed automatically or interactively and can take place in the initial query formulation, or in a query reformulation stage of the online search, or both [Carpineto and Romano 2012].

A broad classification of QE techniques identifies two alternatives for adding terms to the query and their combination [Baeza-Yates et al. 1999]: local and global strategies. Local strategies add terms using feedback based on relevance to the results of the initial query, while global strategies use the entire collection of documents and/or external resources.

The general method of expansion is to construct a thesaurus and use it to effect the expansion of the query. A thesaurus is a list of words with synonyms or similar meanings.The thesaurus can use local or global scope.

The work in [Xu and Croft 2000] is an example of local strategy for QE named in it of Local Context Analysis [Baeza-Yates et al. 1999]. In this method, an expanded query is formulated on the basis of some retrieved documents of search with original query. Thus, the performance of the method depends on the accuracy of the top results.

The correlation between two terms is calculated using co-occurrence statistics. In contrast, [Bai et al. 2005] also uses co-occurrence statistics of terms but calculated over the entire collection of documents. This is an example of a global statistical strategy.

An alternative route to using statistics are QE based on the semantic-lexical thesaurus. In this methods, a thesaurus built on a large external corpus is used for the QE, being Wordnet [Miller 1995a] the most famous. In [Pal et al. 2014, Zhang et al. 2009] and [Voorhees 1994] the Wordnet thesaurus is used to expand queries with higher results in the med and news domains, respectively.

The literature on query expansion is wide. Surveys can be found in [Carpineto and Romano 2012, Azad and Deepak 2017, Bhogal et al. 2007, Dahab et al. 2018]. The works [SanJuan et al. 2007, Jiang and Conrath 1997, Bhagdev et al. 2008] are hybrid based approaches which combine both statistical and lexical approaches. In general, experiments have shown that query expansion is highly topic dependent.

More recently, embedding techniques have been combined with statistical and semantic-lexical approaches. In [Diaz et al. 2016], word embedding was combined with local context analysis, in [Claveau and Kijak 2016] a global statistical thesaurus was constructed using the word2vec trained representation and the work [Schütze 1998] uses a combination of ontology with word embedding. In all these cases, the papers report that the embedding representation increased the results.

It is well documented the phenomenon that simple query expansion results in improvement of recall in sacrifice of the reduction of precision in the recovered documents [Carpineto and Romano 2012]. Based on this empirical observation, the proposal of this work is a two-stage recovery model. In the first phase, a document search is performed using the query as it was presented by the user, without expansion. This prevents the query from being contaminated by expansion failures, preserving accuracy. The top-level documents returned in this first-phase query are used as the local context for query expansion which is used in the second document search, the final retrieval. The argument here is that this form of local expansion will restore the recall with less impairment of accuracy. In the process, a global distributional representation is used to calculate the measure of similarity between concepts and queries.

Section II describes the approach in detail, Section III presents the plan of the experiments, the datasets and the merit figures used and the results. The work is completed in Section IV with the conclusion.

2. The Approach

The general method of query expansion is represented in Figure 1. Traditionally, it consists of constructing and using a thesaurus to add new terms, assign weights, or recalculate them to modify the original representation. One of the best-known thesauri is Wordnet [Miller 1995b] built manually and with global features. It has the advantage of bringing lexical information, which solves problems of ambiguity in some cases. The disadvantage is that they are generic, thus do not bring gains in specific domains and are laborious to include new terms [Ooi et al. 2015]. On the other hand, in this work, the automatically constructed thesauri are based on the distributional hypothesis of [Harris 1954]

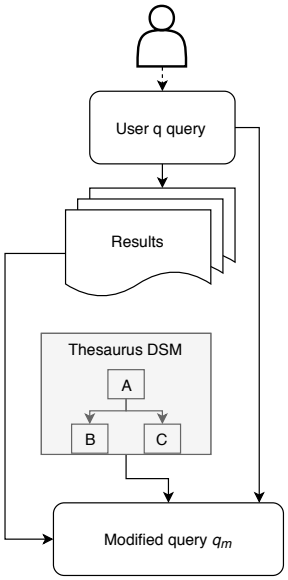

Figura 1. Flow of a query with and without expansion.

which states that the words that are used and occur in the same contexts tend to have similar meanings. From this hypothesis we have constructed theories and methods to represent and quantify the similarity between linguistic items, which we call Distributional Semantics.

