

# Query expansion based on modified Concept2vec model using resource description framework knowledge graphs

**Sarah Dahir, Abderrahim El Qadi**

Department of Applied Mathematics and Computer Engineering, National Graduate School of Arts and Crafts (ENSAM),  
Mohammed V University in Rabat, Morocco

---

## Article Info

### Article history:

Received Dec 28, 2021

Revised Oct 21, 2022

Accepted Nov 20, 2022

---

### Keywords:

Concept2Vector

Databasepedia

Information retrieval

Query expansion

Word embedding

---

## ABSTRACT

The enormous size of the web and the vagueness of the terms used to formulate queries still pose a huge problem in achieving user satisfaction. To solve this problem, queries need to be disambiguated based on their context. One well-known technique for enhancing the effectiveness of information retrieval (IR) is query expansion (QE). It reformulates the initial query by adding similar terms that help in retrieving more relevant results. In this paper, we propose a new QE semantic approach based on the modified Concept2vec model using linked data. The novelty of our work is the use of query-dependent linked data from DBpedia as training data for the Concept2vec skip-gram model. We considered only the top feedback documents, and we did not use them directly to generate embeddings; we used their interlinked data instead. Also, we used the linked data attributes that have a long value, e.g., "dbo: abstract", as training data for neural network models, and, we extracted from them the valuable concepts for QE. Our experiments on the Associated Press collection dataset showed that retrieval effectiveness can be much improved when a skip-gram model is used along with a DBpedia feature. Also, we demonstrated significant improvements compared to other approaches.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Abderrahim El Qadi

Department of Applied Mathematics and Computer Engineering

National Graduate School of Arts and Crafts (ENSAM), Mohammed V University

Rabat, Morocco

Email: a.elqadi@um5r.ac.ma

---

## 1. INTRODUCTION

The explosion and diversity of data on the Web has made information more available but difficult to use and less relevant to the user. As well the information retrieval systems (IRS) return a high number of unrelated results to the user because of the imprecise query and the query optimization issue [1]. To express the user's intention more clearly, a step of reformulating the queries is often necessary. Reformulation can be done by expanding queries, i.e., enriching them through adding new terms extracted by different possible term selection methods. This expansion of the original query can solve both the problem of the information's insufficiency in the user's query and the problem of vocabulary mismatch between the query terms and the documents' terms [2].

Query expansion was first proposed in 1960 by Maron and Kuhns [3]. It can be divided into; i) Global approaches, which are considered as query-independent, all documents are analyzed for all queries [4], ii) Local approaches, which find expansion words that are closely related to the original query words [5]. Recently, linked open data (LOD) knowledge bases are used to expand queries by taking into consideration the context. But the main challenges that can be faced is the lack of domain-specific LOD sources since

domain-independent sources do not cover a high number of technical terms. For instance, DBpedia Spotlight fails sometimes to annotate some specialized entities because they are not covered by it e.g., “operating system” (OS). Also, every specialized area has its vocabulary items (e.g., specific features) and characteristics that need to be taken into consideration [6]. In addition to this, many attributes from LOD have a whole list of values [7]. And it can be difficult to determine the adequate value to use for expansion regarding their size.

Also, Word embedding is applied to achieve semantic relatedness in IR, it allows predicting adjacent terms for a certain word or context by capturing term proximity and similarity [8]. One problem of embedding models is the fact that they vary in the quality of the generated embeddings especially that we are lacking metrics to evaluate the quality of embeddings. As for ontological concepts, embeddings of the entities *dbr: Mosco* (an entity of type *DBpedia PrivateCompany*), *dbr: Paris*, *dbr: Dublin* are supposedly close to the concept *dbo: City* (an entity of type *DBpedia ObjectProperty*) and far from entities such as *dbr: Barack\_Obama*, *dbr: Bill\_Clinton* which are associated with the concept *dbo: President*.

To resolve this issue, we propose a new query expansion method that relies on DBpedia concepts to generate the vectors. Our approach extends our previous work on association-based query expansion [9]. And instead of using cosine similarity only at the end of the embedding process as an evaluation metric of the embeddings' quality as in [10]; we use it before the embedding to insure having only semantically related concepts from the start in the training data. Moreover, our approach aims at reducing the number of expansion concepts. Also, it aims at expanding the remaining expansion concepts with their interlinked data using the same modified *Concept2Vec* model and a different DBpedia attribute.

Our paper is divided into 6 sections. Section 2 presents the preliminaries. Section 3 describes the related work. Section 4 details the proposed approaches. Section 5 presents the experimental results. Finally, section 6 concludes this paper.

## 2. PRELIMINARIES

This section presents the necessary background for understanding our proposal. The first element is the term frequency-inverse document frequency (TF-IDF) technique. The second element is the popular word embedding technique called *Word2Vec*.

