A Hybrid Method for Retrieving Medical Documents with Query Expansion

Jiyeon Choi
Knowledge Service Engineering
KAIST
Daejeon, Korea
Email: jeeyeon51@kaist.ac.kr

Youkyoung Park
Knowledge Service Engineering
KAIST
Daejeon, Korea
Email: park60@kaist.ac.kr

Mun Yi
Knowledge Service Engineering
KAIST
Daejeon, Korea
Email: munyi@kaist.ac.kr

***Abstract*—Query expansion is a well known method widely used to improve the efficiency and precision of information retrieval in diverse fields. However, information retrieval in the medical domain still confronts many challenges due to the vastness of jargons and inconsistency of terms, leading to poor performance in information retrieval. To circumvent these problems, the main strategy that has been used is to rely on medical ontologies known as an intensional approach despite the fact that it lacks range of expressions. In contrast, an extensional approach is based on external resources such as documents, notwithstanding its own weaknesses. Thus in this paper we propose a hybrid approach, which combines the two approaches along with a refinement technique, in order to overcome each approach's weaknesses while creating a synergistic effect that maximizes each approach's strengths. The effectiveness of this framework is tested through an experiment using TREC-CDS Track 2014 data. Based on the positive results, we suggest this hybrid approach as a viable solution in query expansion for the medical domain.**

***Keywords*—** query expansion, hybrid approach, medical document search

I. Introduction

Query expansion has been widely studied to improve information retrieval performance in document search and retrieval operations [1], [2], [3]. This approach has been proven to be helpful in improving the efficiency and precision of information retrieval in various studies. However, most of the work focuses on search with regard to the website or general documents, leaving many challenges still in specialized areas such as the medical domain. The medical domain holds its own domain-specific characteristics which needs to be carefully considered. For instance, the use of terms across experts, textbooks, and individuals is not consistent, and also the terms of old and new are mixed. It is why search performances are poor within this domain when using relatively simple queries, meaning that query expansion can play an important role in this domain. Thus, in this paper our goal is to establish a query expansion framework appropriate for the medical domain addressing its challenges and verify its effectiveness through an experiment.

According to [4], the approaches of query expansion can be categorized into three different branches. First is an intensional approach, which augments the query based on the meaning of the words used in the initial query. This is preferred when an ontology or a thesaurus exists, on account of the fact that the keywords' relationships are well organized. Prior studies such as [5] and [6] are examples of this approach using an ontology. By contrast, rather than narrowing the scope of focus to the keywords of the query itself, an extensional approach utilizes external resources to reformulate a query. Some typical examples of this approach are relevance feedback and local analysis, each of which builds upon the outcome of the initial query. In case of relevance feedback, it extracts new query words from users feedback, which are documents that users found it relevant. The local analysis method relies on the top-ranked results. Lastly, collaborative approach tries to discover the current user's intention by analyzing the historical data in terms of prior queries entered by one or more users.

The medical sector is a field with a vast number of professional words and jargons, and its expressions vary by source. Hence, there has been much effort to organize the link of medical terms, for which UMLS(Unified Medical Language System) and MeSH are two popular solutions - the former as an ontology and the latter as a thesaurus. With the availability of these solutions, query expansion in the medical domain has relied on the use of mainly intensional strategies. However, when queries are tied to an existing net of words it is difficult to construct a new query outside the dictionaries and it is mostly unfeasible to consider the context of the document. Thus, in this paper we propose a hybrid approach, which combines the two approaches along with a refinement technique, in order to overcome each approach's weaknesses while seeking to create a synergistic effect that maximizes each approach's strengths. Considering the limitless range of terms the extensional approach can cover, the points the intensional methods overlook can be captured by the hybrid approach. Meanwhile, it is difficult to conclude whether or not the modified query is relevant to the initial query when using the extensional method only, which can be solved by utilizing a reliable ontology. We expect that this mixed approach will increase the search performance compared to the baseline of involving one approach only. We carried out an experiment to test our framework, and evaluated its effectiveness in searching performances within the medical domain.

