Medical Query Expansion using Semantic Sources DBpedia and Wikidata

Sarah Dahir[a], Jalil ElHassouni[b], Abderrahim El Qadi[c], Hamid Bennis[a]

[a]*IMAGE Laboratory, SCIAM Team, Graduate School of Technology, Moulay Ismail University of Meknes, Morocco*
[b]*LRIT-CNRST (URAC'29), Faculty of Sciences, Rabat IT Center, Mohammed V University in Rabat, Morocco*
[c]*ENSAM, Mohammed V University in Rabat, Morocco*

Abstract

Since Query Expansion (QE) is known for its effectiveness in increasing query relevancy in IR Systems, and LOD are currently used in different domains for various objectives like suggesting, to users, alternative options based on features of their previous searches and interests. We suggest an approach to enhance Information Retrieval (IR) in the medical domain through QE using two Linked Open Data (LOD) bases: DBpedia and Wikidata. We use DBpedia entities within the PubMed abstract, as candidates for expansion, along with their associated labels ("rdfs:label") in DBpedia base. We evaluate our suggested approach, using MEDLINE collection and Indri search engine. Our expansion approach lead to significant improvements; especially in terms of precision and Mean Average Precision (MAP) compared to related approaches; using only one domain dependant/independent source.

Keywords

DBpedia, Information Retrieval, PubMed, Query Expansion, Wikidata.

1. Introduction

Information Retrieval Systems (IRS) match the user query to a collection of documents. As a result, a subset of documents is returned. This subset is considered relevant because it contains the query terms. But, sometimes words from the us[1]er query are different from those contained in the relevant document set. This issue has been shown in various studies; one from the medical field.

Covid-19 symptoms (fever, sore throat, shortness of breath, loss of taste, and loss of smell) as well as testing for coronavirus, and preventing measures (face mask, hand sanitizer, social distancing, and hand washing) have become some of the most trending queries, along with other search trends related to the aftermath of the pandemic on several other domains such as economy (e.g. unemployment and stock market) and education e.g. school closure [1]. For instance, queries on the loss of smell attained 8% in Mars 23rd, 2020 and testing for corona queries attained 97% on April 13th [1].

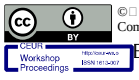
The lockdown caused by the pandemic, increased, more than ever, our need for better IR for medical queries in general. Especially that this type of queries lacks technical terms that domain experts use in web pages. This problem is often referred to as vocabulary mismatch.

One way to overcome this problem is to use query expansion. This process is done through adding new terms to the user query based on association rules between the terms [2]. However, adding so many terms to the query can be more harmful than adding few ones [3].

Linked Data[2] take advantage from the Web to connect related data [4]. For this purpose Uniform Resource Identifier (URI) and Resource Description Framework (RDF) are used among other technologies and Linked Data standards. Some of them are open and others require a license agreement:

- DBpedia: is a knowledge base that contains structured information from Wikipedia. This knowledge base describes 6 million entities; including 5000 diseases [5]. And it allows among other things: annotation of a text through the Web interface DBpedia Spotlight[3] that performs Named Entity Recognition. Yet, we noticed throughout our multiple accesses to DBpedia Spotlight that the annotation stops functioning from time to time. For more preci-

ISIC'21: International Semantic Intelligence Conference, February 25-27, 2021, Delhi, India
EMAIL: sarah.dahir2012@gmail.com (S. Dahir);
abderrahim.elqadi@um5.ac.ma (A. El Qadi))
ORCID: 0000-0001-6000-2428 (A. El Qadi)



CEUR Workshop Proceedings (CEUR-WS.org)

[2] http://linkeddata.org/
[3] https://www.dbpedia-spotlight.org/demo/

sion, we were unable to annotate texts using this Web application for three times in four years. And whenever, it stops functioning, it stays that way for three to four days in a row.

- Wikidata: is one of the largest datasets. It is a free knowledge database project hosted by Wikimedia with 90,478,674 data items [6] (including concepts). Unlike other knowledge bases, Wikidata may be edited by users. Furthermore, it usually gives links that allow browsing the resource in other databases like MeSH, PubMed, Freebase, etc.