### 2.1. Term frequency-inverse document frequency

TF-IDF is a popular statistical measure, which is widely used in text mining. It aims to weight the importance of each word in a particular document [11], [12]. Term frequency  $tf(t, d)$  stands for the number of times the term  $t$  appears in the document  $d$ . Inverse document frequency  $idf(t)$  depends mainly on the total number of documents in the corpus  $C$ , and the document frequency  $df(t)$  representing the number of documents that contain the term  $t$ . The term  $idf(t)$  is expressed as in (1),

$$idf_{(t)} = \log \left( \frac{N}{df_{(t)}} \right) \quad (1)$$

TF-IDF weight combines both the term frequency (TF) and the inverse document frequency (IDF) to estimate the weight for each term  $t$  in the document  $d$  as in (2),

$$tf-idf_{(t,d)} = tf_{(t,d)} \times idf_{(t)} \quad (2)$$

### 2.2. Word embedding

With the success of deep learning, several techniques based on neural networks have become increasingly popular to convert words into meaningful vectors. The so-called word embedding stands for a class of predictive models, which is commonly adopted to learn vector representations of words from a corpus of documents. It is designed to capture the context of words, semantic and syntactic similarity, and the relationship with other words. The underlying rationale behind it is that the distance between vectors determines the similarity in terms of meaning and semantic.

In general, *Word2vec* is a well-known word embedding technique, which was introduced by Mikolov *et al.* [13]. It has gained a great popularity due to its good performance. It is a shallow neural network model designed for learning distributed representations of words, where each word can be represented as a vector. The continuous representations of words are meaningful for a variety of real-world problems, including, machine translation, information retrieval, text classification, and so on.

In particular, there are two variants of *Word2Vec* called skip-gram model and continuous bag-of-words (CBoW) model,

- Skip-gram model [13], [14]: It is a neural network model, which is composed of three layers, called, input layer, projection layer, and output layer. It is designed to learn representations of words and predict the context (i.e., the surrounding words of a target word). Figure 1 shows the architecture of the skip-gram model.

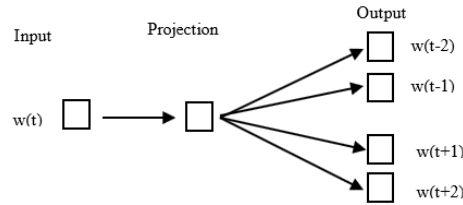


Figure 1. Architecture of the skip-gram model [13]

Let  $k$  be a parameter representing the size of the context window on a single side. Also, let  $w_{(1)}, w_{(2)}, \dots, w_{(t-1)}, w_{(t)}$  be a sequence of words. For the target word denoted  $w_{(t)}$ , the context window can be expressed by:  $[w_{(t-k)}, \dots, w_{(t-1)}, w_{(t)}, w_{(t+1)}, \dots, w_{(t+k)}]$ . The target word  $w_{(t)}$  is converted into a vector using one hot encoding by placing 1 in the position of that word and 0 for the other words, shows in Figure 1. There is only one hidden layer that performs the dot product between the weight matrix and the input vector  $w_{(t)}$  [15]. The result of the dot product at the hidden layer is passed to the output layer that computes the dot product between the output vector of the hidden layer and the weight matrix of the output layer. The softmax activation function is used to infer the probability of predicting the word  $w_{(t+j)}$  appearing in the context given the target word  $w_{(t)}$  as in (3),

$$p(w_{(t+j)} | w_{(t)}) = \frac{\exp(u_{w_{(t+j)}}^T v_{w_{(t)}})}{\sum_{l=1}^V \exp(u_l^T v_{w_{(t)}})} \quad (3)$$

$V$  represents the number of words in the vocabulary. The vectors  $u_w$  and  $v_w$  denote the input and output vector representations of the word  $w$ . This model aims to maximize the average log probability, which is written as in (4) [16],

$$\frac{1}{T} \sum_{t=1}^T \left[ \sum_{j=-k}^k \log p(w_{(t+j)} | w_{(t)}) \right] \quad (4)$$

where  $k$  is the size of the context window on a single side.

- Continuous bag of words (CBoW) model: It is a neural network architecture, which is an alternative to skip-gram model. It is designed to predict the target word in a sentence based on the context. CBoW is much faster than skip-gram and gives a better frequency for frequent words [13].
- Resource description framework to vector (RDF2Vec) [17]: it was introduced to learn embeddings from the resource description framework graphs after converting them into a set of sequences, because such algorithms require a propositional feature vector representation of data, where each instance is represented by a vector of features that are binary, numerical or nominal (symbols) [18].

### 3. RELATED WORKS

Many query expansions approaches have been studied, some of them focused on local and vocabulary analysis. Schütze *et al.* [4], authors analyze word occurrences and relationships in the whole corpus to automatically derive a thesaurus. They are inability to handle ambiguous terms from the query; because they process each query term separately from the others. As a result, the expansion of the query “apple computer” will cause a query drift. Alternatively, grammatical relations can be used, e.g., entities that are grown, cooked, eaten, and digested, are more likely to be food items. However, even if using grammatical dependencies is more accurate, using word co-occurrence is more robust because it cannot be misled by parser errors.

