MeSH-based Semantic Query Expansion

Wided Selmi
MIRACL: Computer science
FSEGS
Sfax, Tunisia
wided.selmi.fseg@gmail.com

Hager Kammoun
MIRACL: Computer science
FSS
Sfax, Tunisia
hager.kammoun@gmail.com

Ikram Amous
MIRACL: Computer science
ENET'COM
Sfax, Tunisia
ikram.amous@isecs.rnu.tn

***Abstract*— In the medical field, medical terms are not necessarily used in the original query. However, they are the key allowing the information retrieval. Query Expansion approaches try to address the problem by providing a way to add words having a "similar meaning" in the original query. In this paper, we discuss various Query Expansion approaches with some of our case study. We focus on the "semantic query expansion" approach which exploits the semantic medical resource MeSH. We propose a new approach that provides the opportunity to improve the information retrieval efficiency, measured by the mean precision at the top 5, 10 returned documents using the TREC EVAL software. This approach offers different treatments (by synonymy, by specialization, and by generalization) according to the entity MeSH nature of the term for the expansion, with the intent to improve the query specification and to evaluate the effect of these treatments. The results show an improvement of retrieval performance on both OHSUMED and Cystic Fibrosis collections.**

Keywords—Semantic query expansion; medical domain; MeSH thesaurus

I. INTRODUCTION

Information Retrieval (IR), or particularly Document Retrieval (DR), consists in selecting documents from a large collection according to criteria specified in the user's query expressing his information need. Thus, it is necessary to develop Information Retrieval Systems (IRS) capable of extracting the greatest amount of relevant documents satisfying the various needs and matching their current query.

In practice, the conventional IRS proposed is far from ideal for two main reasons. On one hand, for most users, it is difficult to formulate optimal and sufficient query relating and specifying the desired information need, which causes a noise by the retrieved irrelevant documents and a silence by the absence of the relevant documents in the list of the retrieved documents. This is especially true if such documents belong to a domain far from user's specialty. On the other hand, IRS were not sophisticated enough to process the query efficiently. The gap between the user's query keywords and the documents words is a well known problem in information retrieval community called "mismatch" problem: the indexers and the users do not use the same words. This is known as the vocabulary problem, compounded by synonymy (different words may represent the same meaning) and polysemy (same word may represents different meanings: ambiguity). A solution would be to introduce a query reformulation step which makes it possible to modify the original query or to expand it with additional information: words or expressions having a "similar meaning" and to generate a new "more adapted" one.

In our work we are interested in query reformulation approaches.

In the literature, the query reformulation has been widely studied in different ways. For instance, according to the role of the user, adding terms to query can be interactive or automatic. However, in the interactive approaches referred to as relevance feedback, and in response to the original query, the user is provided with a list of matched documents deemed relevant (system relevance). Hence, the ideal list of documents which are considered relevant by the user are selected. The system extracts terms from these documents that will be added to the original query to perform a new research [3]. It should be noted that it is possible to guide the user in choosing terms of the new query, by proposing terms derived from external resources [14]. For the automatic approaches, the query expansion is made automatically by using either external resources represented by a controlled vocabulary [5] or the selected relevant documents (Pseudo Relevance Feedback) [1], [17], [21].

Proposed techniques for query reformulation can be subdivided into two categories: local methods and global methods.

The local methods, meanwhile, is based on the initial retrieval results. This strategy selects terms from the top ranked documents in response to the initial query (pseudo relevance feedback) [28], [17] or from marked relevant documents by the user (relevance feedback) [25], [26]. One of the classical algorithm for implementing relevance feedback was done by (Rocchio, 1971) [25] which is based on the vector space model. Rather than reweighting the query in a vector space, one way of doing this is with a probabilistic model. (Robertson et al., 1996) [24] proposed an automatic query expansion algorithm in conjunction with Okapi system. It is similar to Rocchio's, while using a different term weighting functions.

The global methods are techniques for expanding queries, which don't take into account the results obtained from the original query. It is also known as "query-independent"

978-1-5386-9120-5/18/$31.00 ©2018 IEEE

method. Global analysis can be classified into two categories: corpus-based and linguistic-based.

Corpus-based approaches refer to the use of information previously established in the collection. They are statistical approaches for query expansion. Some of these approaches use the term co-occurrence [6], the mutual information [16].

Linguistic-based approaches are the semantic approaches who exploit external resources, such as terminology, thesaurus or ontology. Some of them express general knowledge, such as WordNet, Cycl, YAGO, etc. while others express specific knowledge, like MeSH, SNOMED-CT, and UMLs that consider the medical field.

