Query Expansion in Text Information Retrieval with Local Context and Distributional Model

Fabiano Tavares da Silva, José E. B. Maia
State University of Ceará - UECE
Brazil
{fabiano.tavares@aluno.uece.br} {jose.maia@uece.br}

Journal of Digital
Information Management

ABSTRACT: *The Semantic Distributional Model is based on the frequency of contexts of use of language terms in large open corpus such as the web, to establish similarity or the relationship between words. These relationships or similarities can be used to add terms when expanding queries. The idea explored in this paper is that, for queries in closed collections of text documents, a posterior filter based on the restricted vocabulary of the collection can improve the effectiveness of automatic query expansion. This idea is developed and evaluated in publicly available benchmarks presenting promising results.*

Subject Categories and Descriptors: [H.3.3 Information Search and Retrieval]; Query formulation: **[I.2.7 Natural Language Processing]** Text analysis **[F.4.2 Grammars and Other Rewriting Systems];** Grammar types

General Terms: Distributional Semantic Model, Information Retrieval, Local Context Analysis.

Keywords: Distributional Semantic Model, Information Retrieval, Local Context Analysis

Received: 14 July 2019, Revised 30 September 2019, Accepted 9 October 2019

Review Metrics: Review Scale- 0/6, Review Score-4.22, Inter-reviewer Consistency- 78.5%

DOI: 10.6025/jdim/2019/17/6/313-320

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 [17]. 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 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 [7].

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 [7].

Figure 1 shows the conceptual diagram of a typical ISR. Both documents and queries are transformed into a representation oriented to the IR algorithm. The collection of documents is previously transformed and indexed and the query is expanded before being transformed into the working representation. The figure also shows that after the initial retrieval of the documents the user can provide the system with relevancy feedback. In relevancy feedback the user informs which of the retrieved documents are indeed relevant in such a way that the ISR can thus perfect his search for that query.

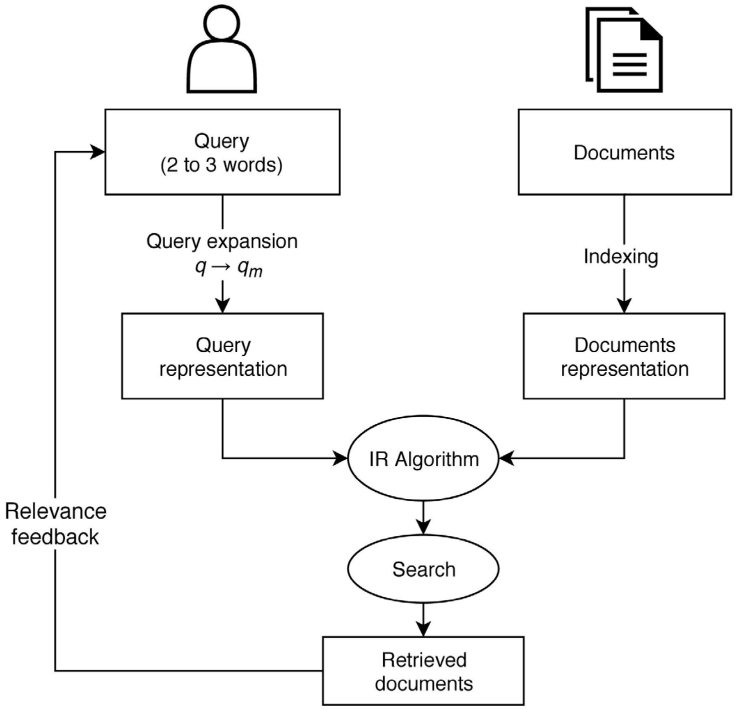

Figure 1. Conceptual diagram of a typical IRS

A broad classification of QE techniques identifies two alternatives for adding terms to the query and their combination [2]: 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 affect the expansion of the query. A thesaurus is a list of words with synonyms or similar meanings. The thesaurus can use the local or global scope.

The work in [27] is an example of a local strategy for QE named in it of Local Context Analysis [2]. In this method, an expanded query is formulated on the basis of some retrieved documents of search with the original query. Thus, the performance of the method depends on the accuracy of the top results. The correlation between the two terms is calculated using co-occurrence statistics. In contrast, [3] 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 these methods, a thesaurus built on a large external corpus is used for the QE, being Wordnet [19] the most famous. In [22, 28] and [25] the Wordnet thesaurus is used to expand queries with higher results in the med and news domains, respectively.