Rocchio [5], authors refer to the relevance feedback (RF) and pseudo-RF (PRF). RF is based on the user’s manual judgment on some of the retrieved documents and the use of this feedback information to expand the query. PRF is considered only the most retrieved documents as relevant, this may decrease the quality of results for difficult queries, in particular since the top retrieved documents may be irrelevant. In

other words, the efficiency of PRF depends directly on the quality of the feedback documents. Other related query expansion approaches used Word embedding and linked open data knowledge bases Table 1.

Table 1. Comparison between related work and our approach in terms of significant improvements

Related work	Related work	Limits	Significant improvements of our approach
Zamani and Croft [8] expanded queries using the relevance-based word embedding approach: Relevance likelihood maximization (RLM). They trained word embedding models using queries as well as the top-ranked feedback documents of each query.	Even though useful information may be captured from the top retrieved documents [19]–[21], pseudo relevance feedback (PRF) may decrease retrieval performance, especially for difficult queries, unlike relevance feedback (RF) [22] which is very effective in improving retrieval performance [5], [21].	None use of linked data has.	Use of training data that are related to the used queries instead of using random queries for training.
Baroni <i>et al.</i> [23] expanded queries either by: (i) directly using candidate terms that are closest to the query in the embedding space to compute the mean cosine similarity between a candidate term and all the query terms, or by (ii) first restricting the search domain for the candidate expansion terms to the top-ranked documents instead of the whole collection of documents then applying the previous approach.			Use of linked data in different stages of the process: both before and during the embedding.
Rattinger <i>et al.</i> [24] evaluated the quality of embeddings using the following metrics: (i) semantic relatedness based on human judgment, (ii) synonym detection which uses cosine similarity to compare the target word with all the choices displayed, (iii) concept categorization that groups concepts in taxonomic order, (vi) selectional preference that uses noun-verb pairs to capture the relevancy of a noun.	Cosine similarity is used as an evaluation metric only at the end of the embedding.		Cosine similarity is used early in the process, before the embedding, to make sure that the training data are relevant to the query.
Imani <i>et al.</i> [25] trained the skip-gram model either on the whole corpus or on the English-language edition of Wikipedia.	Since the used dataset is small, the default number of iterations is set from 5 to 20.		The top documents and the query are considered along with their interlinked data.
Kuzi <i>et al.</i> [26] suggested a deep expansion classifier (DEC) that used pre-trained word embeddings as inputs for the classifier that classified candidate expansion terms into good terms for expansion, bad terms, and neutral ones.	None exploitation of feedback documents.		Exploitation of feedback documents.
Lavrenko <i>et al.</i> [18] converted linked open data graphs into a set of entities' sequences using graph walks and Weisfeiler-Lehman Subtree RDFgraph kernels. Then they used those sequences to train a neural language model estimating the likelihood of an entities' sequence appearing in a graph.	The relation between the query and the feedback documents is not exploited.		Exploitation of the relation between the query and the feedback documents through the use of linked data.
Alshargi <i>et al.</i> [10] evaluated the quality of concepts in the embeddings based on: (i) the categorization aspect by considering the rdf: type (resource description framework) property, (ii) the relational aspect through considering a relation as valid based on the entity's types.	The enormous number of properties may make it difficult to judge the embedding only based on one property.		Cosine similarity is used early in the process, before the embedding, to make sure that the training data are relevant to the query.
Dahir <i>et al.</i> [27] used the query as a whole in CBoW to determine expansion terms from the entire list of feedback documents. Then, they integrated these terms with the pseudo-feedback-based relevance model (RM).	None use of external sources.		Use of DBpedia to obtain interlinked data.
Patel <i>et al.</i> [28] employed WordNet synonyms for the query title and DBpedia features for the query description field.	WordNet synonyms may lead to lower results if not picked carefully. Not all datasets have queries that have a query description field.		None use of WordNet. None use of any other field of the query.
Dahir <i>et al.</i> [9] proposed an approach, which is designed to annotate feedback documents of the initial query using DBpedia. Then, for each annotated DBpedia entity, the relevant "dct: subject" (an entity of type Thing) is adopted to find all entities having this subject as one of the "dct: subject" attribute's values using SPARQL protocol and RDF query language (SPARQL).	The SPARQL approach returns a long list of concepts and needs to be further exploited since it carries valuable information.		Use of the SPARQL approach results as training data for the skip-gram model.

#### 4. METHOD

This section details our proposal, which is an improved variant of our previous method "cosine-similarity (COS-SIM) on linked vectors" [9]. In particular, we proposed two approaches called Label-based modified Concept2vec Approach and abstract-based modified Concept2vec approach. Each one is based on

the skip-gram model illustrated in Figure 1 and a DBpedia feature (“dbo: abstract”) that was not exploited in the earlier work. Our modified Concept2vec approaches consist of the following steps, which are shown in Figures 2 and 3.