In continuity to the work of [Xu and Croft 2000] this work uses Local Context Analysis (LCA) combined with Semantic Distributional Model. This technique proposes to use the first results of a query to construct a representation by co-occurrence of concepts (groups of nouns) and by similarity of these with the query find those candidates to be aggregated to the query expansion, that is, to combine local and global analysis for QE. This is shown in Algorithm 1.

A brief description of this algorithm is given in this paragraph. In line 1, the best ranked documents are retrieved with the original query. In rows 2 to 8, the top-ranked documents retrieve the n most well-ranked passages using the original query. This is achieved by breaking the documents initially retrieved by the query into *passages* and ranking those passages as if they were documents. In lines 9 to 11, for each concept (group of nouns) in the top passages of the results, the similariy $simqc(q, c, th)$ between the entire query q (not the individual query terms) and the concept c, is calculated using a variant of TF-IDF. In lines 12 and 13, the n most well-ranked concepts, according to $simqc(q, c, th)$, are added to the original query q. For each added concept a weight given by $(1 - 0.9i/m)$ is assigned, where i is the position of the concept in the ranking of concepts. The terms in the original query q can be emphasized by assigning a weight equal to 2 for each of them.

Although the structure of algorithm 1 is similar to that of a standard LCA, its implementation is significantly different. It differs in two points: in the concept of context window considered and in the calculation of similarity.

The local context analysis (line 3) uses the notion of passage. A passage is a sentence enclosed in a punctuation mark. This results in a context window of variable

Algorithm 1: Pseudocode for the QE proposed for IR.

Data: query q, thesaurus th
Result: modified query q_m

```
1:  documents ← search_top_ranked(q);
2:  for each documents do
3:      passages ← window(document);
4:      for each passages do
5:          concepts ← find_concepts_in_context(passage);
6:      end for
7:  end for
8:  sort(concepts);
9:  for i ← 1 to N do
10:     m[i] ←simqc(q, concepts[i], th);
11: end for
12: sort(m);
13: q_m ← q + m[1..n];
```

size dependent on the statements present in the document being parsed, unlike the fixed size window used in standard LCA.

The second point on which this algorithm differs from the standard LCA is in the calculation of the similarity between query terms and concepts ($simqc(q, c)$, line 10). Algorithm 1 receives as one of its inputs a previously calculated distributional thesaurus which is taken into account. When it comes to an enclosed collection of documents the thesaurus is calculated for the collection. Already in web applications this thesaurus can be global. The calculation of similarity is given by Equation 1 [Xu and Croft 1996, Croft 2002]:

$$simqc(q, c) = \prod_{k_i \in q}^{t} (\delta + \frac{\log(f(c, k_i) \times IDF_c)}{\log n})^{IDF_i}. \quad (1)$$

In this equation, δ is a small constant (0.1 in [Croft 2002]) to avoid the zeroing of the expression in some cases, $f(c, k_i$ is a function that quantifies the correlation between a concept and a query term considering the distribution in the thesaurus DSM, and n, c and t are as already defined.

The core of the automatic query expansion (AQE) method is the similarity measure $simqd(q, d)$. Given a collection D of documents where each document is represented in the vector space of words of t terms, the terms are indexed forming the vocabulary $V = \{k_1, k_2, ..., k_t\}$, each document $d_i = (w_{1,i}, w_{2,i}, ..., w_{t,i})$. Let q be the query represented as a pseudo document also in the same space $d_0 = (w_{1,0}, w_{2,0}, ..., w_{t,0})$ where w_t is the weight associated with the term in the query. The similarity between query q and document d is expressed as follows:

$$simqd(q, d) = \sum_{t \epsilon q \cap d} w_{t,d} \cdot w_{t,q}. \quad (2)$$

In (1), $w_{t,q}$ e $w_{t,c}$ are weights calculated by the TF-IDF (frequency of the term by the inverse of the frequency in the documents): $w_{i,j} = (1 + log f_{i,j}) \times log(N/n_i)$, if $f_{i,j} > 0$.

Tabela 1. Datasets

Dataset	Language	Matrix size ($terms \times docs$)	No. of topics No. of queries
MED	EN	7876 x 1033	30
LISA	EN	11710 x 6004	35
NPL	EN	7861 x 11429	100

The AQE process consists in using the terms of q to find new terms, without user participation, that solve problems of disambiguation which best discriminate the documents. The initial query q added by the new terms becomes q' and the similarity in (1) becomes $simqd(q', d)$.

