Query expansion using Wikidata attributes' values

Sarah Dahir[1], Abderrahim El Qadi[2], Hamid Bennis[3]
{sarah.dahir2012@gmail.com[1], elqadi_a@yahoo.com[2], hamid.bennis@gmail.com[3]}

High School of Technology of Meknès, Moulay Ismail University of Meknès, Morocco[1], High School of Technology of Salé, Mohammed V University of Rabat, Morocco[2], High School of Technology of Meknès, Moulay Ismail University of Meknès, Morocco[3]

Abstract. One of the well known techniques that are used to increase the relevancy of user queries is query reformulation through the expansion of queries. Also, Linked data are being used in various fields and available for different domains. This paper aimed at improving Information Retrieval results using Linked data from Wikipedia in general and Wikidata in particular to expand the queries with attributes' values. The experiments that we have done, using Medline collection and Indri search engine to compare our suggested approach with baseline, lead to improvements in precision at different ranks.

Keywords: Information Retrieval, Query Expansion, Linked Data, Wikipedia, Wikidata.

1 Introduction

Information Retrieval (IR) is based on a matching between the user's query and a collection of documents that results in returning a subset documents judged as relevant and that contain the terms of the query [1].

A common problem in this domain is the vocabulary mismatch problem that refers to the use of different terms, that can be synonyms, polysemes, or inflections, to refer to the same concept which may lead to low recall (i.e. the non retrieval of relevant documents) in the case of synonymy as well as inflections, and low precision (i.e. the retrieval of non-relevant documents) in the case of polysemy [2].

One of the well known methods to improve the relevancy of the results is: query expansion that is done through adding new terms to the initial query based on association rules between the terms [1]. Query expansion is either interactive i.e. expansion terms are suggested to the user and it is up to him or her to use them or not, or automatic which means that the expansion terms are automatically added to the query without notifying the user [3]. However, adding too many terms to the query can affect negatively the results of the expansion method more than adding few terms [3].

In this paper, we suggest using linked data from Wikidata as expansion terms by searching for corresponding Wikipedia [4] articles to determine valuable top links to use in the Wikidata knowledge base; for determining expansion terms from the available attributes values. For instance, Wikidata is a free collaborative and secondary database with 53,615,165 data items that can be read and edited by both humans and machines since it contains linked data and uses linked data standards [5].

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

ICCWCS 2019, April 24-25, Kenitra, Morocco

DOI 10.4108/eai.24-4-2019.2284070

2 Related work

In general, Automatic query Expansion (AQE) techniques can be classified into 4 categories:

- Linguistic analysis [6] - [7]: that can not solve ambiguity issues [2] since they deal with each term seperatly from the others using WordNet [4] - [8] for example. This lexical database which can be used in approaches that are ontology-based [9] - [10] has liabilites that include among others the limited coverage of concepts [11] as well as the very few number of the available relationships (e.g. synonymy) ;
- Query-log analysis: that exploits log files information e.g. the click activity of the users. The information in the logs of ancient queries that may be used to expand the user's query is the relationship between queries and selected documents [12]. In [13] the authors extract probabilistic correlations between query terms and document terms, by analyzing query logs, in order to determine expansion terms. Yet, this technique is not efficient for systems that do not have large logs [12]. In the last decades, the IR field has also known an integration of the context aspect to query expansion. In the work presented in [14], a new model for measuring similarity between web queries was proposed. The Language Model (LM) is used to build the query context, which is composed of the most similar queries to the query to expand and their top-ranked documents. Then, they apply a query expansion approach based on the query context and the Latent Semantic Analyses (LSA) method;
- Linked data techniques [15]: in the domain of health for example; linked data help in corresponding terms that patients use with those used by specialists of the medical domain. In [16], authors use the "Unified Medical Language System" (UMLS) database to dermine synonyms of users' queries phrases. In [17] authors show the importance of Linked Open Data Cloud properties in finding semantically similar alternatives, to user query keywords, to be used for expansion, but in there work they explore only a few number of Dbpedia [18] properties which means that valuable properties may not have been explored. In our work we do not select properties to use depending on their names, we select them depending on their attributes' values. In [3], authors use Wikipedia to solve ambiguity problems of queries buy finding correspondies articles to each query n-grams