II. Related Works

A. Medical Domain

1) Intensional Approach: The medical community has been putting much effort to organize their domain language, resulting in several ontological resources available, which has promoted active research in intensional query expansion. For instance, [5] constructs a new query by using the synonym and heading relationship extracted from MeSH. Image search was tested simultaneously to enhance the search performance, but using MeSH turned out to be better, while the integration of both sources presented the best result. The work by [6] focuses on the clinical sector and rebuilt an ontology to find new semantic relationships for their query reconstruction rather than relying on a existing ontology. Recently, not only synonyms and high level terms but also new methods such as random walk were applied to diversify the query words. This is the case of [7], in which UMLS and graph algorithm based on random walk were used to formulate queries.

As seen above, to improve search qualities in the medical domain, the mainstream approach of query expansion was to concentrate on the meanings of the terms, therefore searched for a new term that involved initial terms meanings when expanding queries. However, these studies reveal weakness in adding queries based on conventional terms with which semantic relationships do not exists.

2) Extensional Approach: As considering intensional approaches cannot capture the context of the document, recent studies attempted to extract context terms from collected documents. [8] observed an increase in search performance by adopting various terms from a mix of documents to expand a query. [9] experimented the use of pseudo relevance feedback, which is one of the extensively applied method in the general domain, to find a way to efficiently utilize an external corpuse in the medical domain. In this study various types of documents were put together while different weights were given according to feedback. One type of document was considered as a cluster based on the cluster-based document model, which was calculated through probability distribution of the terms to search proper words for query augmentation. These studies assume that terms can be detected sufficiently from the documents thus they do not refer to any ontologies. However, it is obvious that there is a high chance of low quality in the relationships of terms found from documents when compared with the relationships found within an ontology, which are defined by human.

B. General Domain

1) Word Refinement: In the general domain, many researchers primarily chose extensional approach using numerous documents to formulate queries as it is almost impossible to construct an ontology of everything or disambiguate words explicitly. As a result of extensive work in this area, this line of research now has evolved from extracting terms to refinement of terms that are more useful for query formulation.

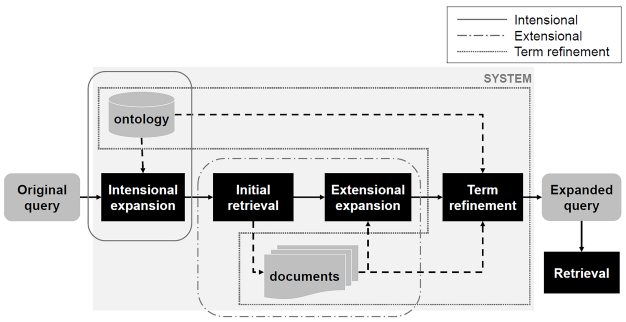

Fig. 1. Framework of hybrid expansion method

III. Framework

In this paper we propose a hybrid framework of query expansion as shown in Fig.1, which contains 3 parts: intensional expansion, extensional expansion, and word refinement. When a user inputs a query, first the intensional module retrieves terms from the ontology that have the same meaning with the words in the input. Next, the query constructed one step before returns searched documents which are the basis for the formulation of the extensional query. Lastly, term features are extracted from the ontology and documents respectively and then machine learning technique is used to rank the terms. Only the top ranked terms are added to the initial query to construct a final query for actual search.

A. Intensional Expansion

The intensional expansion module consists of initial query and ontology. The ontology provides synonyms for each word included in the given query and those words are added to the candidate query list. This process complements the search for documents that could have been ignored due to term differences in expression but with the same meaning. The initial given query and the added words list is the result of this module, which is then sent to the next module, the extensional expansion.

B. Extensional Expansion

This part is composed of a search module, retrieved documents, and a module that expands queries extensionally. The search module runs when it receives words list from the intensional module, and returns documents related to the terms from the list. New words are extracted from the documents that are judged to be related with the initial words and added to the candidate list. This step is required to extract not only the synonyms but also terms that are only searchable in the documents. The processed list is then passed to the final stage where words are refined.

