IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB12781

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour

Query Reformulation with Information-Based Query Expansion for Handling Medical Scenario Queries

Yong Jun Choi, The George Washington University, 11333 Westbrook Mill LN #304, Fairfax, VA 22030,
T 703-352-0666, F 831-597-3767, yongj@gwu.edu\

ABSTRACT

In this paper, I propose an information-based query expansion technique to support scenario specific retrieval in the medical domain. An information-based query expansion technique takes advantage of the UMLS (United Medical Language System) information source to append the original query with additional terms that are specifically relevant to the query's scenario, thus improving upon traditional query expansion approaches. I compare this technique with cross-search method that only refer to the encyclopedia and expand terms that are not necessarily scenario specific. The study on the clinical notes shows that the information based techniques that results in scenario-based expansion outperformed over the cross-search and automatic query expansion method on average in all categories of scenarios.

1. INTRODUCTION

A number of approaches to query expansion have been studied for decades as an effective method to improve the query document mismatch problem. The basic idea behind all techniques is to supplement the original query with additional terms related to the original query topic so that the modified query has a better chance to match relevant documents.

In clinical practices, doctors are often interested in answers relevant to certain scenarios that correspond to common tasks in medical practice such as "diagnosis", "treatment", "symptom" etc. As a result, queries they pose are frequently scenario specific like "liver cancer, diagnosis". Studies show that 60% of doctors' queries center around a limited number of medical scenarios such as "treatment", "diagnosis" etc in clinical practice. [1]

Retrieving documents that are specifically related to the query's scenario is referred to as scenario-based retrieval. Scenario terms in the queries are typically general such as "diagnosis, liver cancer", while full-text medical documents often discuss the same topic using much more specialized terms such as "chemoembolization". Such general scenario terms fail to match with the specialized terms in relevant documents, resulting in poor retrieval performance.

The fundamental challenge is that scenario terms in the query are too general to match specialized terms in relevant documents like "chemoembolization" which is one of treatment options for liver cancer. Therefore it is often desirable to retrieve only documents pertaining to a based medical scenario where a scenario is typically defined as a frequently reappearing medical task. For example, a doctor can pose a query "liver cancer, diagnosis" to find out the latest diagnostic techniques about the disease in diagnosing a potential liver cancer patient. In this case, "diagnosis" is the medical task that marks the scenario of the query.

Scenario based retrieval is not adequately addressed by traditional text retrieval systems such as SMART and such systems suffer from the fundamental problem of query document mismatch when handling scenario-based queries. [2]

There has been research on query expansion to improve the query document mismatch problem. [3-6] Those techniques also have difficulties handling scenario-based queries. In principle, query expansion techniques append the original query with specialized terms that have a statistical co-occurrence relationship with original query terms in medical literature. Even if adding such specialized terms makes the expanded query a better match with relevant documents, the expansion is not scenario based. For example existing query expansion techniques can add not only terms such as "chemoembolization" that is relevant to the treatment scenario but also irrelevant terms like "alcohol" simply because the term co-occur with "liver cancer" in medical literature in handling the query "liver cancer, treatment".

Adding non-scenario-based terms leads to the retrieval of documents that are irrelevant to the original query's scenario, diverging from the goal of scenario-based retrieval. Moreover, expanding just synonyms without considering the scenario-based information embedded in the original query is not sufficient in dealing with such queries. For example, previous methods will exclude "chemoembolization" from the expansion list for query "liver cancer, treatment" simply because "chemoembolization" is not a synonym of any original query concept.

For the demand of doctors such as finding clinical documents in the event of scenario-based retrieval, I propose specially developed technique called information based query expansion and cross-search query expansion that allow more convenient searching of clinical documents. However, the conventional methods of informational retrieval [7] are not suitable for this task. For this design, I have applied the associative access method (ASSA). [8] The ASSA method is based on constructing of bit-attribute matrix by transforming a free text into a set of trigrams. [9] This gives rise to possibilities for approximate matching of pieces of distorted text and fuzzily formulated queries because clinical documents have lots of spelling mistakes.