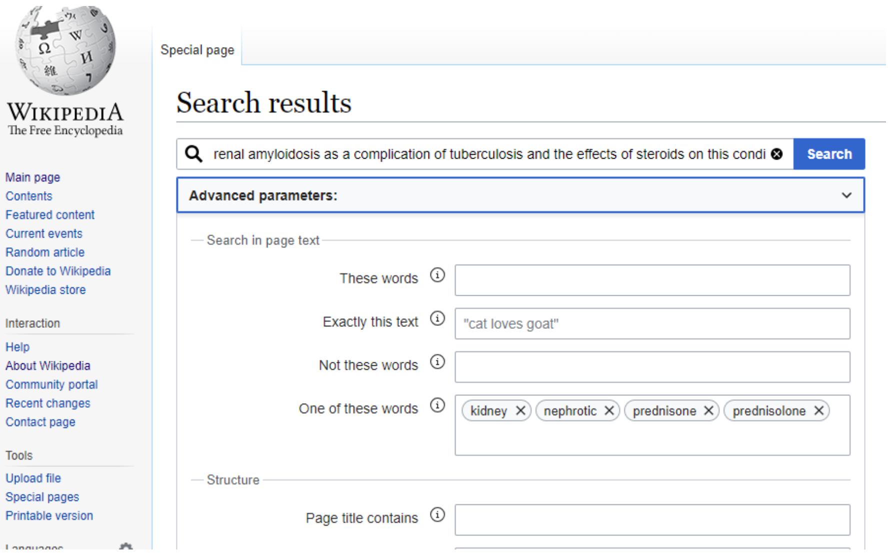
3 Proposed Method

To expand queries using Linked data from Wikipedia we:

1. Looked for the query in the English Wikipedia. When the query contained many sentences (table 1); we opted for the advanced search to write the first sentence in the search text field and write the terms in the other sentences in the "One of these words" text field (as shown in **Figure 1**) to avoid not getting any results;

Table 1. Example of used queries, from Medline collection, that contain more than one sentence.

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 and nephrotic syndrome were selected by the requester. prednisone and prednisolone are the only steroids of interest.



Search results

renal amyloidosis as a complication of tuberculosis and the effects of steroids on this condi

Advanced parameters:

Search in page text

These words:

Exactly this text: "cat loves goat"

Not these words:

One of these words: kidney ✕ nephrotic ✕ prednisone ✕ prednisolone ✕

Structure

Page title contains:

Fig. 1. Use of “One of these words” text field in the case of query number 14 that contains more than one sentence.

2. Selected from the top 10 links (because the most relevant results are believed to be the first ones): the titles that were equal to the query, the ones that contained a sub-part of it (if it was not a stop word), and the ones that contained a sub-part of the query written between parentheses but not preceded by expressions like: “redirects from”, “category”, and “section” ;
3. Used the “Wikidata item” link in the Wiki page (which is available in the left as one of the “tools” section links) to get its corresponding Wikidata page from which we determined: the attributes’ values, from the section “Statements”, that contained a query term (or a sub part of it), e.g. “renal carcinoma” which is the value of the property “subclass of”, unless it was a stop word and deleted from them the expressions “science/”, “subject/”, “-pro” (**Figure 2**).