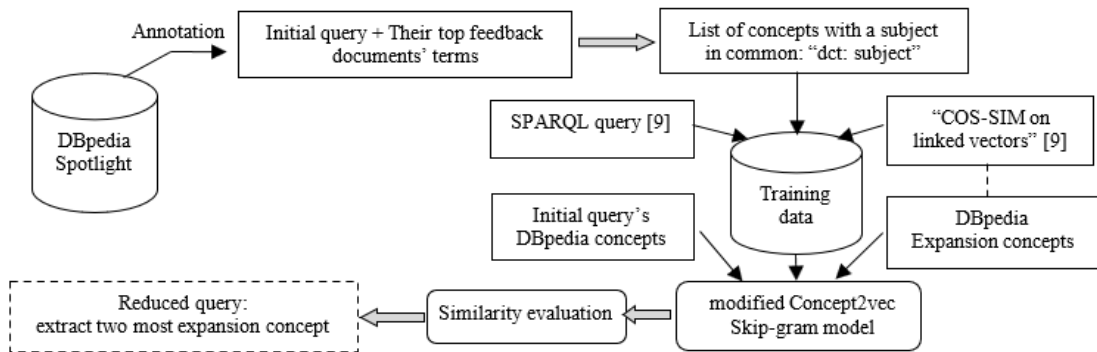


Figure 2. Flowchart of our proposed label-based modified Concept2vec approach

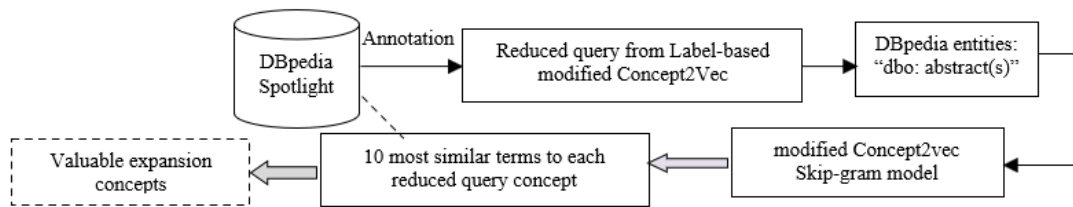


Figure 3. Flowchart of our proposed abstract-based modified Concept2vec approach

Label-based modified Concept2vec approach,

- Annotation of both the query’s expressions and the top associated feedback documents’ terms using DBpedia Spotlight;
- Determination, for each annotated DBpedia concept, of a “dct: subject” that contains the concept’s label. If there is not any, we use the first subject;
- Use of the SPARQL Protocol and RDF Query Language to find a list of concepts that have a determined dct: subject in common. It is the SPARQL [9] approach;
- Comparison between the lists of concepts (previous step) that are related to query concepts with those related to the top documents using cosine similarity to determine expansion concepts. It is the “COS-SIM on linked vectors” [9] approach;
- Use of the SPARQL [9] lists of concepts as well as the “COS-SIM on linked vectors” [9] expansion concepts as data for the modified Concept2vec model;
- Creation of the skip-gram model using "(3) and (4)". The number of context words we are looking at is 5 concepts before the input and 5 concepts after it;
- Evaluation of the similarity between the initial query concepts and the “COS-SIM on linked vectors” expansion concepts from DBpedia using the following line on python: Similarity(['QueryConcept'],'[COS-SIMonlinkeddataConcept]')

We used the "(5)" to calculate the similarity between the input vector A of the input word and the output vector B of the target word, and the normalization softmax [2] "(6)" to transform a set of given real values in the range of [0,1], such that the combined sum is " 1 " [29].

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \tag{5}$$

Where  $A_i$  and  $B_i$  are components of vector A and B respectively.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \text{ for } i = 1, \dots, K \text{ and } z = (z_1, \dots, z_k) \in \mathbb{R}^K \tag{6}$$

- Reduction of the “COS-SIM on linked vectors” expansion concepts to 2 based on the similarity results. In other words, only 2 expansion concepts from the earlier work [9], having the highest similarity with initial query concepts, are kept in the label-based modified Concept2vec approach shows in Figure 2. We opted for only 2 concepts because the number of expansion concepts in the “COS-SIM on linked vectors” [9] approach is already small.

Abstract-based modified Concept2vec approach,

- Annotation of the label-based modified Concept2vec expanded query using DBpedia;
- Determination of the DBpedia entities within the query from the previous step;
- Use of the found entities’ abstracts (i.e., values of “dbo: abstracts”) as the new data for the earlier skip-gram model (step 6);
- Determination of the 10 most similar terms to each concept from the previously reduced query. Table 2 shows an example of the results of the previous function on each DBpedia concept from the reduced query number 46 “tracking computer virus outbreaks” in Text Retrieval Conference Associated Press 88-90 (TREC AP88-90).
- Use of the DBpedia entities within the found terms, in the previous step, as expansion concepts for the abstract-based modified Concept2vec query shows in Figure 3. And further expansion of the expansion entities using their associated “rdf: type”, “dct: subject”, and domain-dependant ontology that have a maximum of 2 values.

