Using Tag-Neighbors for Query Expansion in Medical Information Retrieval

Frederico Durao, Karunakar Bayyapu, Guandong Xu, Peter Dolog, Ricardo Lage
Department of Computer Science
Aalborg University
Selma Lagerlöfs Vej 300
Email: fred,kreddy,xu,dolog,ricardol@cs.aau.dk

***Abstract*—In the context of medical document retrieval, users often under-specified queries lead to undesired search results that suffer from not containing the information they seek, inadequate domain knowledge matches and unreliable sources. To overcome the limitations of under-specified queries, we utilize tags to enhance information retrieval capabilities by expanding users' original queries with context-relevant information. We compute a set of significant tag neighbor candidates based on the neighbor frequency and weight, and utilize the most frequent and weighted neighbors to expand an entry query that has terms matching tags. The proposed approach is evaluated using MedWorm medical article collection and standard evaluation methods from the text retrieval conference (TREC). We compared the baseline of 0.353 for Mean Average Precision (MAP), reaching a MAP 0.491 (+39%) with the query expansion. In-depth analysis shows how this strategy is beneficial when compared with different ranks of the retrieval results.**

I. Introduction

In the context of medical document retrieval, users often under-specified queries lead to undesired search results that suffer from not containing the information they seek, inadequate domain knowledge matches and unreliable sources. For instance, when a user wants to search for a recent outbreak of *influenza* on the web, a search with the query *influenza* will return a list of documents containing the query term, ranked by a set of criteria defined by the search engine. In this case, at least three issues may affect the quality of the search result. One, a query with only one or two terms may be under-specified, that is, it may not contain enough terms for the search engine to retrieve the desired information to the user. Second, in the document repository of the search engine, there might exist more than hundreds of thousands articles matching the requested query. In such an amount of information, it is impossible to locate the desired information by simply browsing through all contents of returned results. The third reason is related to domain knowledge requirements. Because conventional search engines focus on generic information search, domain specific results are usually not taken into consideration during the search. Thus, a simple word based search does not produce relevant search results in specific domains such as the medical domain [1]. As a consequence of these issues related to query-based searches, only one fourth to one half of the relevant articles on a given topic are retrieved in searches performed in specific domains [2]. In other words, the sparse and incomplete query terms may result in information overload increasing the noise present in search results. Hence, the importance of refining a query is increased in such scenarios.

To overcome the limitations of under-specified queries, we utilize **tag** *neighbors* to enhance information retrieval capabilities by expanding the user's original query. Tags are free style terms to make annotations indicating the user's own perceptions or conceptual judgments about the tagged resources. We focus on medical document collections, e.g. *PubMed*[1]and *MedWorm*[2], because in searching these collections it is often desirable to retrieve only those documents pertaining to a specific medical area. To this end, tags given by the users to the documents in the collection are typically related to the domain(s) each user is interested in. That is, users are able to choose their own free style terms (i.e. *tags*) which are associated to the domain(s) of their interest.

The purpose of query expansion is to fill the gap between the users entered queries and extracting the relevant documents. In a nutshell, we compute a set of significant tag neighbor candidates based on the tag neighbor frequency and weight and utilize the most frequent and weighted tag neighbors to expand an entry query that has terms matching tags. For instance, if a user submits a query *influenza*, the query will be automatically mapped to the higher frequency tag neighbor term *contagious* by our method. Thus, the search will be refined by retrieving documents having the words *influenza* and *contagious* in their contents. Furthermore, neighbor terms also searchable. Take the previous query, for example, documents indexed with medical terms that include the word *influenza (e.g. influenza contagious viral)* will also be returned depending on the neighbor frequency and weight.

In this paper, the expansion terms we used are selected from a large amount of tags provided by the users. Then we propose to use the tag neighbors method for a high frequency term selection. Based on this method we tried to choose good expansion terms from the candidate neighbors, according to their potential impact on retrieval effectiveness. We implement our method in a search system with contents extracted and indexed from the *MedWorm* medical article database. We

[1] www.ncbi.nlm.nih.gov/pubmed

[2] www.medworm.com

carry out experiments with *MeSH*[3](Medical Subject Headings) vocabulary search queries to evaluate the performance of our method in the developed system. The experimental results show that the retrieval effectiveness can be improved with the Mean Average Precision (MAP) by +39% over traditional information retrieval. The main contributions of this paper are summarized as follows:

- An algorithm for tag neighborhood generation, that searches for the neighboring words of all tags occurring in the document corpus and computes the neighbor's frequency and weight;
- Tag neighborhood selection for the expansion of a given query based on the neighbor frequency and weight. The expanded query seeks to obtain documents that not only refer to the tags but also to related concepts based on their neighbors;

The rest of the paper is structured as follows. Section II reviews the related research undertaken in this area. Section III presents our approach of expanding query for document search with tag neighbors. Section IV describes the experimental setup, evaluation methodology, metrics and results. Section V discusses the results of the query expansions carried out into the evaluation. Section VI concludes the paper and outlines future works.