QE applications in specific domains can benefit from hybrid strategies. This is the case of the paper [15] in which the information of the most relevant documents of the initial query is combined with a thesaurus to expand the query in medical documents.

When using a thesaurus for query expansion there are some problems. The thesaurus is not dynamic. For some specific subdomains, they are not comprehensive. Context terms are not included. However, for closed collections, the thesaurus can play a good role in expanding queries.

The literature on query expansion is wide. Surveys can be found in [7, 1, 5, 10]. The works [23, 14, 4] 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 [11], word embedding was combined with the local context analysis, in [8] a global statistical thesaurus was constructed using the word2vec trained representation and the work [24] 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 the sacrifice of the reduction of precision in the recovered documents [7].

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

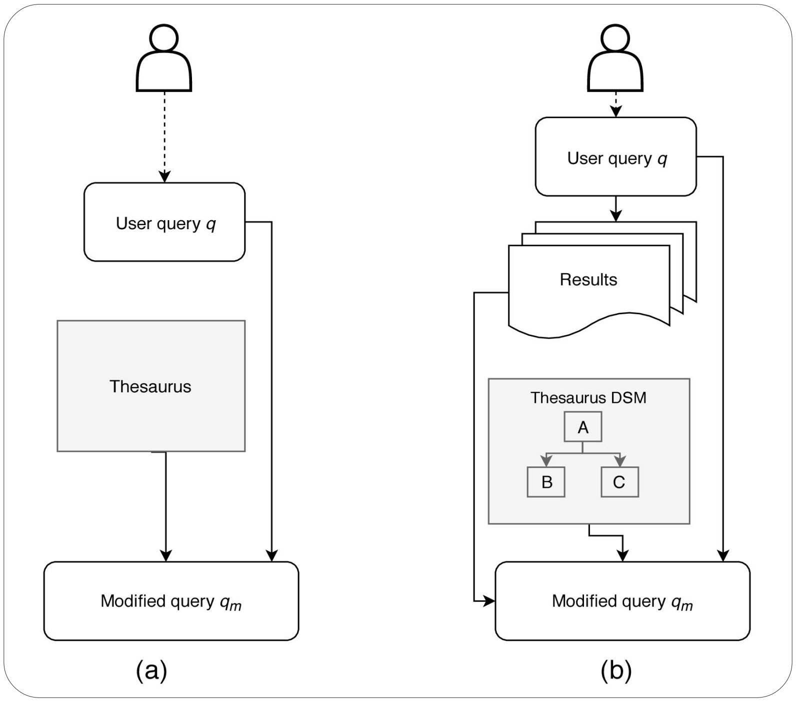

Figure 2. The flow of a query with and without expansion

this first-phase query is 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 2 describes the approach in detail, Section 3 presents the plan of the experiments, the datasets and the merit figures used and the results. The work is completed in Section 4 with the conclusion.

2. The Approach

The general method of query expansion is represented in Figure 2a. 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 [20] 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 [21]. On the other hand, the automatically constructed thesauri are based on the distributional hypothesis of [13] 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 [27], 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 the co-occurrence of concepts (groups of nouns) and by the 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 Figure 2b.

Figure 2b shows this process performed in two phases. After retrieving documents in an initial search the expanded query q_m is constructed using information from a thesaurus and the documents returned at the top.

The procedure is represented 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 (the 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 the context window considered and in the calculation of similarity.