The label-based modified Concept2vec method is used as a refinement of the earlier method to minimize the expansion terms and keep only the most efficient ones. Whereas the abstract-based modified Concept2vec method further expands the new reduced query by employing the long values of “dbo: abstract” in skip-gram. In other words, the abstract-based modified Concept2vec method is a continuation of the label-based modified Concept2vec method.

In DBpedia, annotated concepts can change depending on lower and upper case. Consequently, were annotated as two separate concepts. But it is clear that from the most similar terms in Table 2 that DBpedia does take into account the context. Thus, for “virus” we have terms like “software” and “malware” which are semantically related to the query.

Table 2. Ten most similar words to query number 46, of TREC AP88-90, using our abstract-based modified Concept2vec approach

Query concepts	Most similar terms and cosine similarity
computer	monster (0.2483), trojan (0.2199), vector (0.1844), fk (0.1701), beast (0.1609), strang (0.1530), fester (0.1492), mac (0.1396), dunihi (0.1123), comparison (0.1108)
virus	multipartite (0.2167), zero (0.1925), goat (0.1874), comparison (0.1382), multigrainmalwar (0.1346), download (0.1117), infect (0.1107), cooki (0.1069), rabbit (0.0908), softwar (0.0877)

## 5. RESULTS AND DISCUSSION

### 5.1. Dataset and evaluation measures

In this work, we used the collection of documents TREC AP88-90 [30] shows in Table 3 and 10 of its associated queries. We opted for a sample of queries because we believe it is enough to do the experiement. The length of these queries ranges from 1 to 6 keywords as in the previously mentioned query number 46 “tracking computer virus outbreaks”. Also, we used TF-IDF as a retrieval model.

Table 3. Description of the dataset

Number of documents	Average document size	Document relevancy	Topics (queries) numbers
158,240	261	0 (non-relevant), 1 (relevant)	1-50, 51-100, and 101-150

To evaluate the performance of our proposal, we employed three popular metrics, called, precision, recall, and mean average precision [31]. They are widely used to measure the prediction accuracy:

- Precision: it is defined as in (7),

$$\text{Precision} = \frac{\text{Number of relevant retrieved documents}}{\text{Number of retrieved documents}} \quad (7)$$

- Recall: it is known as the true positive rate. It shows the capability of the system for returning all the relevant documents. It is expressed as in (8),

$$\text{Recall} = \frac{\text{Number of relevant retrieved documents}}{\text{Number of relevant documents}} \tag{8}$$

mean average precision (MAP): For a set of queries, MAP represents the mean of the average precision (AP) scores for each query. It is represented as in (9),

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q} \text{ and } \text{AveP} = \frac{\sum_{k=1}^n (P(k) * \text{rel}(k))}{\text{Number of relevant documents}} \tag{9}$$

where Q is the number of queries. Rel(k) is an indicator function that is equal to 1 if the element at rang « k » is a relevant document and zero otherwise.

**5.2. Results and discussion**

This section is intended to compare the performance of the two proposed models named label-based modified Concept2vec approach and abstract-based modified Concept2vec approach with existing competitors. Figures 4-7 show the results obtained using our proposed query expansion approaches, based on modified Concept2vec and RDF knowledge graph. Figure 7 is intended to check the MAP of the methods in terms of the top retrieved documents (MAP@10). Since users mainly check the results within the first page (the top ones). Whereas Figure 6 checks the MAP of the whole set of retrieved documents and not only the top ones.

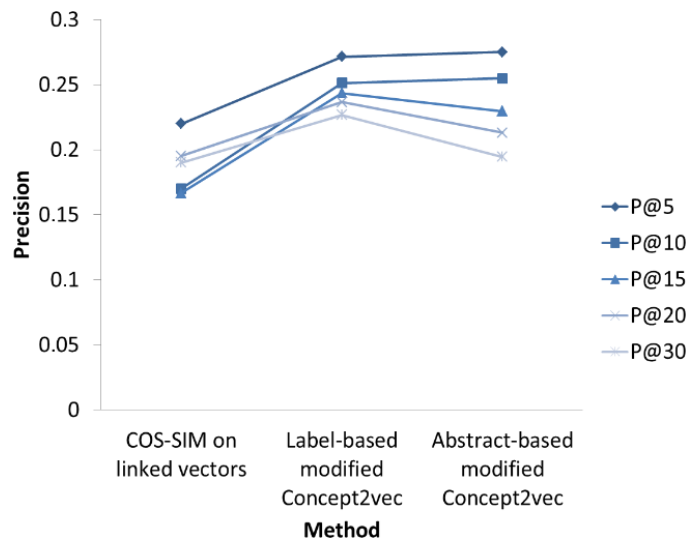


Figure 4. Precision at different cut-off ranks, for our two approaches and the baseline method