II. Related Work

A. *Query expansion in the medical domain*

Query expansion requires a term selection stage where the system presents the query expansion terms to the users in some reasonable order [3]. The order should preferably be one in which the terms that are most likely to be useful are close to the top of the list. In addition, heuristic decisions can also be applied during this stage, for example, poor terms are excluded from the term list instead of being given low weights. [4] proposed an information-theoretic approach to automatic query expansion, which is based on Information Theory, assigning scores to various candidate expansion terms. These scores are used to select and weight expansion terms within Rocchio's framework for query re-weighting. This approach was compared with other query expansion techniques via empirical studies. They claimed that their approach was able to achieve better retrieval effectiveness on several performance measures. Similarly to our model, [4] weight candidate terms for query expansion with the goal of achieving better retrieval effectiveness. On the other hand they do explore tags as means of providing additional semantics to the query expansion.

[5] investigated the effectiveness of using MeSH in PubMed through its automatic query expansion process: Automatic Term Mapping (ATM) and concluded that retrieval performance was improved but the improvement may not affect to end PubMed users in realistic situations. Although our approach has applied a different technique, our evaluation shows that we also achieved improvements on the performance with the query expansion. Likewise [5], we outline some drawbacks of our approach. Specifically, we indentify which medical categories our information retrieval does not perform properly and achieve low precision rates.

A knowledge-based query expansion method [6] exploits UMLS (Unified Medical Language System), a large thesaurus in the biomedical domain constructed by Library National of Medicine knowledge source. They goal is to append the original query with additional terms that are specifically relevant to the query's scenario. The exploration of UMLS knowledge is done by mapping a large text of collection ImageCLEFMed(CLEF-Cross Language Evaluation Forum) to UMLS concepts, and expanding queries and documents automatically base on semantic relations in the UMLS hierarchy [7]. The exploitation of semantic relations from a knowledge based is what most differ their work from ours. In our work we do not rely on any existing database to perform the query expansion. Instead, we harvest the implicit semantic relations inherit within the nearest candidate terms.

[8] presented a method to expand queries with a medical ontology in order to improve an IR system. The aim is to improve a multimodal retrieval system by expanding the user's query with MeSH descriptors. They have combined two independent subsystems to retrieve textual and visual information. The evaluation of this system is carried out using the collection queries, and relevance judgments provided by the ImageCLEF medical task organization. Moreover, they compared the use of a traditional Information Retrieval (IR) system, an IR system with medical knowledge, a Content-Based Information Retrieval(CBIR) system and a mixed system with information from these systems. Finally, the results show that the use of medical ontology to expand the queries greatly improves the system. Similarly to [6], [8] also relies on an ontology (as a knowledge database) to perform the query expansion. Unlike our approach they benefit from a second resource, which is the visual information. This information enriches the semantics of the search queries with taxonomic concepts.

Concept-based query expansion for retrieving gene related publications from MEDLINE was investigated by [9]. The approach is based on exploiting the direct links between genes and other biological concepts obtained from public biological databases. These networks of associations are implemented through direct relations in the integration database. Images are also more important and varied in the medical domain, as they become available in a digital form. Despite the fact that images are language-independent, they are often accompanied by textual features (associated captions, titles and articles) strongly improve the retrieval quality. Link relations are not considered in our approach. However we see a potential study of tagging activity between documents. This analysis could bias the weighing the search score.

In [5], [6], [9] authors used approximately the same data source for query expansion on different knowlsedge domains. However, to the best of our knowledge, this is the first attempt to use *tags* for query expansion to enhance information retrieval performance in the medical domain.

[3] www.nlm.nih.gov/mesh

B. Tagging and Folksonomy

A *tag* is a one-unit word or label that describes a piece of information. In social tagging, in contrast to taxonomies developed by subject specialists using authorized terms (determined by professionals), people use their own keywords to describe websites for future discovery and retrieval [10]. The resulting list of tags of information and objects is often termed a "folksonomy," a classification done by untrained individuals (folks) [11]. The term "folksonomy" is a combination of folk and taxonomy to describe the social classification phenomenon. Folksonomy provides user-created metadata rather than the professional created an author created metadata [12]. The tags are the core of folksonomy can be seen as good keywords for describing the respective web pages from various aspects of medical ontology [13].

Tags in Medical bookmarking systems are usually assigned to organize and share resources on the Web. By tagging, users label resources freely and subjectively, based on their sense of values [14]. [15] method shown that an effective tagger for medical terms related to diseases, injuries, drugs, medical devices, and medical procedures can be built using words from a robust medical term list along with a probabilistic term classifier that uses local context to disambiguate terms being used in a medical sense from terms being used in a non-medical sense. [16] proposed a semantic tagger that provides high level concept information for phrases in clinical documents, which enriches the medical information tracking system that supports decision making or quality assurance of medical treatment. Tagging systems in the medical field have focused on the lexical level of syntactic and semantic tagging. [17] and [18] performed semantic tagging on terms lexically using the Unified Medical Language System (UMLS).