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];
```

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 size depending 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 ($simqc(q,\ c,\ th)$, 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 [26, 9]:

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

In this equation, δ is a small constant (0.1 in [9]) to avoid the zeroing of the expression in some cases, $f(c,\ k_i,\ th)$ is a function that quantifies the correlation between a concept and a query term considering the distribution in the thesaurus $DSM\ th$, 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:

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

Dataset	Lang.	Matrix size (terms × docs)	No. of topics	Size (%) Top.X Doc.
MED	EN	7876 x 1033	30	14.2
LISA	EN	11710 x 6004	35	67.7
NPL	EN	7861 x 11429	100	26.0

Table 1. Datasets

In (2), $w_{t,q}$ 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 $f_{i,j} > 0$. The AQE process consists of using the terms of q to find new terms, without user participation, that solve problems of disambiguation which best discriminates the documents. The initial query q added by the new terms becomes q' and the similarity in (2) becomes $simqd\ (q',\ d)$. The choice of the new terms is made through a thesaurus in Algorithm 1. The global thesaurus ise automatically constructed using Semantic Distributional Model with Word Embedding representation [18].

3. Results and Discussion

3.1 Data

Results are obtained on three publicly available datasets in [12] 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 short size, 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 wase used, used by the TREC community: MAP, BP, MRR and the Precision-Recall curve. These metrics are precisely defined in [2, 17, 6]. 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. The MAP corresponds approximately to the area under a noninterpolated recall-precision curve (AU-RPC) providing the single score of this tradeoff.

3.3 Results

Tables 3, 4 and 5 present the results of the experiments for the MED, LISA and NPL datasets. Each table shows four results columns: baseline, wordnet, LCA, and LCADSM. 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 [16]. The LCA column records the results where query expansion is based on the local context analysis [27]. Finally, the LCA-DSM column applies semantic distributional representation (Word Embedding) to the local context analysis to construct a thesaurus and uses this thesaurus in the query expansion. The experiments were performed using the VSM (Vector Space Model) and BM25 (Problylistic Model) search models. Table 2 presents the parameterization of the algorithms.

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 global thesaurus. On the other hand, when local context analysis is used associated with the distributional model an additional improvement is achieved in almost all cases.

Taking MAP as a reference, which is a metric of the compromise between precision and recall, it can be seen from Tables 3, 4, and 5 that LCA-DSM is consistently superior to other methods in either the BM25 model or the VSM model. From Table 3, for example, the LCA-DSM MAP is 0.5459 versus 0.5033 from the model BM25 baseline for the MED dataset. These higher numbers are repeated in tables 4 and 5 for the LISA and NPL datasets.

Algorithm	Parameters
VSM	S: mode = 'OR', $R = TF$
BM25	S: mode = 'OR', $k1 = 2.0$, $k3 = 1.0$, $b = 0.75$ R = TF.IDF
WordNet	L = English, limit = 2 by term
LCA	$W = 300$, $N = 10$ No. passages: 50, $m = 5$
LCA-DSM	W = Dynamic select, $N = 10$ No. passages: 50, $m = 5$
word2vec	$W = 5$, Vector size = 300 CBOW $sg = 0$, min_count = 2, workers = 4

S - Select mode, R - Model of representation , W - Words window

Table 2. Settings of algorithms during the experiments

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

Table 3. Results for MED 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

Table 4. Results for LISA 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

Table 5. Results for NPL dataset

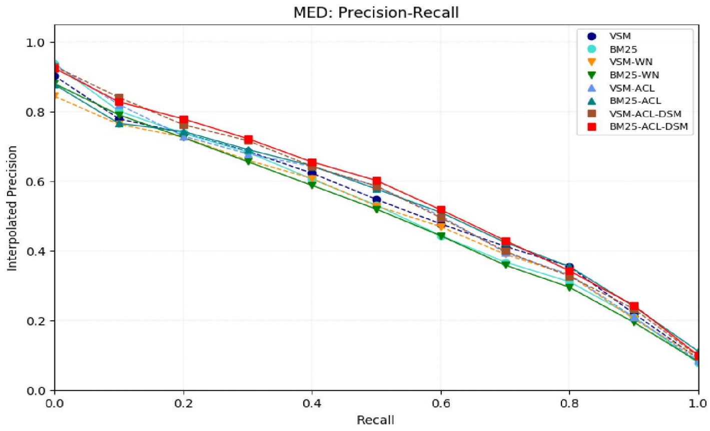

Figure 3. Precision x Recall curve for MED dataset

The consistency of these results can be examined through the precision-recall curve. Figure 3 shows the precisionrecall 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).

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 ex tracted 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 the frequency of use.