Our research focuses on the medical field; the latter is a center of interest of a large population of specialists and non-specialists. Our goal is to introduce the Medical Subjects Headings (MeSH) thesaurus as the knowledge resource into the query expansion process.

According to the hierarchical structure of these resources, terms in semantic relationships (synonym, hypernym, hyponym) with the terms of the original query are to be extracted and added. Indeed terms of the original query can be interpreted in multiple meaning depending on the context in which they appear (ambiguities of language). Therefore, Word Sens Disambiguation methods (WSD) allow to select the correct sense of a term from a set of possible senses offered by external resources according to a given context [11], [2]. The WSD methods require Semantic Similarity (SS) measures between terms. Different SS measures have been well established and studied in the literature, we notice those edge counting based [18], and those information content based [22].

Another popular approach is hybrid approaches, expansion terms are extracted by applying different techniques (semantics and statistics) [23], [27].

In this paper, we propose a new automatic semantic query expansion approach. In fact, we aim to improve the performance of medical information retrieval when we are interested in the choice of the expansion terms. Our proposition is based on the exploitation of the MeSH structure and especially on the different semantic relations that exists between the different entities. We also consider the nature of these entities to the resource in question.

The rest of the paper is organized as follows. We review different research studies about the semantic query expansion in section 2 and then explain our proposed method in section 3. We present the test collections and measures used to evaluate our method in section 4 and we detail a series of experiments which were run in section 5. We finish this paper by conclusion and some future work in section 6.

II. Related Works

There are different studies dealing with the semantic query expansion for both general and specific domains mainly the medical field.

A. Semantic Query Expansion for a General Domain

The most often used resource in the general field is the WordNet, as a large lexical database. It groups terms into sets of synonyms called synsets. All synsets are interconnected by semantic relationships, such as, hypernym/hyponym (is-a/kind-of), meronym (has-part) and holonym (part-of) relationships.

(Gong et al., 2006) [13] proposed an approach for query expansion. In their approach, they expanded the simple query terms by all hypernyms, hyponyms and synonyms. They noticed that this method includes many noise words, thus, reduce the precision of the query results. Thus, they proposed to divide terms in the same query into groups using the $SS_{Jian-Conrath}$ measure to obtain multi-terms. Then, they expanded these multi-terms by hypernyms, hyponyms and synonyms. To evaluate their methods, Web pages are used to calculate the precision/recall value. It is concluded that multi-term query expansion method show significant increases in the number of correct retrieved documents compared to term query expansion.

(Audeh et al., 2013) [4] also proposed a new approach to incorporate additional new terms to a query whenever the query terms are disambiguated using the $SS_{Wu\& \ Palmer}$ measure. The authors added the synonyms and in case where the term doesn't have synonyms, the hypernyms are considered for possible additions to the query. The authors evaluated their approach with the INEX2009 collection and topics. In their experiments, they showed that the approach can enhance both the recall and the precision.

B. Semantic Query Expansion for a Specific Domain

As the general semantic expansion improves the query performance, the medical semantic expansion can also improve it. The MeSH thesaurus and the UMLS Metathesaurus are broad use in the medical field. MeSH uses Category/Descriptor/Concept/ Term or Category/ Descriptor/ Entry Terms structure. MeSH contains also other terms mentioned as Consider Also or See Related. There are three types of explicit relationships between the preferred concept and subordinate concepts: broader, narrower and related but not broader or narrower.

Figure 1 presents an example which illustrates these relationships. However, implicit relationships are defined between descriptors ("is-a", "part-of", "conceptual-part-of", and "process-of").