In the related works mentioned above, different kinds of expanded query information and tagging activity were considered in order to enhance the retrieval performance on different systems and collections in the medical domain, showing the potential application of query expansion. However, the gap between the users required information and extracting relevant document information is not yet clarified efficiently in the medical domain.

III. Tag Neighbor based Query expansion

In this section, we discuss the query expansion approach. We first overview the query search system and then explain the generation of tag neighbors and the query expansion procedure. In order to test our approach, we develop a keyword-based search system, which supports search by user entry queries. The framework of the proposed approach is depicted in Figure 1. Steps (1) and (2) are pre-processing phase that concerns with data extraction, database storage and neighborhood forming. On the Step (3), the user query is entered, and that is expanded in the Step (4). The last step, (5) presents the results to the user.

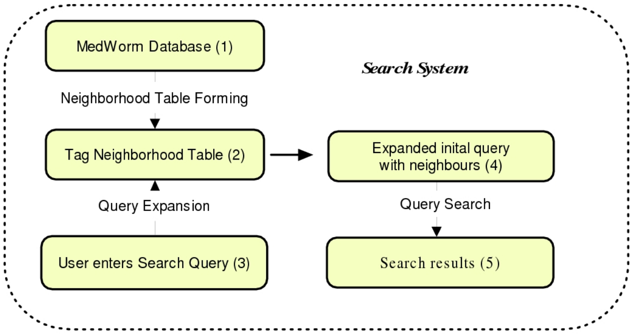
MedWorm Database (1)
Search System
Neighborhood Table Forming
Tag Neighborhood Table (2)
Expanded inital query with neighbours (4)
Query Expansion
Query Search
User enters Search Query (3)
Search results (5)

Fig. 1. The framework of expanded query search system with tag neighbor expansion

A. Generating Tag Neighbors

The basic idea proposed in this paper is to utilize the neighbors appearing before or after the query terms that match one or more tags within the document corpus to expand it with related terms to the original query. The rationale behind the approach is the observation that such kinds of tags appear in specific blocks or in pairs within the content of a document. In the context of medical documents, this phenomenon sometimes is dominantly observed. For example, the words "pandemic, H1N1, influenza, contagious, viral" are often used concurrently. In such a scenario *expanding the initial query with the related tag neighbors is expected to facilitate the retrieval of more closely related documents.*

To realize the task of expanding the initial query terms provided by a user during a search, we first define and retrieve tag neighbors from the document corpus. This is a pre-processing algorithm as described below:

1) Let T be the set of all tags assigned to the whole corpus. For each tag $t \in T$, we search for documents using t as an entry query.
2) For each retrieved document, we fetch the n terms before and after each occurrence of the tag t in the document.
3) To assure minimal quality of the neighbor candidates, we analyze the fetched terms, remove possible invalid characters and constraint those present in the list of stop words. The qualified terms are now represented by C, i.e. the set of candidate neighbors for the tag $t \in T$.
4) For each term $c \in C$, we calculate its weight based on its distance from the tag $t \in T$. We assume 1 unit the distance from t to the immediate term c after or before it. The distance to the second term is 2 units and so forth until the $|n|th$ unit. We assume that the furthest away the term c is from tag t, the less relevant it is. The function $w(t, c)$ that weighs the distance between a tag $t \in T$ and a candidate neighbor c is defined as:

$$w(t,c) = \sum_{i=1}^{|c|} \frac{1}{d(t,c_i)}, \forall c \in C \tag{1}$$

where $d(t, c_i)$ is the distance between the occurrence i of candidate neighbor c and the tag $t \in T$, and $|c|$ is

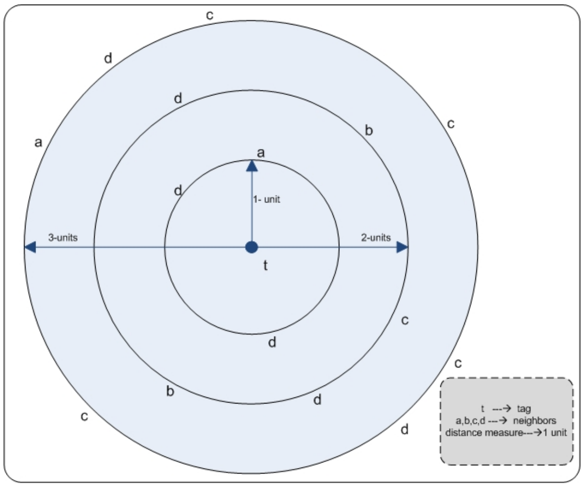

Fig. 2. Example of a concept view of a tag and neighbors distance relationship graph

T, C	w(t,c)	f(c)	nf(t,c)
t, a	0.665	0.13	0.086
t, b	0.5	0.13	0.065
t, c	1.83	0.3	0.549
t, d	3.66	0.4	1.464

TABLE I
TAG NEIGHBORS WEIGHT AND FREQUENCY FROM FIGURE 2

the total number of occurrences of candidate neighbor c near tag t.

5) For each candidate neighbor $c \in C$, we calculate the neighborhood frequency for its respective tag $t \in T$. i.e.

$$f(c) = \frac{|c|}{|C|} \tag{2}$$

where $|C|$ is the total number of candidate neighbors for the particular tag t.