C. Word Refinement

The final form of query expansion is completed after the word refinement module where the keywords for the query are selected from the candidate list. This part involves the module that refines terms, ontology, and retrieved documents. The ontology and the documents are employed to rank the

TABLE I
TYPES OF QUESTIONS

Type	Generic Clinical Quesion	Number of Topics
Diagnosis	What is the patient's diagnosis?	10
Test	What tests should the patient receive?	10
Treatment	How should the patient be treated?	10

TABLE II
EXAMPLE OF A TOPIC

No.	Type	Summary	Relevant Articles
1.	Diagnosis	A woman in her mid-30s presented with dyspnea and hemoptysis. CT,scan,revealed a cystic mass in the right lower lobe. Before she received, treatment, she developed right arm weakness and aphasia. She was, treated, but four years later suffered another stroke. Follow-up CT scan, showed multiple new cystic lesions.	3148967, 3082226, 2987927

candidate terms. Several features extracted from the ontology and documents are calculated based on the given data, which then machine learning technique is applied to compute the rank. Only the top k terms are selected from the list and determined as the final query terms.

IV. EXPERIMENT

A. Data

In our experiment we used the TREC-CDS(Clinical Decision Support) Track 2014 [10] data to test our system performance. This data set contains 733,138 non-image text documents from PubMed Central(PMC) collected till Jan. 21, 2014. It is provided in NXML format, which are files encoded in XML by NLM Journal Archiving and Interchange Tag Library. Each document is labelled with PMCID that identifies the records. In the track, 30 topics that describe a medical status are provided as shown in Table I and the task is to find a suitable document that might be a possible answer to the given query. There are 10 topics for each 3 types of question. Table II shows an example of a topic. The answer set and the PMCID of relevant documents are also provided for each topic.

B. System

Because the given input, a patients status, is presented in lines of sentences, we first went through a process that extracts a few keywords for the search task. This keyword extraction process was done by a team of two coders consisted of a medical student and a non-medical student. First each member read the description of a patient and then selected some query words. Second they selected the words that both person chose and, if not sufficient, words were added after discussion. The query words were limited to five separate words.

The ontology used in this study in conjunction with the intensional expansion is the UMLS, which consists of 3 parts: metathesaurus, semantic network, and specialist lexicon. The metathesaurus covers biomedical concepts and its names, and the relationships of the concepts. The semantic network component connects the semantic types by semantic relationships. The specialist lexicon provides lexical information needed for the natural language processing system. In this paper we used the MetaMap, which is a tool that can extract terms from the metathesaurus to expand the query words based on their meaning.

Using Lucene as the search engine required for the extensional expansion, the given data set of 733,138 NXML files were parsed and indexed, and basic search provided in Lucene was used. When extracting words from the retrieved documents the term dependency model [3] was performed. It is basically determined that the higher chance terms co-occur together has a closer relationship. The co-occurrence of the terms was measured by Equation 1. Fundamentally, the term dependency model assumes that terms are dependent on each other, but in this study we considered two different cases of 1) completely independent and 2) dependent of two words near each other, and calculated the linear interpolation of the two to measure the words properness. In case of complete independence Equation 2 was applied while Equation 3 was used for the case of dependent relationship. The linear interpolation is calculated by Equation 4.

$$cooc(t_i, q_j) = \frac{\sum_{d \in D} \log(tf(t_i|d) + 1) \log(tf(q_j|d) + 1)}{\log N} \tag{1}$$

$$coof_{single}(t_i, Q) = \sum_{q_j \in Q} idf(t_i) idf(q_j) \log(cooc(t_i, q_j) + 1) \tag{2}$$

$$coof_{bigram}(t_i, Q) = \sum_{j=1}^{n-1} idf(q_i) idf(q_{j+1}) idf(t_i) \tag{3}$$

$$coof(t_i, Q) = (1 - \lambda) coof_{single}(t_i, Q) + \lambda coof_{bigram}(t_i, Q) \tag{4}$$

Finally, the words that are more suitable were selected through a machine learning technique, rather than having all the words expanded by intensional and extensional methods. The Learning to Rank [11] algorithm of machine learning ranked the words using several features such as overall term frequency, document frequency, and feedback term frequency. These features can be divided into two different categories, ontology and document features. The document features were adopted from a recent paper [12] regarding word refinement while the features from ontology were values of similarity and relationship between extracted words from documents and the terms included in the query given by the user. The similarity and relationship values are calculated based on UMLS Similarity.