The choice of the new terms is made through a thesaurus. The global and static thesaurus WordNet [Miller 1995b] and another one automatically constructed using Semantic Distributional Model with Word Embbeding representation [Mikolov et al. 2013].

3. Results and Discussion

3.1. Data

Results are obtained on three publicly available datasets in [Fox 1990] and have been compared to a baseline method and two reference methods on the same dataset showing the competitiveness of the proposed algorithm. The subjects of documents in datasets are medicine (MED), library science (LISA) and electrical engineering (NPL). The characteristics of this datasets are shown in Table 1.

In MED documents are small and test queries are short, but they generate responses with high accuracy even without query expansion. In LISA the documents are a little larger, but the test queries are long and generate answers with low precision, setting the need to expand queries that are already long. In NPL the documents are short, but in greater quantity. Test queries range from long to short and generate responses with intermediate precision. These are three distinct application scenarios that make it possible to evaluate the merit of IR algorithms in different dimensions.

3.2. Metrics

In the evaluation of performance, four metrics defined in the Granfield paradigm were used, used by the TREC community: MAP, BP, MRR and the Precision-Recall curve. These metrics are precisely defined in [Baeza-Yates et al. 1999, Manning et al. 2008, Buckley and Voorhees 2004]. The Binary preference (BP) measure the number of retrieved documents judged nonrelevant before some relevant document, normalized by the number of relevant judged documents. Mean reciprocal rank (MRR) is the mean, calculated over all queries, of the reciprocal rank of the highest ranking relevant document. To avoid uniqueness, it is zero for a topic if no relevant results were returned. The average precision (AP) of a single query is the mean of the precision scores at each relevant item returned in a search results list. Mean average precision (MAP), then, is the mean of average precision calculated over all queries (topics). In general, there is a tradeoff between precision and recall which is captured by the recall-precision curve (RPC). It is desirable that both accuracy and recall are high, however, this is generally not the case.

Tabela 2. Results for MED dataset

Model	Metric	Baseline	WordNet	LCA	LCA-DSM
VSM	map	0.5142	0.4949	0.5255	0.5348
	bpref	0.8985	0.9318	0.9418	0.9406
	RR	0.8537	0.7726	0.8889	0.8889
BM25	map	0.5033	0.4873	0.5262	0.5459
	bpref	0.8985	0.9318	0.9660	0.9712
	RR	0.8992	0.8294	0.8253	0.8944

Tabela 3. Results for LISA dataset

Model	Metric	Baseline	WordNet	LCA	LCA-DSM
VSM	map	0.2641	0.2034	0.2475	0.2602
	bpref	0.9981	1.0	1.0	0.9981
	RR	0.5184	0.4618	0.5006	0.5038
BM25	map	0.3495	0.2520	0.3577	0.3627
	bpref	0.9981	1.0	1.0	0.9981
	RR	0.6459	0.5085	0.6400	0.6693

Tabela 4. Results for NPL dataset

Model	Metric	Baseline	WordNet	LCA	LCA-DSM
VSM	map	0.1886	0.1428	0.1968	0.2282
	bpref	0.9767	0.9886	0.9878	0.9333
	RR	0.4437	0.3583	0.5014	0.4267
BM25	map	0.2124	0.1756	0.2640	0.2580
	bpref	0.9766	0.9819	0.9891	0.9815
	RR	0.5987	0.5432	0.6142	0.6373

The MAP corresponds approximately to the area under a noninterpolated recall-precision curve (AU-RPC) providing the single score of these tradeoff.

3.3. Results

Tables 2, 3 and 4 present the results of the experiments for the MED, LISA and NPL datasets. Each table shows four results columns: baseline, wordnet, LCA, and LCA-DSM. The baseline column refers to information retrieval by a query without expansion. The results in the wordnet column refer to query expansion with a wordnet thesaurus. The LCA column records the results where query expansion is based on the local context using thesaurus wordnet. Finally, the LCA-DSM column applies semantic distributional representation to the local context analysis to construct a thesaurus and uses this thesaurus in the query expansion. The experiments were performed using the VSM and BM25 models.