6) The neighborhood frequency function $nf(t, c)$, that takes into account the weights and frequency is defined as follows:

$$nf(t, c) = w(t, c).f(c) \tag{3}$$

The calculus of the tag neighbors runs on the dataset as a data pre-processing procedure. After computing the most frequent and highest weight tag neighbors, we select those whose final score is above a threshold α to ensure a minimal quality during the expansion. The threshold α is defined by least nf(t,c) within the highest standard deviation from the best ranked neighbor. This avoids the need for defining an absolute number of neighbors to be retrieved and guarantees low frequent neighbors are selected. The set of tag neighbors will be finally saved in a (hash) tag neighborhood table. In this table, each tag points to the its respective set of neighbors.

1) Working Example on Generating Tag Neighbors: In Section III-A, we have described the method to compute the tag neighbor frequency and weight. Based on this, we select which neighbors have higher frequency and weight for the original query. Formally, let t be the tag and $C = \{c_0, c_1, \ldots, c_{|C|}\}$ the candidate neighbors set. Then, every $c_i (1 \leq i \leq |C|)$ can be instituted as a potential expansion for tag t. According to Figure 2, t is the tag and $\{a, b, c, d\} \in C$ are the candidate neighbors. The total number of candidate neighbors is 15 (i.e. $|C|$ = 15). The tag a occurs twice, b occurs twice, c occurs five times, and d occurs six times. According to these occurrences, we compute the weight function $w(t, c)$ by considering the distance for each candidate neighbor from the original tag. Particularly, in the case of the candidate neighbor a, two distances are evidence, the fist 1 unit and the second 3 units. Thus $w(t, c) = 0.665$ according to the equation (1). The frequency of each candidate neighbor is the division of each candidate neighbor by the total number of candidate neighbor with respect to tag t, i.e. $\frac{|c|_{each_a}}{|C|} = 2/15 = 0.13$. Finally the neighborhood function that multiplies the frequencies by the weights is $nf(t, a) = w(t, a) * f(a) = 0.086$. This procedure repeats all over the remain candidate tag neighbors. Table I shows the results of all candidates for each function as explained in the previous section.

The selection of most qualified neighbors is based on the $nf(t, c)$ value analysis. For instance, (see Figure 2), the candidate neighbor d is more likely to be selected rather than the candidate a, since $nf(t, d) > nf(t, a)$. The selection ordering of tag neighbors follows: $\{d\ ,\ c\ ,\ a\ and\ b\}$.

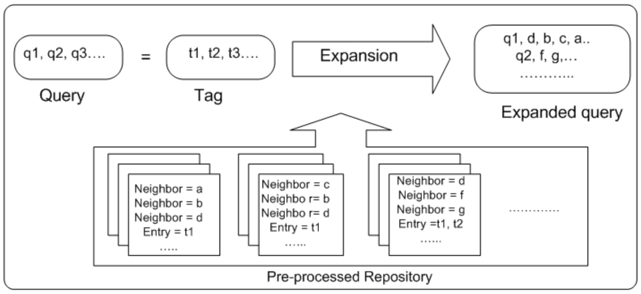

Fig. 3. Query expansion process with tag neighbors

B. Expanding Query with Tag Neighbors

In theory, every query has neighbors in the content and has the chance to get an expansion. Once the tag neighbors are processed, queries provided by the user for searching documents can benefit from the query expansion, which denote the relevant degree between the expansion query and original query. We consider only some set of neighbors to expand queries and select top k neighbors to generate a new expanded query by adding all the terms (see in Figure 3).

The important structure of the query expansion is: Given T as set of all tags occurring in the document corpus and a query consisting of terms $Q = \{q_i\, |q_i \in Q, i = 1, 2, \ldots z\}$, where z is the number of terms that occur in Q. For each term $q_i \in Q \cap T$ we define $Q' \subset Q$ as
$Q' = \{t_i\, |t_i \in T, i = 1, 2, \ldots m\}$, where m is the number of tags that occur in Q'.

For each $t_i \in Q'$ we retrieve all previously processed neighbors C ranked by $nf(t, c)$, where $c \in C$. We then select all neighbors $C' \subset C$ where the least ranked neighbor is the one above the threshold α (see Section III-A). Finally, we add all retrieved neighbors C' to Q expanding it to $Q = Q \cup C'$. Once this process is repeated for all tags t_i, we perform the search using the newly expanded Q expressed in the form of a query vector with a traditional information retrieval algorithm, such as in [19].

1) Working Example on Expanding Query with Tag Neighbors: In order to illustrate the query expansion, we continue working on the previous example from Figure 2 where t is the tag and a, b, c, d are the candidate neighbors. Assuming that a given query q matches the tag t the query expansion procedure begins.