The tests and strategy used are designed for closed collections of documents. Experiments are conducted in controlled environments which are the datasets. A natural evolution of this work is to extend it to retrieving information on the live web where results can deviate.

References

[1] Azad, Hiteshwar Kumar., Deepak, Akshay. (2017). Query expansion techniques for information retrieval: a survey. arXiv preprint arXiv:1708.00247.

[2] Baeza-Yates, Ricardo., Ribeiro-Neto, Berthier. (1999). Modern information retrieval, vol 463. ACM press New York.

[3] Bai, Jing., Song, Dawei., Bruza, Peter., Nie, Jian-Yun, Cao, Guihong. (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, 688–695. ACM.

[4] Bhagdev, Ravish., Chapman, Sam., Ciravegna, Fabio., Lanfranchi, Vitaveska., Petrelli, Daniela. (2008). Hybrid search: Effectively combining keywords and semantic searches. *In:* European Semantic Web Conference, 554–568. Springer.

[5] Bhogal, Jagdev., MacFarlane, Andrew., Smith, Peter. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43 (4) 866–886.

[6] Buckley, Chris., EllenMVoorhees. (2004). Retrieval evaluation with incomplete information. *In*: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, p 25–32. ACM.

[7] Carpineto, Claudio., Romano, Giovanni. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys* (*CSUR*), 44(1) ,1.

[8] Claveau, Vincent., Kijak, Ewa. (2016). Distributional thesauri for information retrieval and vice versa. *In: Language and Resource Conference, LREC*.

[9] Croft, Bruce., W. (2002). Combining approaches to information retrieval. *In:* Advances in Information Retrieval, 1–36. Springer.

[10] Dahab, Yehia, Mohamed., Alnofaie, Sara., Mahmoud, Kamel. (2018). A tutorial on information retrieval using query expansion. *In:* Intelligent Natural Language Processing: Trends and Applications, 761–776. Springer.

[11] Diaz, Fernando., Mitra, Bhaskar., Craswell, Nick. (2016). Query expansion with locally-trained word embeddings. *In*: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol 1: Long Papers), 1, 367–377.

[12] Fox, E. (1990). Virginia disc one. Blacksburg, VA.

[13] Harris, Zellig. S. (1954). Distributional structure. Word, 10 (2-3) 146–162.

[14] Jiang, Jay., J., Conrath, David., W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *In*: Proceedings of 10th International Conference on Research in Computational Linguistics, ROCLINGâ A Z97. Citeseer.

[15] Khatoon, Thayyaba., Govardhan, A. (2018). Query expansion with enhanced-bm25 approach for improving the search query performance on clustered biomedical literature retrieval. *Journal of Digital Information Management*, 16(2).

[16] Li, Wei., Ganguly, Debasis., Jones, Gareth., J. F. (2016). Using wordnet for query expansion: Adapt@ fire 2016 microblog track. *In*: FIRE (Working Notes), p. 62–65.

[17] Manning, Christopher., D., Raghavan, Prabhakar., Schütze, Hinrich. (2008). Introduction to information retrieval. Cambridge University Press Cambridge.

[18] Mikolov, Tomas., Corrado, Greg., Chen, Kai., Jeffre, Dean. (2013). Efficient Estimation of Word Representations in Vector Space. *In*: Proceedings of the International Conference on Learning Representations (ICLR 2013).

[19] Miller, George., A. (1995). Wordnet: a lexical database for English. *Communications of the ACM*, 38 (11) 39–41.

[20] Miller, George., A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38 (11) 39–41, (November).

[21] Ooi, Jessie., Ma, Xiuqin., Qin, Hongwu., Liew, Siau Chuin. (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, p 112–117.

[22] Pal, Dipasree., Mitra, Mandar., Datta, Kalyankumar. (2014). Improving query expansion using wordnet. *Journal of the Association for Information Science and Technology*, 65 (12) 2469–2478.

[23] SanJuan, Eric., Ibekwe-SanJuan, Fidelia., Torres-Moreno, Juan- Manuel., Velázquez- Morales, Patricia. (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*, 252–263. Springer.

[24] Schütze, Hinrich. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24 (1) 97–123.

[25] Voorhees, Ellen., M. (1994). Query expansion using lexicalsemantic relations. *In*: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 61–69. Springer-Verlag New York, Inc.

[26] Xu, Jinxi., Croft, Bruce., W. (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, 4–11, New York, NY, USA, ACM.

[27] Xu, Jinxi., Croft, Bruce., W. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems* (TOIS), 18 (1) 79–11.

[28] Zhang, Jiuling., Deng, Beixing., Li, Xing. (2009). Concept based query expansion using wordnet. *In*: Proceedings of the 2009 international e-conference on advanced science and technology, 52– 55. *IEEE Computer Society*.