These tables clearly show two directions. First, the results for the BM25 models are consistently better than those for the VSM models for all datasets. And also that the results of the experiments that use LCA are consistently superior to those using the global thesaurus. On the other hand, when local context analysis is used associated with distributional model an additional improvement is achieved in almost all cases. The consistency of these results can be examined through the precision-recall curve. Figure 2 shows the

precision-recall curve for the dataset MED. Note from these graphs that the BM25-LCA-DSM combination is consistently above the others which results in a larger area under the precision-recall curve (AUPR).

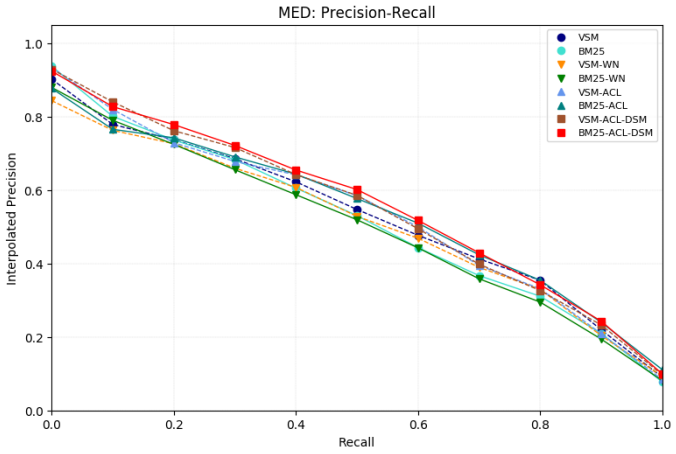

Figura 2. Precision x Recall curve for MED dataset.

4. Conclusion

Distributional semantic models are algorithms that draw their strength from the use of large linguistic corpus to construct dense representations of words which capture their meanings from the use. The hypothesis is that large linguistic corpus contains the records of contexts of language use from which the meanings of words can be extracted and represented. This work showed that the distributional semantics algorithms can also be used in local contexts with representation gain over those of the models based only on frequency of use.

The tests and strategy used are designed for closed collections of documents. A natural evolution of this work is to extend it to information retrieval on the web.

Acknowledgment

The authors would like to thank...

Referências

Azad, H. K. and Deepak, A. (2017). Query expansion techniques for information retrieval: a survey. *arXiv preprint arXiv:1708.00247*.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 688–695. ACM.

Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., and Petrelli, D. (2008). Hybrid search: Effectively combining keywords and semantic searches. In *European Semantic Web Conference*, pages 554–568. Springer.

Bhogal, J., MacFarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.

Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32. ACM.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.

Claveau, V. and Kijak, E. (2016). Distributional thesauri for information retrieval and vice versa. In *Language and Resource Conference, LREC*.

Croft, W. B. (2002). Combining approaches to information retrieval. In *Advances in information retrieval*, pages 1–36. Springer.

Dahab, M. Y., Alnofaie, S., and Kamel, M. (2018). A tutorial on information retrieval using query expansion. In *Intelligent Natural Language Processing: Trends and Applications*, pages 761–776. Springer.

Diaz, F., Mitra, B., and Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 367–377.

Fox, E. (1990). Virginia disc one. *Blacksburg, VA*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*. Citeseer.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*. Cambridge university press Cambridge.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Miller, G. A. (1995a). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Miller, G. A. (1995b). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Ooi, J., Ma, X., Qin, H., and Liew, S. C. (2015). A survey of query expansion, query suggestion and query refinement techniques. *2015 4th International Conference on Software Engineering and Computer Systems, ICSECS 2015: Virtuous Software Solutions for Big Data*, pages 112–117.

Pal, D., Mitra, M., and Datta, K. (2014). Improving query expansion using wordnet. *Journal of the Association for Information Science and Technology*, 65(12):2469–2478.

SanJuan, E., Ibekwe-SanJuan, F., Torres-Moreno, J.-M., and Velázquez-Morales, P. (2007). Combining vector space model and multi word term extraction for semantic query expansion. In *International Conference on Application of Natural Language to Information Systems*, pages 252–263. Springer.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69. Springer-Verlag New York, Inc.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1):79–112.

Zhang, J., Deng, B., and Li, X. (2009). Concept based query expansion using wordnet. In *Proceedings of the 2009 international e-conference on advanced science and technology*, pages 52–55. IEEE Computer Society.