According to Equation (3), the tag neighbors (whose neighborhood function values $nf(t, c) > \alpha$) of tag t are selected from the pre-processed neighborhood table. The selection order will depend on the neighborhood function values. In Table I, the candidate neighbors selection order follows d, c, b, a since their $nf(t, c)$ values are $1.464, 0.549, 0.065, 0.086$, respectively, i.e. the given query q expands with d first and next follows c, b, a. Therefore, the expanded query q' formation is equal to $\{t \wedge d, t \wedge c, t \wedge b, t \wedge a\}$. Finally, the search system performs hits the search space with the expanded query q'.

IV. Experimental Evaluation

In this section, we describe the experimental design that supports the evaluation of our proposal and the evaluation results achieved with the experiment. The goal of this evaluation is *to show that the document search enhanced by the tag neighbor expansion result in an improvement of document retrieval performance in comparison with a base line performance*. In the following parts, we detail the data collection, methodology, and metrics that were used in the experiment.

A. Dataset and Experimental Setup

1) Data: In order to test our approach, we crawled the article repository in *MedWorm* system during April 2010. MedWorm is a *medical RSS(Really Simple Syndication) feed provider as well as a search engine built on data collected from RSS feeds*[4]. We downloaded the contents into our local database. After stemming out the entity attributes from the data, four data files, namely user, resource, tags and quads, were obtained for our experiments. The fourth file represents the links between users, resources and tags. Using these data files, we generated SQL scripts to insert all data into the database. The resulting dataset comprises 949 users, 13,509 tags and 26,1501 documents. Currently, this data is available at *sourceforge*[5] .

2) Queries: The proposed approach is evaluated with MeSH (Medical Subject Headings) vocabulary(i.e queries), which are in MedWorm dataset. Selected queries are related to MeSH tree structure-2011 *C-diseases*[6]. The MeSH thesaurus is a National Library of Medicine's (NLM) controlled vocabulary for subject indexing and searching of journal articles in MEDLINE, and books, journal titles, and non-print materials in NLM's catalog [8]. These headings, also known as descriptors are organized in 16 categories: category A for anatomic terms, category B for organisms, C for diseases, D for drugs and chemical, etc., but we only consider C-diseases category due to our specific project purpose. This category is further divided into 26 sub categories as C01, C02, ..C26. Table II shows sample sub category *MeSH queries* and their candidate neighbors with frequencies.

3) Experimental setup: After data storage, we indexed the stemmed words with the help of *Lucene* Library in order to build up the search space. Lucene is a *high performance, full featured text search engine library written in Java* [7]. We utilize the processed indexing and content database to compute the tag neighborhood and calculate the neighbor frequency and weight as described in section III-A. As a result, we ended up with 15,175,334 tag neighbors with an average of approximately 9 highly ranked candidate neighbors per tag by selecting a higher value of a neighborhood function threshold $\alpha = 0.6$ III-A.

B. Evaluation Methodology and Metrics

The methodology for inferring relevance assessments is based on the ranked lists of documents submitted in response to a given query and the number of documents relevant to the query. For the human relevance assessment, we chose 106 MeSH queries according to [6] from 26 sub categories of selected C category of MeSH-2011 vocabulary. However, 6 queries did not have information in our database. So, we compute the relevant assessment for 100 queries. Each query is evaluated by top-10 and top-20 retrieved documents. For the relevance assessment, we invited experts who are familiar with medical domains to browse through the whole content of documents and assess whether they were relevant. The experts are invited biological and medical PhD students and employees of FP7 ICT project M-Eco: Medical Ecosystem Personalized Event-Based Surveillance. At last, we average the precision rate from the expert evaluations.

1) Analysis of Precision: The focus of our analysis was based on the observation of precision of our search engine. We compared our precision results with results from a baseline query search that rely on the simple user entry query (without expansion). It is very important to justify why we not addressing recall in this experiment. Because of the evaluation took place specifically on the top-10 and top-20 retrieved documents, we are unable to come up with a realistic recall analysis. As pointed out by [20], the recall should determine the ability of search engines to obtain all or most of the relevant documents in the corpus. Thus it requires knowledge not just of the relevant and retrieved but also those not retrieved. Since we do not have the precise knowledge of all relevant items within the entire corpus, it is very likely that we

[4] http://www.medworm.com/rss/aboutmedworm.php

[5] https://sourceforge.net/projects/iwis/files/

[6] http://www.nlm.nih.gov/mesh/trees.html

[7] http://lucene.apache.org/java/docs/index.html

TABLE II
Sample MeSH queries(i.e. tags) and their TOP 5 Neighbors followed by their frequencies

Query	1	2	3	4	5
influenza	contagious {0.412}	viral{0.352}	infection{0.312}	inflammation{0.284}	cold{0.185}
cancer	metastasize {0.342}	malignant{0.331}	infection{0.312}	tumor{0.198}	collon{0.1135}
diabetes	insulin {0.452}	hormone{0.421}	glucose{0.387}	blood{0.302}	pancreas{0.271}
overdose	drug {0.311}	blood{0.274}	hemoglobin{0.191}	injecttion{0.114}	cocaine{0.076}
diarrhoea	intestinal{0.324}	bacteria{0.288}	fluid{0.257}	digestion{0.184}	children{0.165}
chemotherapy	blood{0.401}	cancer{0.367}	treatment{0.247}	tumor {0.121}	cure{0.103}

TABLE III
Comparison of the performance over 100 queries

Metrics	Baseline	Our approach	% improvement
MAP	0.353	0.491	39%
P@10	0.399	0.523	31%
P@20	0.367	0.509	38%

might perform an inaccurate and non-realistic recall analysis. On the other hand, we understand that this demanding analysis must be conducted in a future experiment.