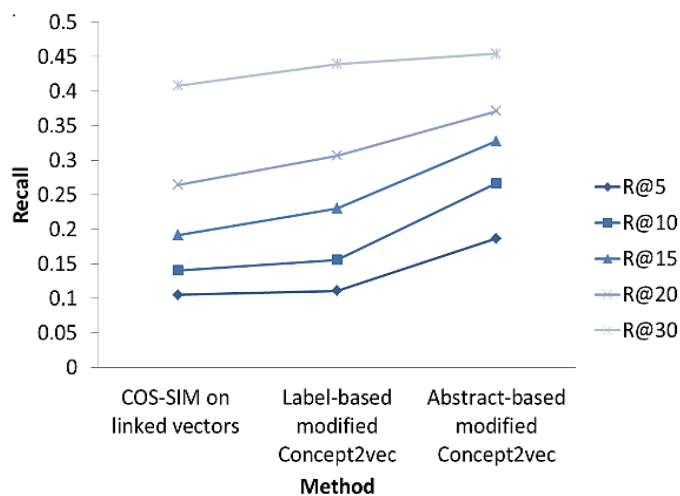


Figure 5. Recall at different cut-off ranks, for our two approaches and the baseline method

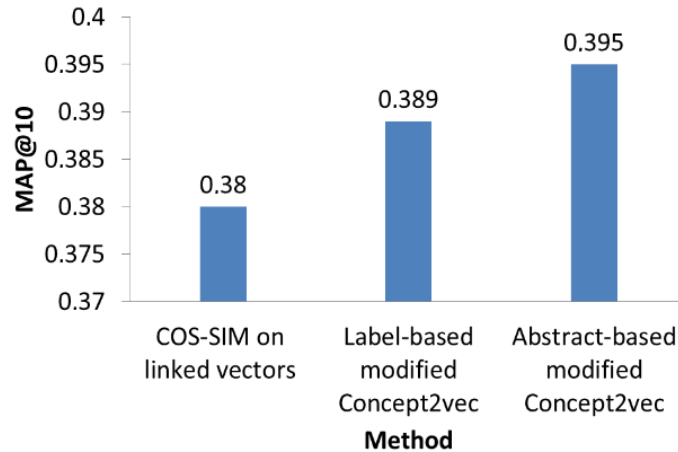


Figure 6. MAP@10 for our two approaches and the baseline method

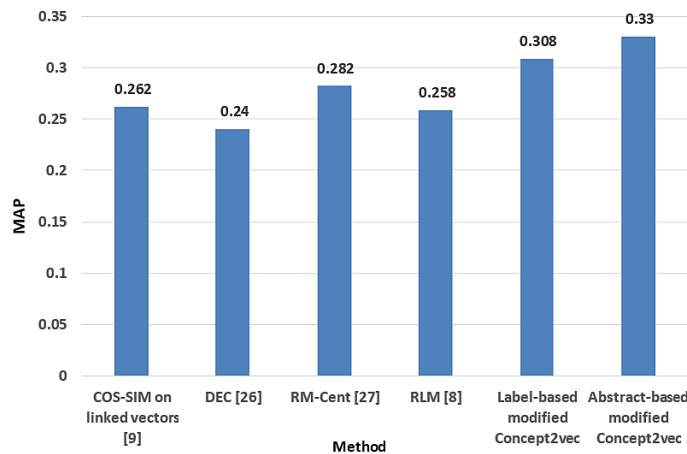


Figure 7. Results of the comparison of the proposed approaches with related works

From Figures 4-6, we noticed significant improvements in terms of precision at all cut-off ranks for both approaches compared to the baseline. Furthermore, the abstract-based modified Concept2vec bettered slightly the label-based modified Concept2Vec for P@5 and P@10 only. Whereas the label-based modified Concept2vec; improved the abstract-based modified Concept2vec in terms of P@15, P@20, and P@30. In addition, we have achieved significant improvements at all recall levels; in particular, for the abstract-based modified Concept2vec. For instance, the R@10 improved from 0.140 for the baseline to 0.156 for label-based modified Concept2vec and 0.267 for abstract-based modified Concept2vec. Moreover, the MAP@10 increased from 0.380 for the baseline to 0.389 for our label-based modified Concept2vec approach and 0.395 for the abstract-based modified Concept2vec approach.

From Figure 7, it is clear that our approaches outperform DEC [26] because although the expansion candidates are taken from an external source. They are semantically linked to feedback documents, whereas Kuzi *et al.* [26] no initial retrieval is performed. Consequently, good (if not better) terms from feedback documents are not exploited Kuzi *et al.* [26]. Another reason why our approaches perform better than DEC is our use of query-specific or dedicated training data. In other words, we used different training data for every query which is not the case for DEC where a general and non-domain specific corpus is used.

Similarly, RLM [8] leads to lower results because initial retrieval is performed on publically available queries and not on AP queries. Also, no semantic source is used nor dedicated training datasets are chosen depending on the query. The approach still performs slightly better than DEC since it uses a considerably large number of queries for training. As for RM-Cent [27], lower results are due to expansion using only initial retrieval results, i.e., external sources that may contain good interlinked terms are not used. However, it is clear that including initial retrieval corpus in expansion methods in general, e.g., COS-SIM on



linked vectors [9], and Word2Vec methods in particular, e.g., RM-Cent [27], gives better results than using a random and general training corpus e.g., DEC [26] and RLM [8].