In this work, we suggest expanding queries using two linked data sources (DBpedia, Wikidata) along with a search engine (PubMed) that allows the search in the MEDLINE database, and the National Library of Medicine (NLM) controlled vocabulary thesaurus (Medical Subject Headings (MeSH)) which is used to index PubMed articles.

This paper is organized as follows: Section 2 discusses related work. Section 3 gives methodological details of our suggested approach, and section 4 presents its evaluation results, and gives an outlook on future work.

2. Related work

Query Expansion (QE) plays a crucial role in improving Web searches. The user's initial query is reformulated by adding additional meaningful terms with similar significance. There are many queries expansion techniques:

Linguistic analysis [7] - [8]: deals with each query keyword separately from the others using for example the lexical database WordNet [9] - [10] - [11] that has a limited coverage of concepts [12] and a very small number of relationships (synonyms, hypernyms, and hyponyms). Consequently, this kind of techniques cannot solve ambiguity issues [13];

Query-log analysis: exploits log files' information of earlier queries; like the click activity of the user. But, this technique requires large logs [14];

Linked Data techniques [15]: take into consideration the context of keywords. In [16] authors explore just a small number of Dbpedia properties which means that important properties may not have been exploited. In [17] and [18] DBpedia is used to expand queries by using indexed terms from feedback documents that share similar DBpedia features with query terms.

In the medical domain; Linked Data allow corresponding terms used by patients to those used by domain experts. In [19], authors used the "Unified Medical Language System" (UMLS) database to determine synonyms for phrases within the user query.

In [20], authors expanded medical queries using only MeSH thesaurus. After that, they extracted documents based on the similarity between those expanded queries and clusters of medical documents. And In our previous work [21], we used attributes (features) values from Wikidata to expand medical queries. For this purpose, we considered only values that contained a query term. However Wikidata is not domain specific. Thus it lacks emphasis on the medical data. But, since Wikidata has links to numerous ontologies and databases from different domains, we decided to exploit one of those links that is specific to the medical domain. It is the PubMed's link.

3. Proposed method

Be it domain dependant or independent, linked data or not; every external source has its advantages, its limits, and its specificities. As a result, we suggested in this work a medical query expansion approach (Figure 1) that combines various sources; including two knowledge bases from Linked Data, a medical database, and a medical thesaurus as explained in the following steps:

1. We first look for the longest n-gram that covers most (if not all) DBpedia entities within the query and returns results in the Wikidata search engine. In case the n-gram does not feature all of the entities within the query; use other n-grams too; featuring those entities. Table 1, shows an example of used queries from the MEDLINE collection. Most of those queries are long (more than 4 keywords) and consist from many sentences. As a result, we must shorten them to avoid not getting any results at all and to make sure that we kept the most valuable keywords while shortening them.
2. Then, we search the n-gram(s) in Wikidata.
3. After that, we browse the PubMed identifier ("PubMed ID") of the first result in Wikidata.
4. Next, we perform Named-entity Recognition on the PubMed abstract, of the previously browsed page, using DBpedia.
5. Then, we consider DBpedia entities within the PubMed abstract, as candidates for ex-

pansion; along with their associated labels ("rdfs:label") in DBpedia.

6. Finally, we expand the query using the entities as well as their associated labels, from the previous step, that are also available in the MeSH terms of the PubMed page.

Table 1
Example of a long query, from MEDLINE dataset, containing several sentences.

Query number	Query content
14	renal amyloidosis as a complication of tuberculosis and the effects of steroids on this condition. only the terms kidney diseases andnephrotic syndrome were selected by the requester. prednisone and prednisolone are the only steroids of interest.