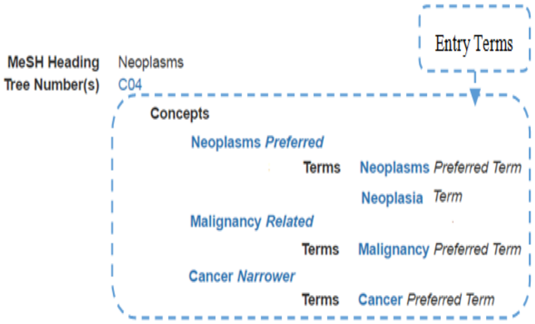


Fig. 1. An example for MeSH entities and relationships

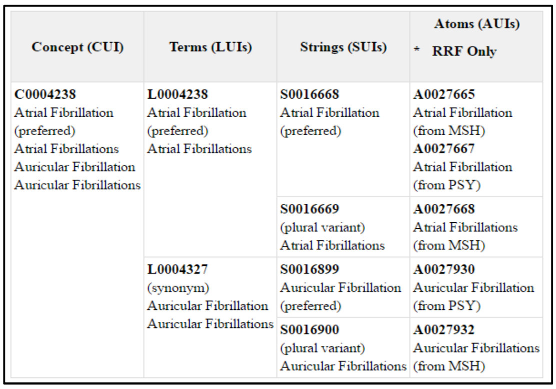
UMLS Metathesaurus uses Concept (CUI)/Terms (LUI)/String (SUI)/Atom (AUI) structure, as shown in the example below (Figure 2). There are different relationships between them such as “part-of”, “conceptual-part-of”, “process-of”, “surrounds”, etc. which the “is-a” relationship organizes all the concepts into a hierarchical structure as in MeSH.

Therefore, there are several research studies which dealt with the effect of introducing MeSH and UMLS for query expansion.

(Diem et al., 2007) [10] evaluated the effect of expanding queries and documents automatically based on semantic relationships in the UMLS hierarchy. They added more specific concepts to queries and more general concepts to documents while choosing the “is-a” relationship. To perform their study, they used the ImageCLEFMed2005 collection. They obtained the best enhancement of performance for documents, while query expansion by specific concepts gives less favorable result. They explained this by the fact that when adding more specific concepts to the query, they take the risk of changing the user’s information need and then the expanded query tend to lose its focus.

(Díaz-Galiano et al., 2007) [9] proposed a method to expand the queries with MeSH thesaurus. They expand the original query by the Entry terms. So, as they reduce the number of terms to add, they worked on some categories (A, C and E categories of MeSH). They expanded both the simple and compound words (giant cell: extending cell and giant cell). The results of their method on the ImageCLEFmed2007 collection gives a lower improvement in MAP (Mean Average Precision), because the expansion queries have many repeated words that don’t add new information, forever their weight increases in the query, which may not be desirable since the standard Lemur query Parser assumes that a very common word is very important in the query.

(Mata et al., 2011) [19] have developed two techniques based on the hierarchical MeSH structure for query expansion. The first technique added the SeeRelated Descritor or ConsiderAlso to the initial query. The second technique exploits various strategies. The first expansion strategy added the entry terms of descriptor. In the second strategy, the entry terms of the preferred concept is added to the initial query. The third expansion strategy added the descriptors. And to avoid noise, they apply a filtering method. However, it is proven that MeSH hierarchical structure could be used to achieve a slight improvement in both MAP and F-measure. The authors employed the collections used in the medical image retrieval task (ImageCLEF2009 and 2010).



Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

Fig. 2. An example for UMLs Metathesaurus

(Sharef et al., 2013) [20] also expanded the queries by using MeSH thesaurus. The proposed method extracts medical terms from the clinical records and expands the original queries with their synonymous terms. To verify the efficiency of their method, they used the ImageCLEF2010 collection. Experimental results suggest that the query expansion using synonymous terms can improve the MAP value.

(Dramé et al., 2014) [12] used MeSH thesaurus or UMLS to expand queries. Therefore, synonymous terms, descendants and related MeSH terms have been used to expand the queries. Moreover, to perform their study, they used the CLEFeHealth2014 collection. They noticed that the use of related terms decreases retrieval performance. On the contrary, query expansion using only synonymous terms from MeSH, improves the retrieval performance.

C. Discussion

Those different studies on the query expansion, led us to identify different types of query expansion [30]:

- The expansion by synonymy: consists in reformulating the query by adding synonyms, which guarantees that reformulation does not change the semantic of the initial query [7].
- The expansion by specialization: consists in adding terms to the query that specialize the original ones (that is, their descendants). Therefore, this method can contributes to the increase of the number of relevant documents to the returned ones and subsequently to an increase of precision [12] [8].
- The expansion by generalization: consists in adding each term of the original query with more general concepts present in the semantic resources (that is, its ascendants), which generates more general answers and far than what is desired by the user [29]. This type of expansion is used for the document expansion in most approaches [10].

In Table I, we try to classify studied works by considering the choice of expansion terms for the medical field, the ontological resource and the expansion type directly related to the choice of expansion terms. We notice that different studies on the query expansion did not focus on the medical terms nature (Terms, Concepts and Descriptors) and they use these different types of query expansion, as we saw above, in an independent way.