TABLE III
EXPERIMENTS COMPARISON

	Intensional	Extensional	Term Refinement	
			with Ont.	with Doc.
Baseline				
Model 1	*		*	
Model 2		*		*
Proposed Model	*	*	*	*

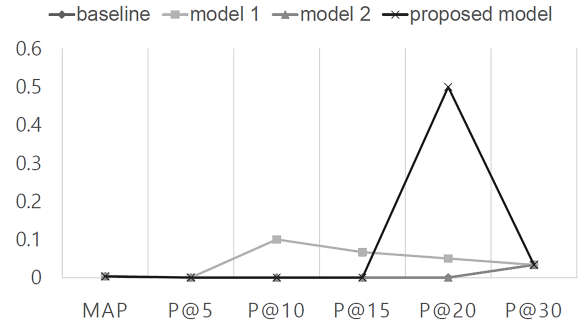

Fig. 2. Experimental results

C. Evaluation

As shown in Table III four models were tested including the baseline. The baseline model is the case in which query expansion has not been conducted. In Model 1 only the intensional expansion and the use of ontology for word refinement based on documents was implemented while in Model 2 only the extensional expansion and word refinement based on documents was implemented. Our main objective in this study is to propose Model 3, which included both the intensional and extensional expansions, and assess its effectiveness over the other models. Both the ontology and the documents were used for word refinement to extract feature values. The search performance was compared by using the final query, which corresponds to the last column of the table, and MAP and P@n were used as evaluation criteria.

D. Results

As shown in Fig. 2, there have been noticeable improvements in search performance in Model 1, in Model 2, and mostly in our proposed model. Although the experiment was conducted with only one topic, or a query, all three models' performances were enhanced or at least maintained its baseline level. Furthermore, the enhancements can be adjusted by the parameters: 1) k in top-k used when words were extracted from documents 2) lambda from term dependency model when conducting linear interpolation 3) inner parameter and LTR algorithm from term refinement.

V. CONCLUSION

In this paper, we proposed a hybrid framework of query expansion that can improve information retrieval performance in the medical domain. We assumed that applying both the intensional and extensional approaches to expand queries would result in improved performance. We tested our proposed model using TREC-CDS Track 2014 data and demonstrated the superiority of the proposed model. It is expected that by adjusting certain parameters the observed positive effects can be larger. In summary, we argue that research on query expansion, especially for the medical domain, should be further developed in the line of integrating intensional and extensional approaches so as to overcome the known weaknesses of each approach alone. However, as our evaluation was conducted using a small number of queries, more validation of the proposed framework is needed. Notwithstanding this limitation, the current study shows promise for an integrative approach of query expansion.

ACKNOWLEDGMENT

This work was supported by the Industrial Strategic Technology Development Program, 10052955, Experiential Knowledge Platform Development Research for the Acquisition and Utilization of Field Expert Knowledge funded by the Ministry of Trade, Industry & Energy (MI, Korea).

REFERENCES

[1] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 4–11.

[2] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Probabilistic query expansion using query logs," in *Proceedings of the 11th international conference on World Wide Web*. ACM, 2002, pp. 325–332.

[3] Y. Lin, H. Lin, S. Jin, and Z. Ye, "Social annotation in query expansion: a machine learning approach," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 405–414.

[4] F. A. Grootjen and T. P. Van Der Weide, "Conceptual query expansion," *Data & Knowledge Engineering*, vol. 56, no. 2, pp. 174–193, 2006.

[5] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. Ureña-López, "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.

[6] A. Babashzadeh, J. Huang, and M. Daoud, "Exploiting semantics for improving clinical information retrieval," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 801–804.

[7] D. Martinez, A. Otegi, A. Soroa, and E. Agirre, "Improving search over electronic health records using umls-based query expansion through random walks," *Journal of biomedical informatics*, vol. 51, pp. 100–106, 2014.

[8] D. Zhu, S. Wu, B. Carterette, and H. Liu, "Using large clinical corpora for query expansion in text-based cohort identification," *Journal of biomedical informatics*, vol. 49, pp. 275–281, 2014.

[9] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 186–193.

[10] M. S. Simpson, E. Voorhees, and W. Hersh, "Overview of the trec 2014 clinical decision support track," in *Proc. 23rd Text Retrieval Conference (TREC 2014). National Institute of Standards and Technology (NIST)*, 2014.

[11] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 129–136.

[12] B. Xu, H. Lin, and Y. Lin, "Assessment of learning to rank methods for query expansion," *Journal of the Association for Information Science and Technology*, 2015.