In the previous work [9], the use of the “SPARQL” method to expand queries did not give good results. We believed that it was due to the high number of concepts that may be used in that case for expanding a query. In the label-based modified Concept2vec, we tried to benefit from the earlier work’s “SPARQL” method [9] by using separately each of its expanded queries as data for our modified Concept2vec method. We believe that the results improved because each used “SPARQL” expanded query is query dependent unlike available training data, used in other related work, that tends to be general and not specific to a particular query. However, the “SPARQL” [9] queries may be short in some cases which may not be beneficial for our method that requires large training data. As for the abstract-based modified Concept2vec method, it further improved the results because it employs the interlinked data of the already bettered queries. Furthermore, we believe that our choice of the attribute to use as data for skip-gram was successful since “dbo: abstract” is one of the DBpedia attributes that are always available no matter what the domain of the entity is. Also, “dbo: abstract” is always mono-valued. Consequently, we will not have multiple values to choose from. Moreover, this attribute is often long enough to be used as data for neural networks. And in the future, we suggest varying the Linked data attributes for the “SPARQL” method [9] to increase the size of the data.

## 6. CONCLUSION

This work is designed to improve the query expansion method “COS-SIM on linked vectors” using two modified Concept2vec methods based on the skip-gram model. One advantage of our work is the use of training data that are semantically related to the query and relevant to it. To our best knowledge, unlike existing works that rely on general training data for Word2vec like Google News articles, our approaches improved significantly the results of related studies. On the one hand, our label-based modified Concept2vec approach that uses long vectors from the SPARQL approach as training data helped to improve the results of “COS-SIM on linked vectors” through query reduction. This approach restricts the number of expansion concepts to 2 instead of keeping all concepts with a cosine similarity higher than 0 as in the earlier work. On the other hand, our abstract-based modified Concept2vec approach further improves the reduced query using DBpedia’s “dbo: abstract” of the entities within the new query as the new training data. Also, we judge better practice that the quality of word embeddings should benefit from bigger datasets. However, one limitation of our approach is the size of the data, which varies depending on the size of the vectors generated by the SPARQL query in the previous work, and on the size of “dbo: abstract” values. One way to alleviate this problem is the variation of used linked data attributes for the SPARQL query to increase the size of the data. And another way to improve our approach is by varying the parameter values of ‘size’ and ‘window’ and observing the variations in the results. These two perspectives represent the main topics of our current researches.




## REFERENCES

- [1] K. M. Wen, Z. D. Lu, X. L. Sun, and R. X. Li, “Research of semantic search: Overview,” *Computer Science*, vol. 35, no. 5, pp. 1–4, 2008.
- [2] W. Li, S. Wang, and Z. Yu, “Deep learning and semantic concept space are used in query expansion,” *Automatic Control and Computer Sciences*, vol. 52, no. 3, pp. 175–183, 2018, doi: 10.3103/S0146411618030082.
- [3] M. E. Maron and J. L. Kuhns, “On relevance, probabilistic indexing and information retrieval,” *Journal of the ACM (JACM)*, vol. 7, no. 3, pp. 216–244, 1960, doi: 10.1145/321033.321035.
- [4] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval,” *Introduction to Information Retrieval*, 2008, doi: 10.1017/cbo9780511809071.
- [5] A. Hamid, “Relevance feedback in information retrieval,” *the SMART Retrieval System*, no. October, pp. 313–323, 2017.
- [6] J. Choi, Y. Park, and M. Yi, “A hybrid method for retrieving medical documents with query expansion,” *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, pp. 411–414, 2016, doi: 10.1109/BIGCOMP.2016.7425959.
- [7] R. Abbes, A. Koplaku, K. Pinel-Sauvagnat, N. Hernandez, and M. Boughanem, “Apport du web et du web de données pour la recherche d’attributs,” *CORIA 2013-Conference in Research and Applications Informations-10th French Information Retrieval Conference*, 2013.
- [8] H. Zamani and W. Bruce Croft, “Relevance-based word embedding,” *SIGIR 2017-Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 505–514, 2017, doi: 10.1145/3077136.3080831.
- [9] S. Dahir, A. El Qadi, and H. Bennis, “An association based query expansion approach using linked data,” *9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018-Proceedings*, pp. 340–344, 2018, doi: 10.1109/ISIVC.2018.8709216.
- [10] F. Alshargi, S. Shekarpour, T. Soru, and A. Sheth, “Concept2vec: Metrics for evaluating quality of embeddings for ontological concepts,” *CEUR Workshop Proceedings*, 2018.
- [11] B. Liu, “Web data mining,” *Web Data Mining*, 2011, doi: 10.1007/978-3-642-19460-3.
- [12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *1st International Conference on Learning Representations, ICLR 2013-Workshop Track Proceedings*, 2013.
- [13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in Neural Information Processing Systems*, 2013.