TABLE I. CLASSIFICATION OF SEMANTIC QUERY EXPANSION WORKS

Works	Ontological resources	Expansion terms	Expansion Types
Diem et al., 2007	-UMLS Metathesaurus	- Descendants concepts	-Expansion by specialization
Stokes et al., 2009	-MeSH thesaurus -SNOMED-CT -UMLS Metathesaurus	-Synonyms -Descendants concepts -Ascendants concepts	-Expansion by synonymy -Expansion by specialization -Expansion by generalization
Crespo et al., 2011 Crespo et al., 2012	-MeSH thesaurus -SNOMED-CT	-Descendants descriptors -Entry terms -Entry terms of the preferred concept	-Expansion by synonymy -Expansion by specialization
Sharef et al., 2013	-MeSH thesaurus	- Synonyms	-Expansion by synonymy
Dramé et al., 2014	-MeSH thesaurus -UMLS Metathesaurus	-Synonyms -Descendants descriptors -Relatedterms	-Expansion by synonymy -Expansion by specialization

Our goal is to take the MeSH resource into consideration the relationships between these entities and their natures and to combine these different expansion types in order to propose an approach for query expansion which focuses on the choice of the appropriate expansion terms.

III. THE PROPOSED APPROACH

In the medical field, medical terms are not necessarily used in the original query although, they are crucial for information retrieval. To solve this problem, we apply a semantic query expansion method. In this context, we do not aim to propose methods for query expansion because many techniques already exist since the 90's. The goal of this paper is how to choose the expansion terms based on the characteristics of the hierarchical structure of MeSH thesaurus. We fix this resource as knowledge semantic resource since its broad use in the medical field.

A. MeSH Structure

The Medical Subject Heading (MeSH) is created by the National Library of Medicine in 1954, and it is a controlled vocabulary used to index Medline papers. MeSH is composed of thirteen hierarchical structures having various levels of specificity. The highest level consists of 16 Categories (**S_{Cat}**): Category A for anatomic terms, Category B for organisms, C for diseases, D for drugs and chemicals, etc. Each Category contains a set of Descriptors **(S_D)**, Concepts **(S_C)**, and Terms **(S_T)**. The Descriptors are identified by a number that indicates their tree location, for example, "Pain" has these trees numbers C10.597.617.140, C23.888.592.612.107. Each Descriptor consists in one or more Concepts, in which the Descriptor name is that of the Preferred Concept. Each Conceptof the set **(S_C)** consists in one or more synonymous Terms and has a Preferred Term, which is also said to be the name of the Concept. In contrast, the Terms in one Concept are not strictly synonymous with the Terms in another Concept.

We distinguish between these sets **S_T**, **S_C**, **S_D** by the definitions and properties as below:

We define **S_T** as:

$$S_T = S_{NPT} \cup S_{PT} \qquad (1)$$

With:

- **S_{NPT}**: Corresponds to the set of terms which are not considered as concepts or descriptors (Non Preferred Terms).
- **S_{PT}**: Corresponds to the set of Preferred Terms.

We define **S_{PT}** as:

$$S_{PT} = S_C = S_{NPC} \cup S_{PC} \qquad (2)$$

With:

- **S_{NPC}**: Corresponds to the set of concepts which are not considered as descriptors. They are subordinate concepts (Non Preferred Concepts).
- **S_{PC}**: Corresponds to the set of descriptors (Preferred Concepts), we assume that:

$$S_{PC} = S_D \qquad (3)$$

We define some properties as follows:

$$S_{NPT} \cap S_C = \emptyset \qquad (4)$$

$$S_{NPC} \cap S_D = \emptyset \qquad (5)$$

We represent the different relationships between elements of these sets **S_{NPT}, S_{NPC}, S_D,** and **S_{Cat}** in Figure 3 below.

B. MeSH-Based Query Expansion Method

Our approach for semantic query expansion is detailed in this section. It includes different steps shown below in Figure 4.

The first step denoted annotation, consists in indexing (1.1) the initial query (q_i) able to extract a set of significant keywords from the query, which uses different treatments: removing stopwords from the query and stemming query terms. Afterwards, we propose to divide the set of keywords