Aim at fully reflecting the performance comparisons across related works, we adopted MAP (Mean Average Precision) and P@n (precision of the first n retrieved documents) as in the *Text Retrieval Conference(TREC)*[8]. In particular, MAP, P@10 and P@20 are utilized as the performance measures during in our evaluation [21]. P@10, P@20 reflect the percentage of documents in the top 10 and top 20 of the ranked list that are relevant to the query. P@n is defined as follows.

$$P@n = \frac{\textit{number of relevant documents in top n results}}{n} \tag{4}$$

MAP stands for the *mean of the average precision* scores for a set of queries. The average precision (AP) for a single query is the mean of the precision after each relevant document is retrieved.

$$AP = \frac{\sum_{n=1}^{N} P@n}{\textit{number of relevant documents}} \tag{5}$$

where n is the rank, N is the number of retrieved documents. Finally MAP is obtained averaging the AP values over the set of queries.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{|Q|} \tag{6}$$

where Q is the number of queries and $|Q|$ is the amount of queries.

2) Empirical evaluation of tag neighbors quality: We performed an empirical analysis of tag neighbors quality on the given dataset since the performance of our approach is highly dependent on the quality of the tag neighbors utilized in the query expansion. In order to empirically assess such quality, first we set up the threshold ($\alpha > 0.6$) to cut off the unwanted tag neighbors and afterwards, we sort the tag neighbors by their neighborhood function values in descending order. Last, we invite the expert to assess the degree of relevance between the tag neighbors and parent tags.

C. Evaluation Results

1) Results: Table IV shows the sample statistics at different precision levels. The first column is about sub categories of the MeSH headings, and second column gives the query information, which we used for search. The third column gives the amount of best (for threshold $\alpha > 0.6$) available tag neighbors to expand the original query. The last four columns give the retrieval precision by P@10 and P@20.

Table III summarizes the overall performance measures MAP, P@10 and P@20. We compared our tag neighbor approach performance metrics with baseline performance metrics MAP, P@10 and P@20. According to the results of the table III, we improved our performance with 39% MAP, 31% P@10 and 38% P@20 respectively.

Figures 4, 5, and 6 show different comparison measures of baseline and our tag neighbor approach. We utilize β [1,10] scale in figures to explain the expected improvement. Specifically, we tried with β = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], where each point is equal to the average performance of the 10 queries.

2) Tag Neighborhood results for query expansion: Already we discussed in section IV-B2, the quality of the search retrieval depends on the tag neighbors quality. In order to validate this, we asked the experts (IV-B) to analyze the 20 most frequent tag neighbors derived from 575 tags (as eventual queries) randomly chosen from our database. The sample size was calculated according to [22] with the confidence level set to 95% and confidence interval set to 4%. Each neighbor was requested to be evaluated individually whether it was related or not to the parent tag. As a result, according to the expert assessment, 83% of the tag neighbors was correctly related with the parent tag while only 17% was senseless. This assessment was crucial to give credibility to the results expressed in Table III. We observed that many terms within the set of items corresponding to the 17% are mostly non-medical terms such as verbs and nouns. As part of our future works, we aim at validating such terms by consulting a medical dictionary or domain ontology vocabularies when generating the tag neighborhood table.

[8] http://trec.nist.gov, the Text Retrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval (IR) research areas, or tracks

TABLE IV
STATISTICS ABOUT A SAMPLE MESH SUB CATEGORIES, QUERIES AND PERFORMANCE MEASURES.

MeSH subcategory	Query	No.of expansion terms	Baseline(P@10)	P@10	Baseline(P@20)	P@20
Bacterial Infections and Mycoses[C01]	infection	16	0.689	**0.798**	0.673	**0.789**
Virus Diseases[C02]	arbovirus	11	0.646	**0.783**	0.641	**0.769**
Parasitic Diseases[C03]	helminthiasis	7	0.459	0.632	0.435	0.612
Neoplasms[C04]	hamartoma	1	0.145	0.145	0.0112	0.112
Musculoskeletal Diseases[C05]	dysostoses	9	0.654	0.711	0.632	0.697
Digestive System Diseases[C06]	cholangitis	8	0.599	0.688	0.572	0.671
Stomatognathic Diseases[C07]	mouth	13	0.701	**0.796**	0.699	**0.785**
Respiratory Tract Diseases[C08]	lung disease	17	0.764	**0.802**	0.732	**0.798**
Otorhinolaryngologic Diseases [C09]	nose disease	5	0.3.56	0.422	0.325	0.413
Nervous System Diseases [C10]	brain injuries	12	0.712	**0.794**	0.705	**0.786**