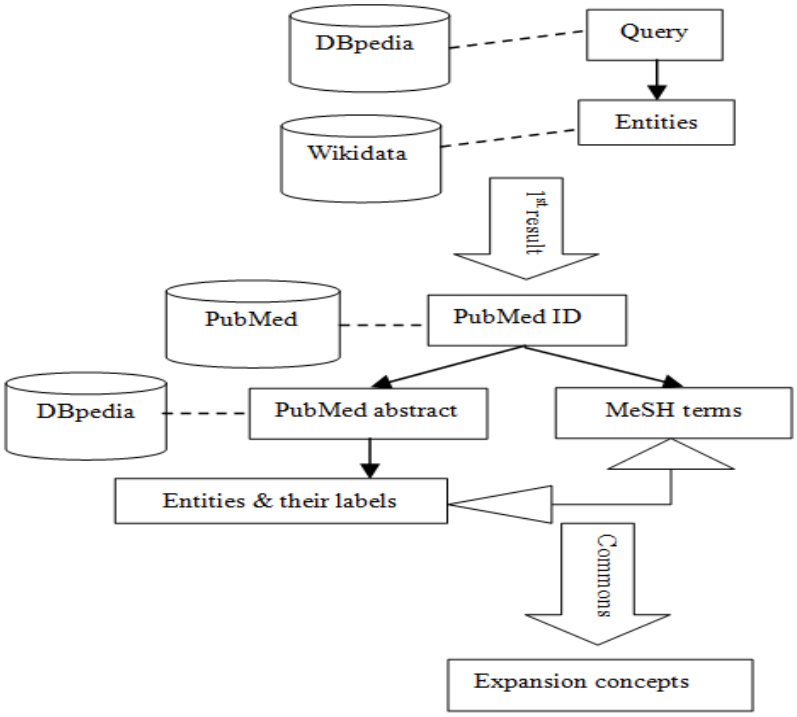


Figure 1: The flowchart of our suggested Query Expansion approach

4. Results and discussion

To evaluate our approach, we used MEDLINE (table 2) collection. It is a set of articles from a medical journal that we indexed with a stop words list using Indri search engine.

Table 2
Description of MEDLINE collection

Total number of texts	1033
Number of topics	30
Total number of tokens	159970
Total number of distinct (unique) tokens	13113
Average number of tokens per text	100

4.1. Retrieval model

For the implementation of our approach, we used Kulback Leibler(KL)[22] IR model [23]. In KL (1), we compare the document's model with the query's model.

$$D_{KL}(P||Q)=\sum_{x\in X} P(x)log\ (\frac{P(x)}{Q(x)}) \quad (1)$$

Where P and Q are discrete probability distributions defined on the same probability space.
And we use smoothing through Dirichlet to avoid getting a null result when a term is not present in the created language model.

4.2. Evaluation metrics

In this work we used the following evaluation measures:

- Precision (2): is a measure that indicates how efficient a system is in retrieving only relevant documents [24]:

$$\text{Precision} = \frac{\text{Number of relevant retrieved documents}}{\text{Number of retrieved documents}} \quad (2)$$

  Precision at rank N is evaluated by considering only top results returned by the system.
- Mean Average Precision (MAP) (3): The MAP for a set of queries is the mean of the Average Precision (AP) scores for every query [25].

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q} \quad (3)$$

  Where Q is the number of queries, and:

$$\text{AveP} = \frac{\sum_{k=1}^{n} (P(k) * rel(k))}{\text{Nombre de documents pertinents}} \quad (4)$$

  Where rel(k) is equal to 1 if the element at rank « k » is a relevant document, and zero otherwise [25].
- Normalized Discounted Cumulative Gain (nDCG) (5): measures the quality of the ranking by dividing the Discounted Cumulative Gain (DCG) by the Ideal Discounted Cumulative Gain (IDCG) [26].

$$\text{NDCG}_P = \frac{\text{DCG}_p}{\text{IDCG}_p} \quad (5)$$

  Where:

$$\text{DCG}_P = \sum_{i=1}^{p} \frac{2^{rel_i - 1}}{\log_2(i+1)} \quad (6)$$

  With rel_i: the relevance score of document i; is obtained after documents retrieval using an IR model. And:

$$\text{IDCG}_P = \sum_{i=1}^{|REL|} \frac{2^{rel_i - 1}}{\log_2(i+1)} \quad (7)$$

  Where |REL| is the list of relevant documents ranked based on their relevancy in the corpus.
- Mean Reciprocal Rank (MMR) (8): The Reciprocal Rank (RR) is the multiplicative inverse of the rank of the first exact answer [27]. And the MRR is the average of the RR of multiple queries Q [27].