- [14] S. Doshi, “Skip-Gram: NLP context words prediction algorithm,” *Medium*, 2019, [Online]. Available: <https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c>.
- [15] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” 2013, [Online]. Available: <http://arxiv.org/abs/1309.4168>.
- [16] P. Ristoski and H. Paulheim, “RDF2Vec: RDF graph embeddings for data mining,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9981 LNCS, pp. 498–514, 2016, doi: 10.1007/978-3-319-46523-4\_30.
- [17] P. Ristoski, J. Rosati, T. Di Noia, R. De Leone, and H. Paulheim, “RDF2Vec: RDF graph embeddings and their applications,” *Semantic Web*, vol. 10, no. 4, pp. 721–752, 2019, doi: 10.3233/SW-180317.
- [18] V. Lavrenko, M. Choquette, and W. B. Croft, “Cross-lingual relevance models,” *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pp. 175–182, 2002, doi: 10.1145/564376.564408.
- [19] V. Lavrenko and W. B. Croft, “Relevance-based language models,” *ACM SIGIR Forum*, vol. 51, no. 2, pp. 260–267, 2017, doi: 10.1145/3130348.3130376.
- [20] C. Zhai and J. Lafferty, “Model-based feedback in the language modeling approach to information retrieval,” p. 403, 2001, doi: 10.1145/502585.502654.
- [21] M. Levene, “Search engines: Information retrieval in practice,” *The Computer Journal*, vol. 54, no. 5, pp. 831–832, 2011, doi: 10.1093/comjnl/bxq039.
- [22] D. Roy, D. Paul, M. Mitra, and U. Garain, “Using word embeddings for automatic query expansion,” 2016, [Online]. Available: <http://arxiv.org/abs/1606.07608>.
- [23] M. Baroni, G. Dinu, and G. Kruszewski, “Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors,” *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014-Proceedings of the Conference*, vol. 1, pp. 238–247, 2014, doi: 10.3115/v1/p14-1023.
- [24] A. Rattinger, J. M. Le Goff, and C. Guetl, “Local word embeddings for query expansion based on co-authorship and citations,” *CEUR Workshop Proceedings*, vol. 2080, pp. 46–53, 2018.
- [25] A. Imani, A. Vakili, A. Montazer, and A. Shakeri, “Deep neural networks for query expansion using word embeddings,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11438 LNCS, pp. 203–210, 2019, doi: 10.1007/978-3-030-15719-7\_26.
- [26] S. Kuzi, A. Shtok, and O. Kurland, “Query expansion using word embeddings,” *International Conference on Information and Knowledge Management, Proceedings*, vol. 24-28-October-2016, pp. 1929–1932, 2016, doi: 10.1145/2983323.2983876.
- [27] S. Dahir, H. Khalifi, and A. El Qadi, “Query expansion using dbpedia and WordNet,” *ACM International Conference Proceeding Series*, 2019, doi: 10.1145/3333165.3333184.
- [28] R. Patel and S. Patel, “Deep learning for natural language processing,” *Lecture Notes in Networks and Systems*, vol. 190, pp. 523–533, 2021, doi: 10.1007/978-981-16-0882-7\_45.
- [29] A. Kulkarni and A. Shivananda, *Natural language processing recipes*, vol. 67, no. 1. Berkeley, CA: Apress, 2019.
- [30] “[https://trec.nist.gov/data/docs\\_eng.html](https://trec.nist.gov/data/docs_eng.html),” [Online]. Available: [https://trec.nist.gov/data/docs\\_eng.html](https://trec.nist.gov/data/docs_eng.html). (accessed Sep. 30, (2020))”
- [31] K. Zuva, “Evaluation of information retrieval systems,” *International Journal of Computer Science and Information Technology*, vol. 4, no. 3, pp. 35–43, 2012, doi: 10.5121/ijcsit.2012.4304.

## BIOGRAPHIES OF AUTHORS



**Sarah Dahir**    holds a Doctor of Informatics degree from Moulay Ismail University, Morocco in 2021. Her research includes ontology, Information Retrieval, Query Expansion, Document Reduction and Artificial Intelligence. She has published 9 papers in international journals and conferences. She can be contacted at email: [sarah.dahir2012@gmail.com](mailto:sarah.dahir2012@gmail.com).



**Abderrahim El Qadi**    is currently Professor in Computer Science, ENSAM, Mohammed V University in Rabat Morocco. He received his Ph.D. from the Faculty of Science, Mohammed V University-Rabat Morocco, in 2002, and his HDR in June 2010, in the subject: “Information Retrieval and Query optimization in Data warehouse”. His research interests include: Information Retrieval, Machine learning, Recommendation System, Context Social web. He can be contacted at email: [a.elqadi@um5r.ac.ma](mailto:a.elqadi@um5r.ac.ma)