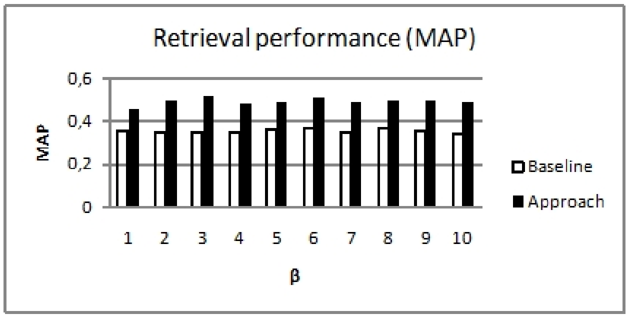

Fig. 4. Overall performance by MAP for baseline and our approach

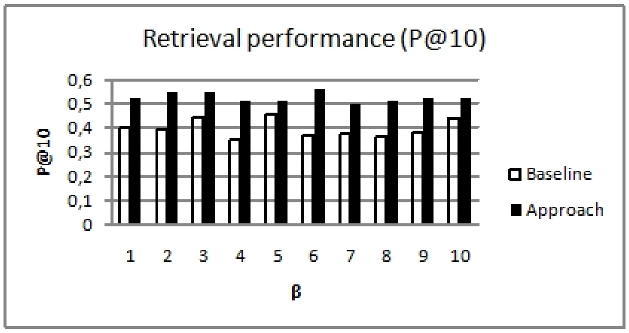

Fig. 5. P@10 measure comparison for baseline and our approach

V. DISCUSSION

In this work, we present the application of tags neighbors as an auxiliary resource towards efficient information retrieval in the medical domain. The satisfactory precision results obtained with the evaluation demonstrate the potential of the approach and allow us to discuss and compare our results with other approaches.

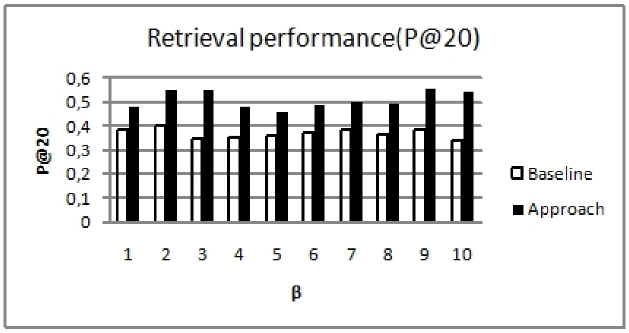

Fig. 6. P@20 measure comparison for baseline and our approach

A. Related work comparison

In order to judge the relative effectiveness of our approach, we compare our results with related studies that also address the query expansion as an instrument for effective information retrieval in the medical domain.

Focused on the mean average precision (MAP) values, [5], which applied MeSH descriptors to expand the queries by adding medical information, obtained 0.3095 on the ImageCLEPmed 2006 dataset. [6], an approach based on a knowledge-based query expansion, achieved 0.474, whereas [9] achieved 0.425. In comparison with such related works [5], [6], [9], we obtained MAP improvements at 58%, 3.5%, 15% respectively. Although we observe better results over compared approaches, this analysis should be moderately judged due to the fact that might exist differences in the evaluation methodology and/or dataset. On the other hand, this comparison gives an overview of the MAP performance among related approaches.

B. Obtained Results and MeSH Sub Categories

As explained previously, the queries issued during the experimental evaluation belongs to specific medical(MeSH) categories. By having this information available, we were able to analyze in which categories our approach performs better. Table IV shows a MeSH sub category followed by sample queries with their performance outcomes at P@10 and P@20. As shown there, the query *lung disease* generates 17 expansions and achieves P@10 at 0.802 and P@20 at 0.798. *Hamartoma* has only one expansion term and P@10 at 0.145 and P@20 is 0.112, which is almost equal to baseline search performance.

We also observe that our approach achieves better performance for the categories Bacterial Infections and mycoses [C01], Respiratory Tract Diseases[C08], Nervous System Diseases [C10], Stomatognathic Diseases[C07], Virus Diseases[C02] MeSH categories. On the other hand, our approach gives very poor results for Neoplasms[C04] MeSH sub category since the query expansion is limited to one tag neighbor. This indicates that queries with more expandable terms have higher retrieval performance, while the low amount of expandable terms has much less performance, close to the baseline. Bold numbers in the table indicate the top performances.

From this analysis, we obtained (in general) satisfactory results however its efficiency for particular categories does not perform as expected. This input opens a request for further improvements on categories such as Neoplasm [C04]. This will likely require a deeper analysis of the dataset that we were utilized for the evaluation.

C. Overall Result and Baseline Comparison

In general the overall performance was satisfactory, our results plus comparisons with related approaches and per-query analysis show the effectiveness of our approach.