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (8)$$

  Where $rank_i$ is the rank position of the first relevant document for the i-th query [27].

4.3. Results and discussion

To evaluate our method (see Table 3, 4 and figure 2) we compared it first with "Wikidata expansion approach" [21]. As we consider this work to be quite comparable to [21], since both works use Wikidata and are suitable for long queries. Also, we compared our approach with a non expansion approach (baseline) and a DBpedia method that uses DBpedia labels of entities within the query for expansion. Second, we compared our work with "Clusters' Retrieval Derived from Expanding Statistical Language Modeling Similarity and Thesaurus-Query Expansion with Thesaurus" (CRDESLM-QET) [20] because it uses MeSH terms and is thus comparable to our work.

We chose to compare the approaches at 30 for most evaluation measures and 10 or 20 for the precision because users are more interested in the top results.

Table 3
Comparison between "Wikidata expansion approach" [21] and our suggested query expansion approach using KL retrieval model on MEDLINE collection in Indri search engine.

Approach	P@20	MAP@30	MRR@30	NDCG@30
Baseline	0,461	0,446	0,818	0,637
DBpedia	0,483	0,450	0,837	0,645
Wikidata [21]	0,471	0,442	0,821	0,627
Our approach	0,525	0,500	0,844	0,671

Table 4
Comparison between our approach and an approach from related work [20]

Approach	P@10	MAP
CRDESLM-QET [20]	0,500	0.361
Our approach	0,600	0,567

Figure 2 shows the impact of using low and high values of C in P@20, we varied the number of expansion concepts to C=1, 2, 5, 10, 15, and 20.

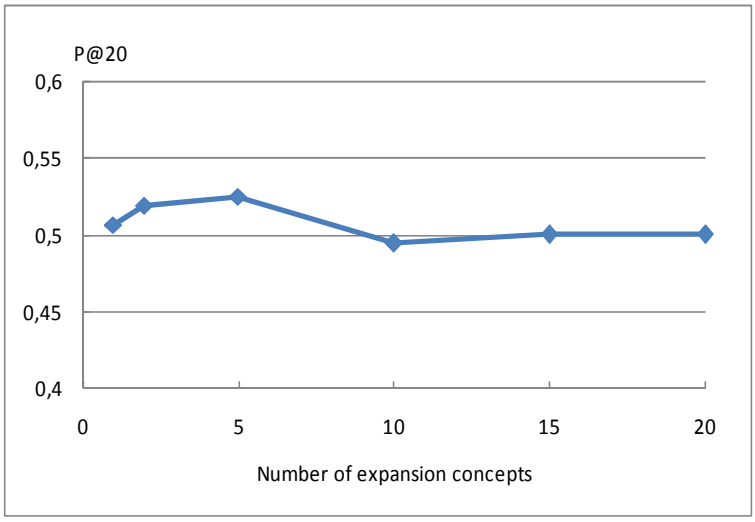


Figure 2: P@20 for different number of expansion concepts

Based on the results in table 3, our approach outperformed the state of art's work in [21].The use of KL to retrieve documents improved the P@20 of our "Multi semantic sources expansion" approach compared to the "Wikidata expansion approach" [21] with 5,4%. Also, our expansion approach in this work gave a 5,8% improvement in terms of MAP, a 2,3% improvement in terms of MRR, and a 4,4% increasement in terms of NDCG compared to the "Wikidata expansion approach" [21]. Similarly, our approach increased the results of both the baseline and the DBpedia approach that performs better than Wikidata.