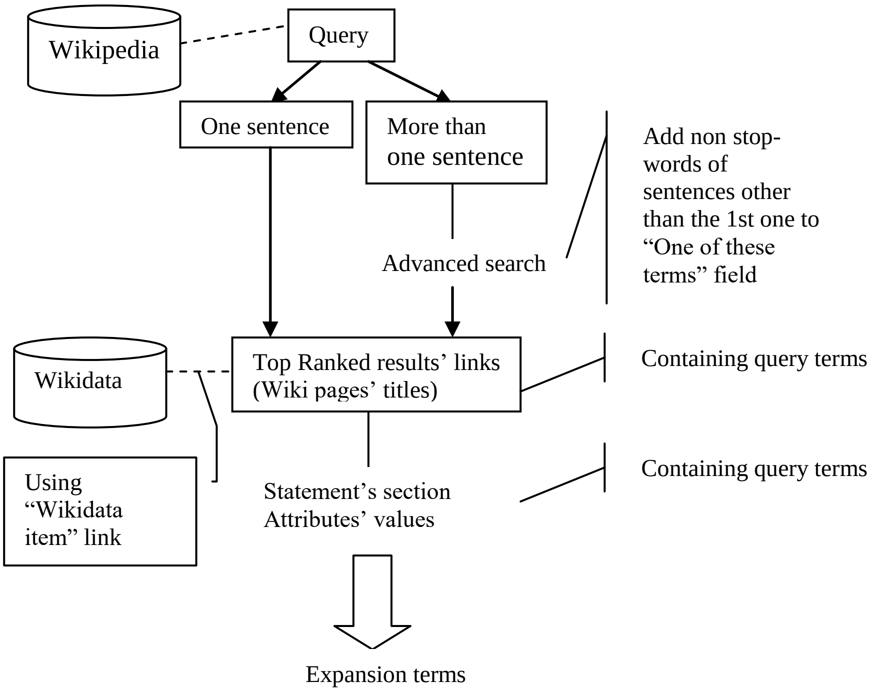

Wikipedia
Query
One sentence
More than one sentence
Add non stop-words of sentences other than the 1st one to "One of these terms" field
Advanced search
Wikidata
Top Ranked results' links (Wiki pages' titles)
Containing query terms
Using "Wikidata item" link
Statement's section Attributes' values
Containing query terms
Expansion terms

Fig. 2. The flowchart of our suggested approach

Our approach solved the problem of the determination of expansion terms' number to be used; since we used only the terms from Wikidata that contained a query term and not all of the attributes.

4 Results and discussion

To evaluate our approach, we used the collection Medline which is a collection of articles from a medical journal (containing 1033 documents, 30 queries and relevance assessments), that we indexed with a stop words list using Indri search engine.

4.1 Used retrieval model

In this work we used Kullback-Leibler (KL) IR language model for the implementation of our approach. And the general idea behind language models in IR, is about considering that a document D represents a sub-language for which we try to construct a language model M_D[19]. The score of D given a certain query Q is determined by the probability that the documents model generates the query. Or, we try to creat a language model for the query and give a score to a document depending on weather this document can be generated by the query's model [19]. Also, we may use smoothing e.g. Dirichlet to avoid getting a null result when a term is abscent in the constructed laguage model;

4.2 Evaluation metrics

Precision. is a measure (1) that shows how capable a system is of returning only relevant documents [20]:

$$Precision = \frac{\text{Number of relevant retrieved documents}}{\text{Number of retrieved documents}} \quad (1)$$

And Precision at rank N is evaluated by considering only top results returned by the system.

Recall. is a measure (2) that shows how capable a system is of returning all relevant documents [20]:

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

4.3 Results and discussion

To evaluate our method we compared it with a baseline for which we used the pre-defined KL model without any expansion of the queries (see **Figure 3**).