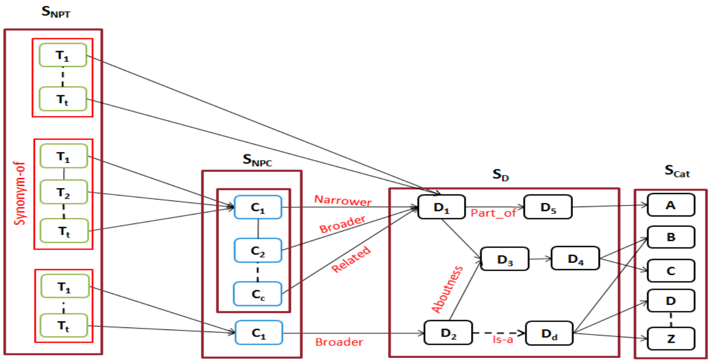


Fig. 3. Representative scheme of relationships between MeSH entities

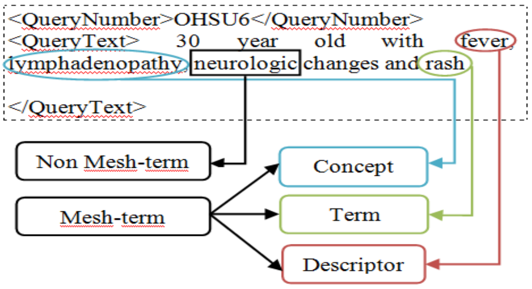
according to their presence in the thesaurus MeSH into two groups: Mesh terms, and non-Mesh terms.

The sub-step (1.2) consists in detecting the entity MeSH nature of the MeSH terms. According to this, we obtain three sets: terms $(\{t_i\}\backslash i=1..t)$, concepts $(\{c_i\}\backslash i=1..c)$ and descriptors $(\{d_i\}\backslash i=1..d)$. Or we get both the simple and compound terms. We choose to consider in our method compound terms. For example, "Pain" and "Back" are descriptors, and also "Back pain" is a descriptor.

In the second step, we are interested in recognizing the semantic relationships between each term in the MeSH terms list and the terms in MeSH thesaurus. Thus, we propose to proceed according to the entity MeSH nature established using the following criteria to expand the Mesh terms:

- Treatment by synonymy (2.1): If the Mesh-term is a term t_i ($t_i \in \mathbf{S_{NPT}}$), it is expanded by the corresponding preferred term.
- Treatment by generalization (2.2): If the Mesh-term is a concept c_i ($c_i \in \mathbf{S_{NPC}}$), it is expanded by at least the corresponding preferred concept (its directly ascendant in the MeSH structure hierarchy).
- Treatment by generalization (2.3): If the Mesh-term is a descriptor d_i ($d_i \in \mathbf{S_D}$) and it is attached to more than one parent, a WSD step is necessary to introduce in order to choose the appropriate one to consider.

We begin by developing the first treatment by synonymy (2.1) and the second treatment by generalization (2.2). In what follows, we propose as illustration an example. Figure 5 present the step of annotation.



<QueryNumber>OHSU6</QueryNumber>
<QueryText> 30 year old with fever lymphadenopathy neurologic changes and rash
</QueryText>
Non Mesh-term
Mesh-term
Concept
Term
Descriptor

Fig. 5. Annotation method applied to a biomedical query

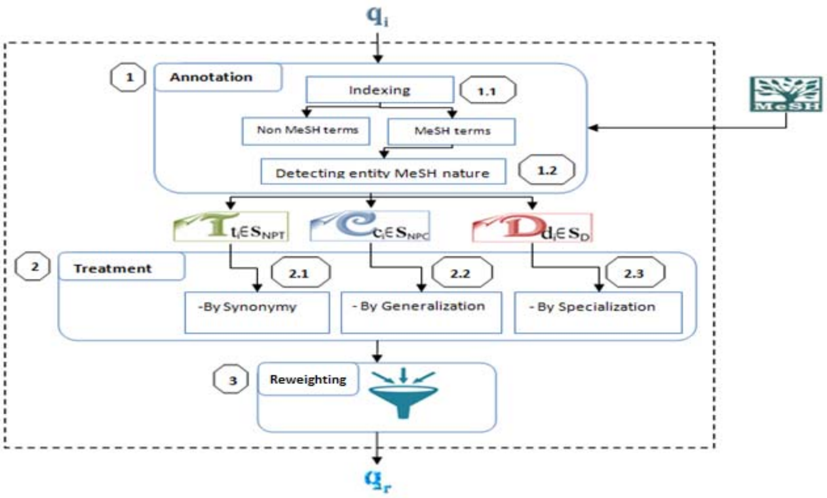
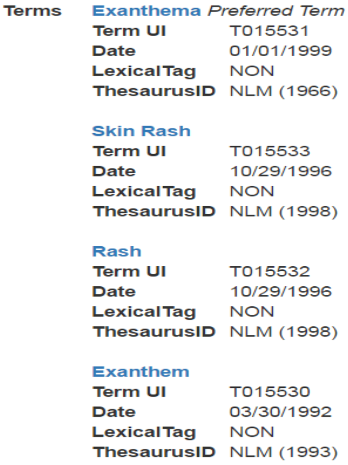


Fig. 4. Architecture of our proposed approach

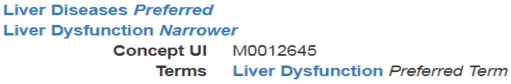
Afterwards, the set of identified terms are expanded by their preferred term. In Figure 6, we take the term "rash" as an illustrative example.