From table 4, our approach outperforms CRDESLM-QET [20] in terms of P@10 with 10% and improves the MAP of CRDESLM-QET [20] with 20,6%.

From figure 2, we noticed that using lower numbers of concepts, especially C=5, leads to better results compared to using higher numbers.

We think that by increasing the number of expansion terms, we increase the possibility of adding non relevant terms to the query.

We believe that DBpedia improves Wikidata results because in the DBpedia approach we use labels that carry important information for the extraction of documents that are relevant but use different terms to refer to the user's query. Whereas the Wikidata approach uses terms that may lead to the extraction of documents that are related to the query but do not necessarily correspond to the user's intent.

Moreover, we reckon that our approach outperformed the Wikidata expansion approach [21] because unlike the previous work [21] that uses only Wikidata to expand queries, our multi semantic sources expansion approach benefits from several semantic sources, some of them are general or domain independent (DBpedia, Wikidata) and others are related to the Medical domain (PubMed, and MeSH). So, along with Wikidata, we decided to use, in this work, some domain specific databases by taking advantage from identifiers' links (e.g. PubMed ID) that are available in almost every Wikidata page of a certain resource or concept. And we had promising results because PubMed is one of the most valuable sources in the medical domain. Furthermore, our approach can be applied on queries of any domain by switching to other identifiers depending of the domain of the query.

As for CRDESLM-QET [20], it did not lead to high results because it uses only MeSH terms. Although MeSH terms are domain specific, they are very short (formed with few words) compared to PubMed abstracts. Also, MeSH is only a thesaurus that follows a tree structure. Consequently, it is not rich in terms of vocabulary compared to linked data sources.

In the future, we consider using other domain specific linked data sources, such as UMLS, for comparison purposes.

5. CONCLUSION

Throughout the lockdown that occurred, nearly, in all of the countries in a row, medical queries became

some of the most trending ones. As a matter of fact, the need for relevant search results in this particular domain, at this moment, pushed us to give more attention to this field and do research in it.

Our approach relies on various sources to determine expansion concepts. Two of these sources are LOD and others are: a search engine on medical databases (PubMed), and a controlled vocabulary (MeSH).

Since our suggested expansion approach that uses domain independent as well as domain dependant semantic sources outperforms our DBpedia approach and expansion approaches from earlier works [20] and [21], we may say that multiplying semantic sources in Automatic Query Expansion and exploiting domain specific sources, like PubMED and MeSH, helps in the improvement of retrieval results. Furthermore, using low numbers of expansion concepts helps in the improvement of retrieval results. Moreover, our new approach can be used for any collection of documents and not only for collections in the medical domain because Wikidata varies links (of identifiers) to a resource in other databases depending on the domain of the query.

In the future, we will try to further improve the results using other specific databases.

References

[1] Coronavirus search trends, Page consultée le 18/04/2020 à partir de : https://trends.google.com/trends/story/

[2] Bouziri, A., Latiri, C., Gaussier, É.: Expansion de requêtes par apprentissage. Conférence en Recherche d'Informations et Applications (2016)

[3] Keikha, A., Ensan, F., and Bagheri, E.: Query expansion using pseudo relevance feedback on wikipedia (2017)

[4] Linked open data, Page consultée le 18/04/2020 à partir de : https://wiki.digitalclassicist.org/Linked_open_data

[5] DBpedia version 2016-04 | DBpedia [Internet]. [cited 2020 Oct 31]. Available from: https://wiki.dbpedia.org/dbpedia-version-2016-04

[6] Wikidata [Internet]. [cited 2020 Oct 31]. Available from: https://www.wikidata.org/wiki/Wikidata:Main_Page

[7] Moreau, F., Claveau, V., and Sébillot P.: Automatic morphological query expansion using analogy-based machine learning. ECIR'07 - 29th Eur. Conf. Inf. Retr., pp. 222–233 (2007)