Figures 4, 5, and 6 shows an comparative analysis with our obtained results with the baseline. On average, the MAP values vary in a range from 0.3 to 0.4 for the baseline approach, whereas for our approach the MAP values are increased and ranges (mostly) from 0.4 to 0.5. Additionally, our approach achieves a mean of 0.5 at precision P@10 while the baseline performs at 0.4. Similarly, our approach achieves a mean of 0.6 at precision P@20 while the baseline performs at 0.5. In summary, our evaluation results verify that *our approach outperforms the baseline query search in terms of mean average precision and precision at different stages* due to the expanded tag expression computation.

VI. Conclusion and Future Work

In this paper, we have proposed an approach of using tag neighbors for query expansion. Such approach supports users with complementary information provided by the most frequent and weighted tag neighbors occurring in the document corpus. We implemented our approach in a search system with contents extracted and indexed from the *MedWorm* medical article database. The evaluation results have shown that our proposed approach achieved substantial improvement on mean average precision and different stages of precisions compared with the same metrics of the traditional information retrieval algorithm.

As a future work, we aim at improving the quality of tag neighbors by comparing then against medical specialized dictionaries or domain ontology vocabularies. Further, we plan to realize more experimental studies necessary to validate the scalability and feasibility of the proposed approach in a broader scope. Finally, we aim at combining the current approach with other techniques previously explored such as collaborative filtering.

VII. Acknowledgments

This work has been partially supported by FP7 ICT project M-Eco: Medical Ecosystem Personalized Event-Based Surveillance under grant number 247829.

References

[1] H. Jain, C. Thao, and H. Zhao, "Enhancing electronic medical record retrieval through semantic query expansion," *Information Systems and E-Business Management*, pp. 1–17, 2010.

[2] W. R. Hersh, M. David, and H. Hickam, "How well do physicians use electronic information retrieval systems? a framework for investigation and systematic review," 1998.

[3] E. N. Efthimiadis, "A user-centred evaluation of ranking algorithms for interactive query expansion," in *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1993, pp. 146–159.

[4] C. Carpineto, R. de Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," *ACM Trans. Inf. Syst.*, vol. 19, no. 1, pp. 1–27, 2001.

[5] Z. Lu, W. Kim, and W. Wilbur, "Evaluation of query expansion using mesh in pubmed," *Information Retrieval*, vol. 12, pp. 69–80, 2009.

[6] Z. Liu and W. W. Chu, "Knowledge-based query expansion to support scenario-specific retrieval of medical free text," in *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2005, pp. 1076–1083.

[7] L. T. H. Diem, J.-P. Chevallet, and D. T. B. Thuy, "Thesaurus-based query and document expansion in conceptual indexing with umls."

[8] M. Daz-Galiano, M. Martn-Valdivia, and L. Urea-Lpez, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Computers in Biology and Medicine*, vol. 39, no. 4, pp. 396 – 403, 2009.

[9] S. Matos, J. Arrais, J. Maia-Rodrigues, and J. Oliveira, "Concept-based query expansion for retrieving gene related publications from medline," *BMC Bioinformatics*, vol. 11, no. 1, p. 212, 2010.

[10] C. E. Bianco, "Medical librarians' uses and perceptions of social tagging," *Journal of the Medical Library Association : JMLA*, vol. 97, no. 2, pp. 136–139, 2009.

[11] J. West, "Subject headings 2.0: Folksonomies and tags," *Library Media Connection*, vol. 25, no. 7, pp. 58–59, 2007.

[12] S. Jin, H. Lin, and S. Su, "Query expansion based on folksonomy tag co-occurrence analysis," *Granular Computing, 2009, GRC '09. IEEE International Conference on*, pp. 300–305, 2009.

[13] L. Gordon-Murnane, "Social bookmarking, folksonomies, and web 2.0 tools," *Searcher Mag Database Prof*, no. 6, pp. 26–28, 2006.

[14] A. Milicevic, A. Nanopoulos, and M. Ivanovic, "Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions," *Artificial Intelligence Review*, vol. 33, no. 3, pp. 187–209, Mar. 2010.

[15] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo, "Fast tagging of medical terms in legal text," in *Proceedings of the 11th international conference on Artificial intelligence and law*, ser. ICAIL '07. New York, NY, USA: ACM, 2007, pp. 253–260.

[16] H. Jang, S. K. Song, and S. H. Myaeng, "Semantic tagging for medical knowledge tracking," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, 30 2006.

[17] P. Ruch, J. Wagner, P. Bouillon, R. H. Baud, A. M. Rassinoux, and J. R. Scherrer, “Medtag: tag-like semantics for medical document indexing.”

[18] J. SB, “A semantic lexicon for medical language processing,” pp. 205–218.

[19] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.

[20] S. J. Clarke and P. Willett, “Estimating the recall performance of web search engines,” *Aslib Proceedings*, vol. 49, no. 7, pp. 184–189, 1997.

[21] H. Jain, C. Thao, and H. Zhao, “Enhancing electronic medical record retrieval through semantic query expansion,” *Information Systems and E-Business Management*, pp. 1–17, 2010.

[22] T. W. Anderson and T. W. Anderson, *An Introduction to Multivariate Statistical Analysis, 2nd Edition*, 2nd ed. Wiley-Interscience, September 1984.