Terms	Exanthema *Preferred Term*	
	Term UI	T015531
	Date	01/01/1999
	LexicalTag	NON
	ThesaurusID	NLM (1966)
	Skin Rash	
	Term UI	T015533
	Date	10/29/1996
	LexicalTag	NON
	ThesaurusID	NLM (1998)
	Rash	
	Term UI	T015532
	Date	10/29/1996
	LexicalTag	NON
	ThesaurusID	NLM (1998)
	Exanthem	
	Term UI	T015530
	Date	03/30/1992
	LexicalTag	NON
	ThesaurusID	NLM (1993)

Fig. 6. Excerpt from MeSH

In the second treatment, the set of identified concepts are expanded by the corresponding preferred concept in the case where does not contain synonymous terms (Figure 7).



Liver Diseases *Preferred*
Liver Dysfunction *Narrower*
Concept UI M0012645
Terms Liver Dysfunction *Preferred Term*

Fig. 7. Example of concept expanded by the preferred concept

IV. Evaluation

In our experimental evaluation, we want to look at the effects of choosing the expansion terms on the retrieval performance. In what follows, we will describe the collection test and the metrics used.

A. Collection Description

The document collections used for the experimental study are the OHSUMED and the Cystic Fibrosis collections.

The OHSUMED collection was obtained by (Hersh et al., 1994) [15]. It consists of the titles and/or abstracts from 270 medical journals published between 1987 and 1991. This collection is subdivided into three sub-directories. The directory that we used contains 36890 documents published from 1987, which contains the "title" and "abstract" fields. Moreover, 63 topics consisting of the "title" field indicate the patient's description and the "description" field announces information request. Each topic is provided with a set of relevant documents judged by a group of physicians in a clinical setting.

The Cystic Fibrosis collection consists of 1239 documents published between 1974 and 1979, discusses Cystic Fibrosis Aspects, which contains the "title" and "abstract" fields. Moreover, there are 100 topics which consist of the "title" field. Each topic includes a set of relevant documents judged by the experts.

B. Evaluation Measures

To measure the IR performance, we use the TREC_EVAL software developed by the TREC (Test REtrieval Conference). This already mentioned software calculates the most widely used measures to assess retrieval efficacy. The P@5, P@10 and MAP are computed. Where P@5, P@10 represent respectively the mean precisions at the top 5, 10 returned documents, and then the P@n for a query is defined as:

$$P@n = \frac{\#relevant\ documents\ in\ top\ n\ results}{n} \qquad (6)$$

Mean Average Precision (MAP) for a set of queries Q is defined as follows:

$$\mathrm{MAP} = \frac{1}{|Q|}\sum \frac{\sum_{n=1}^{N}(P@n*rel(n))}{\#\text{total relevant documents for this query}} \qquad (7)$$

Where:

|Q| is a number of test queries.

rel(n) is a binary function on the relevance of the n-th document:

$$\mathrm{rel(n)} = \begin{cases} 1, \text{ if the n-th document is relevant} \\ 0, \text{ otherwise.} \end{cases}$$

V. Experiments and Results

This section details the experiments made to carry out various runs for each treatment in order to compare their effect on retrieval performance.

In these experiments, documents and queries were indexed using the Terrier platform. It is an efficient, effective and flexible open source search engine, easily deployable on large-scale collections of documents. Terrier is written in Java, and is developed at the School of Computing Science, University of Glasgow.

The indexation process includes stemming algorithm (Porter's stemmer) and stopword (list of SMART system). To rank documents according to their relevance to a given query, we chose the weighting scheme OKAPI BM25, used as the baseline (BL), denoted BM25.

Therefore, there are different runs presented as follows:

1. **BL**: Original queries.
2. **SynonymousTermsMeSH**: Identified query terms expanded with synonymous terms based on MeSH.
3. **PreferredTermMeSH**: Identified query terms expanded with corresponding preferred term based on MeSH.
4. **PreferredConceptMeSH**: Identified query concepts expanded with corresponding preferred concept.
5. The last run combining **PreferredTermMeSH** and **PreferredConceptMeSH** runs.

The tables below (table 2 and 3) show the results obtained with each run. For each query, the first 100 documents for the "Cystic Fibrosis" collection and the first 1,000 documents are considered for the "OHSUMED" collection.

We proceed to perform our runs for each collection. Concerning "Cystic Fibrosis" collection (table 2), in run 5, we get the best performance according to MAP, P@5 and P@10 when we combining both preferred term and preferred concept for each term identified and concept identified respectively.