[8] Bhogal, J., Macfarlane, A., and Smith, P.: A review of ontology based query expansion. Inf. Process. Manag., vol. 43, no. 4, pp. 866–886 (2007)

[9] Jain, A., Mittal, K., and Tayal, D. K.: Automatically incorporating context meaning for query expansion using graph connectivity measures. Progress in Artificial Intelligence, Volume 2, Issue 2–3, pp. 129–139 (2014)

[10] Azad, H.K., and Deepak, A., A New Approach for Query Expansion using Wikipedia and WordNet. arXiv preprint arXiv:1901.10197 (2019)

[11] Dahir, S., Khalifi, H., & El Qadi, A.. Query Expansion Using DBpedia and WordNet. In Proceedings of the ArabWIC 6th Annual International Conference Research Track (pp. 1-6) (2019)

[12] Sinha, R., Mihalcea, R.: Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: Proceedings of ICSC (2007)

[13] Carpineto, C., and Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Comput. Surv., vol. 44, no. 1, pp. 1–50 (2012)

[14] Guisado-Gámez, J., Dominguez-Sal, D., and Larriba-Pey, J.-L.: Massive Query Expansion by Exploiting Graph Knowledge Bases for Image Retrieval. Proc. Int. Conf. Multimed. Retr., no. i, p. 33:33--33:40 (2014)

[15] Abbes, R. et al.: Apport du Web et du Web de Données pour la recherche d'attributs. Conférence en Recherche d'Information et Applications - CORIA (2013)

[16] Augenstein, I., Gentile, A.L., Norton, B., Zhang, Z., and Ciravegna, F.:.Mapping Keywords to Linked Data Resources for Automatic Query Expansion. The Semantic Web: ESWC 2013 Satellite Events. Lecture Notes in Computer Science, vol 7955. Springer, Berlin, Heidelberg (2013)

[17] Dahir, S., El Qadi, A., and Bennis, H.: Enriching User Queries Using DBpedia Features and Relevance Feedback. Procedia Computer Science. Vol.127 Issue C, pp. 499-504 (2018)

[18] Dahir, S., El Qadi, A., & Bennis, H. An Association Based Query Expansion Approach Using Linked Data. In 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC) (pp. 340-344). IEEE (2018).

[19] Le Maguer, S., Hamon, T., Grabar, N., and Claveau, V.: Recherche d'information médicale pour le patient Impact de ressources terminologiques. COnférence en Recherche d'Information et Appli-

cations, CORIA 2015, Mar 2015, Paris, France. Actes de la conférence CORIA (2015)

[20] Keyvanpour, M., & Serpush, F. (2019). ESLMT: a new clustering method for biomedical document retrieval. Biomedical Engineering/Biomedizinische Technik, 64(6), 729-741.

[21] Dahir, S., El Qadi, A., & Bennis, H. Query expansion using Wikidata attributes' values. In Third International Conference on Computing and Wireless Communication Systems, ICCWCS 2019. European Alliance for Innovation (EAI) (2019).

[22] Boughanem, M., Kraaij, W., and Nie, J.Y.: Modèles de langue pour la recherche d'information. In : Les systèmes de recherche d'informations. majid Ihadjadene (Eds.), Hermes-Lavoisier, Lavoisier, 11, rue Lavoisier 75008, pp. 163-182 (2004)

[23] Lemur Retrieval Applications. http://www.lemurproject.org/lemur/retrieval.php

[24] Common Evaluation Measures. https://trec.nist.gov/pubs/trec10/appendices/measures.pdf

[25] Wikipedia contributors, Evaluation measures (information retrieval). Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 23 Mar. 2019. Web. 17 Apr. (2019).

[26] Goharian, N., Information Retrieval Evaluation, COSC 488: https://www.coursehero.com/file/8847955/Evaluation/

[27] Wikipedia contributors. (2018, December 6). Mean reciprocal rank. In Wikipedia, The Free Encyclopedia. Retrieved 12:41, April 28, 2020 from https://en.wikipedia.org/w/index.php?title=Mean_reciprocal_rank&oldid=872349108