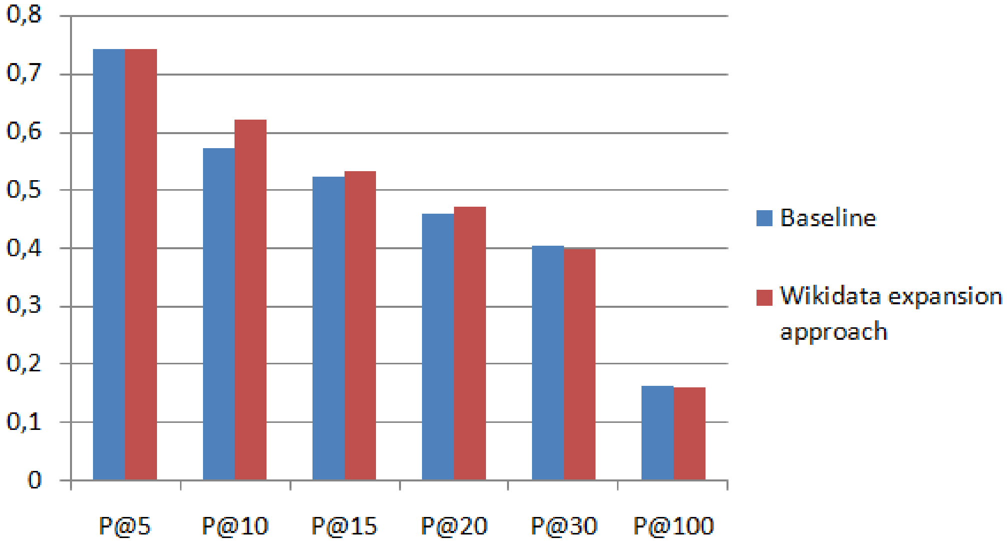


Fig. 3. P@N for our baseline and our suggested Wikidata query expansion approach that both use KL language model to retrieve results for Medline collection using Indri.

Fom the results, we noticed that the use of our suggested query expansion approach improved the baseline in terms of precision at different ranks. For instance, even though the P@5 of the baseline and the expansion method were the same; the P@10 of the suggested approach (which was 0,6214) was much better than that of the baseline (that was 0,5714). In addition to that, the P@15 of the expansion method (0,5333) was higher than that of the baseline (0,5238). And, the P@20 of the expansion approach that attained 0,4714 was better than that of the baseline (0,4607). Moreover, our approach increased slightly the recall at 0,00 since it was 0,8769 for the baseline and became 0,8867 for the Wikidata expansion approach.

Consequently, we believe that our approach is beneficial for Information Retrieval and query expansion; because it is simple and it benefits from linked data advantages through the use of Wikipedia in general and Wikidata in particular.

A possible reason for the unchanged results of P@5, might be the used collection of documents which is Medline that is domain specific and we know that the most appropriate database for this collection of documents is UMLS but the reason why we did not use it is the fact that we want our approach to be applied on any collection and not only on medicine collections. As a result, we think that we could have obtained higher results if we used another collection.

But in order to obtain even higher improvement; we think of using in future work, attributes' names instead of attributes' values and apply the method on other collections of documents. Also, the results vary depending on the retrieved number of documents, the more documents we retrieve; the higher the recall may get. And, the fewer documents we retrieve; the higher the precision may get.

5 Conclusion

To conclude, we may say that the use of Linked data in Automatic Query Expansion helps in the improvement of retrieval results.

Also, our suggested expansion approach that is based on the use of expansion terms from the Wikidata attributes' values increases the precision at different cut-off ranks and improves the recall at 0,00. Moreover, this proposed method is simple and can be used for any collection of documents and not only for collections of a specific domain like the medical domain because it uses Wikipedia and not a medical database like UMLS.

In the future, we will try to improve the relevancy of the results using attributes' names instead of attributes' values.

References

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

[2] 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)

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

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

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

[6] 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)

[7] 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)

[8] 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)

[9] Hajmoosaei, A., and Skoric, P.: Museum Ontology-Based metatdata. In Tenth IEEE International Conference on Semantic Computing, ICSC 2016, Laguna Hills, CA, USA, February 4-6,2016, pp. 100-103 (2016)

[10] Grieser, K., Baldwin, T., Bohnert, F., and Sonenberg, L.: Using ontological and document similarity to estimate museum exhibit relatedness. Journal on Computing and Cultural Heritage (JOCCH). Vol. 3, Issue 3 (2011)

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

[12] 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)

[13] Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y.: Probabilistic Query Expansion Using Query Logs. Proceedings of the 11th International Conference on World Wide Web (2002)

[14] Egozi, O., Markovitch, and Gabrilovich, E.: Concept-Based IR using Explicit Semantic Analysis. ACM Trans. Inf. Syst., vol. 0, no. 0 (2000)

[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] 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 Applications, CORIA 2015, Mar 2015, Paris, France. Actes de la conférence CORIA (2015)

[17] 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)

[18] 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)

[19] 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)

[20] https://trec.nist.gov/pubs/trec10/appendices/measures.pdf