However, adding preferred term for each term identified, in run 3, enhance the results compared with the added of synonymous terms (run 2). So, we see that the addition of a large number of terms decreases the performance.

Concerning "Ohsumed" collection (table 3), (Dinh and Tamine, 2015) [11] tested the query expansion into two ways representing the query. One way by considering only titles. The second way by combining titles and descriptions. They obtained the best results for query expansion when combining titles and descriptions.

The table 3 presents the IR performance obtained by these different runs for queries of the "OHSUMED" collection. In run 2 which uses the synonymous terms that are added for query, we observe a decrement in performance. The cause of the decrement is due to the problem of the large number of added MeSH-terms that are not necessary useful for retrieval. In run 3, we notice that the performance results were improved when adding the preferred term to the initial query in MAP. In run 4, we get the best performance in P@10 when we add preferred concept to the initial query. However, when we combine both preferred term and preferred concept, in run 5, we get a higher improvement in P@5.

In both collections, for some queries, they gave good and even the best results. For example, in "Ohsumed" collection, our Run 5 got the best performance for the query OHSU31 (P@10=0.4) which is even higher than the baseline result (P@10=0.1). This query contains only a concept "Coumadin". So it is extended by preferred concept "Warfarin", sold under the brand name Coumadin. Then, documents containing the more general term "Warfarin", which are considered as relevant, are ranked at the top of the list. On the other hand, for other queries, our method got poorer performance. For example, for the query OHSU13 (P@10=0.4) which is even lower than the baseline result (P@10=0.6).

VI. CONCLUSIONS AND FUTUR WORKS

In this paper, we have proposed a semantic query expansion method in order to retrieve more biomedical relevant documents. Our approach introduces three methods according to the entity MeSH-term nature. For that, we have studied the basic elements of its poly-hierarchical structure (categories, descriptors, concepts, terms and relations). From this study, we have identified different treatments. At this stage we have focused on the evaluation of the first and second treatments (denoted in figure 3 as 2.1 and 2.2). The preliminary results obtained showed the effectiveness of our proposed method in the IR process.

In future work, we plan to evaluate the treatment by generalization (denoted in figure 3 as 2.3) to enhance the IR performance. We plan to introduce a method to disambiguate ambiguous Mesh entities. We will also explore the new techniques to assign a weight to the query terms.

References

[1] M.E.A. Abderrahim, S. Benameur, and M.A. Abderrahim, "The number of terms and documents for pseudo-relevant feedback for ad-hoc information retrieval," International Journal of Computer Science Issues, IJCSI, vol. 10, no 1, pp. 661-667, 2013.

[2] E. Agirre, O. López de Lacalle, and A. Soroa, "Random walks for knowledge-based word sense disambiguation," Computational Linguistics , vol. 40, no 1, pp. 57-84, 2014.

[3] H. Aliane, Z. Alimazighi, R.O. Boughacha, and T. Djeliout, "Un Système de reformulation de requêtes pour la recherche d'information, " La revue de l'Information Scientifique et Technique (RIST), vol. 14, no 1, 2004.

[4] B. Audeh, P. Beaune, and M. Beigbeder, "Expansion sémantique des requêtes pour un modèle de recherche d'information par proximité," In INFORSID, pp. 83–90, 2013.

[5] C. Carpineto, and G. Romano, "A survey of automatic query expansion in information retrieval," ACM Computing Surveys (CSUR), vol. 44, no 1, pp. 1, 2012.

[6] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," ACM Transactions on Information Systems (TOIS), vol. 19, no 1, pp. 1-27, 2001.

TABLE II. RESULTS OF QUERY TITLE EXPANSION FOR "CYSTIC FIBROSIS"

Run		MAP@100	P@5	P@10
1	BM25	0.1466	0.4220	0.3520
2	SynonymousMeSH	0.1500	0.4240	0.3530
3	PreferredTermMeSH	0.1486	0.4260	0.3530
4	PreferredConceptMeSH	0.1495	0.4300	0.3590
5	PreferredTermMeSH-PreferredConceptMeSH	**0.1512**	**0.4340**	**0.3600**

TABLE III. RESULTS OF QUERY TITLE AND DESCRIPTION EXPANSION FOR "OHSUMED" COLLECTION

Run		MAP@1000	P@5	P@10
1	BM25	0.2933	0.3968	0.3016
2	SynonymousMeSH	0.2792	0.3556	0.2810
3	PreferredTermMeSH	**0.3041**	0.4000	0.3032
4	PreferredConceptMeSH	0.3014	0.4032	**0.3159**
5	PreferredTermMeSH-PreferredConceptMeSH	0.3018	**0.4095**	0.3095

[7] J. Choi, Y. Park, and M. Yi, "A hybrid method for retrieving medical documents with query expansion," In 2016 International Conference on Big Data and Smart Computing (BigComp), pp. 411–414.IEEE, 2016

[8] M. Crespo, J.M. Vázquez, and M.J.M. López, "LABERINTO at ImageCLEF 2012 Medical Image Retrieval Task," In CLEF (Online Working Notes/Labs/Workshop), 2012.

[9] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Urena-López, "Integrating mesh ontology to improve medical information retrieval," In Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 601–606, 2007, Springer, Berlin, Heidelberg.

[10] L.T.H. Diem, , J.P. Chevallet, and D.T.B. Thuy, "Thesaurus-based query and document expansion in conceptual indexing with umls," RIVF'07, 2007.

[11] D. Dinh, and L. Tamine, "Identification of concept domains and its application in biomedical information retrieval," Information Systems and e-Business Management, vol. 13, pp. 647-672, 2015.

[12] K. Dramé, F. Mougin, and G. Diallo, "Query Expansion using External Resources for Improving Information Retrieval in the Biomedical Domain," In CLEF (Working Notes), pp. 189–194, 2014.

[13] Z. Gong, and C.W. Cheang, "Multi-term web query expansion using WordNet," In International Conference on Database and Expert Systems Applications, pp. 379–388, 2006, Springer, Berlin, Heidelberg.

[14] N. Hernandez, and J. Mothe, "Ontologies pour l'aide à l'exploration d'une collection de documents," Ingénierie des Systèmes d'Information, vol. 10, no 1, pp. 11-31, 2005.

[15] W. Hersh, C. Buckley, T.J.Leone, and D. Hickam, "OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research," In SIGIR'94 , pp. 192–201, 1994, Springer, London.

[16] J. Hu, W. Deng, and J. Guo, "Improving retrieval performance by global analysis," In Pattern Recognition, ICPR 2006, 18th International Conference on, Vol. 2, pp. 703-706, 2006, IEEE.

[17] N. Ksentini, M. Tmar, and F.Gargouri, "Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms," In Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on, pp. 134-139, 2015, IEEE.

[18] Y. Li, Z.A. Bandar, and D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources," IEEE Transactions on knowledge and data engineering, vol. 15, pp. 871-882, 2003.

[19] J. Mata, M. Crespo, , and M.J. Maña, "Using MeSH to expand queries in medical image retrieval," In MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support, pp. 36–46, 2011, Springer, Berlin, Heidelberg.

[20] N.M. Sharef, and H. Madzin, "Semantic-based medical records retrieval via medical-context aware query expansion and ranking," Journal of Theoretical & Applied Information Technology, vol. 58, no 3, 2013.

[21] H. S. Oh, and Y. Jung, "Cluster-based query expansion using external collections in medical information retrieval," Journal of biomedical informatics, vol. 58, pp. 70-79, 2015.

[22] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," J. Artif. Intell. Res. (JAIR), vol. 11, pp. 95-130, 2011.

[23] A.R. Rivas, E.L. Iglesias, and L. Borrajo, "Study of query expansion techniques and their application in the biomedical information retrieval," The Scientific World Journal 2014, 2014.

[24] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," NIST SPECIAL PUBLICATION SP, pp.73-96, 1996.

[25] J.J. Rocchio, "Relevance feedback in information retrieval," The SMART retrieval system: experiments in automatic document processing, pp. 313-323, 1971.

[26] G. Salton, , and C. Buckley, "Improving retrieval performance by relevance feedback," Readings in information retrieval, vol. 41, no 4, pp.355-363, 1997.

[27] J. Singh, and A. Sharan, "Relevance Feedback Based Query Expansion Model Using Borda Count and Semantic Similarity Approach," Computational intelligence and neuroscience, 2015, pp. 96, 2015.

[28] J. Singh, and A. Sharan, "A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach," Neural Computing and Applications, pp. 1–24, 2016.

[29] N. Stokes, Y. Li, L. Cavedon, and J. Zobel, "Exploring criteria for successful query expansion in the genomic domain," Information retrieval, vol. 12, no 1, pp. 17-50, 2009.

[30] C. Yunzhi, L. Huijuan, L. Shapiro, R.S. Travillian, and L. Lanjuan, "An approach to semantic query expansion system based on Hepatitis ontology," Journal of Biological Research-Thessaloniki, vol. 23, no 1, pp. 11, 2016.