The rest of this paper is structured as follows. In section 2, cross-search and information-based query expansion concepts are described. I experimentally evaluate the techniques and present the results in section 3. Section 4 concludes the paper.

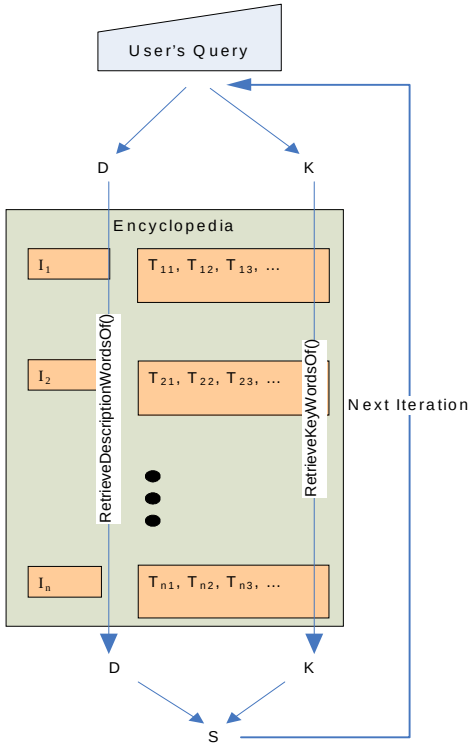
2. PROPOSED QUERY EXPANSION TECHNIQUES

2.1 Describing Cross-Search Query Expansion Concepts

Cross-search in this paper is defined as a search method to find related information by searching for the additional vocabulary. [10] The vocabulary is expanded by collecting relevant words from a relevant context in another knowledge base to which the original inquiry belongs.

An encyclopedia is a knowledge base that is arranged in pairs of a word and its description, and the description is accessible by any key word or word in the description. Cross-search on clinical documents is done against any encyclopedias of choice.

Figure 1. Cross-search mechanism



When users search for documents with the search engine, they type keywords into the engine. The cross-search algorithm is working to find documents.

Figure 1 shows the cross-search mechanism for searching documents.

I_n is a key word. Each key word has an associated description that consists of description words, T_{n1}, T_{n2}, …

There are two operations that can be done against an encyclopedia.

RetrieveDescriptionWordsOf(I_n): The encyclopedia can be searched for a key word, I_n to retrieve the associated description words that are also key words, {T_{n1}, T_{n2}, …}. That way only significant description words are extracted.

RetrieveKeyWordsOf(T_{mk}): The encyclopedia can be searched for a description word, T_{mk} to retrieve the associated key words, {K_p, K_q,…}

Therefore there are two different kinds of cross-searches that can be performed to expand the related vocabulary set, S through the encyclopedia.

For preparation for cross-search by description words, Set D is initialized with each word of the user's query. RetrieveDescriptionWordsOf() search is performed for each element of D and it adds the result to D

For preparation for cross-search by key words, Set K is initialized with each word of the user's query. RetrieveKeyWordsOf() search is performed for each element of K and it adds the result to K.

The related vocabulary set, S is built by unioning D and K.

$S=D \cup K$.

Then the related vocabulary set, S, is searched for in the clinical documents again with the search engine. The clinical documents are displayed with the related vocabulary highlighted. Thus they can be skimmed through by looking at the sections with highlights. As a result, doctors can find useful information in the clinical documents even though they don't know exact words or phrases.

Once the related vocabulary set, S, is built. It can be fed into the beginning of cross-search as an input set instead of the user's query. If the wanted information is not found in iteration, more related vocabulary can be found in the next iteration. As more iteration of cross-searches is performed, more information related to the initial user query would be found.

2.2 Describing Information-Based Query Expansion Concepts

Cross-search expansion concepts that are related to the original query have been described. Only a subset of these candidate concepts is relevant to the original query's scenario. I have developed a special query expansion technique called information-based query expansion for doctors to find relevant clinical documents in the scenario specific query. A method that automatically takes advantages of the knowledge structures in the semantic network and UMLS is designed to identify concepts that are specifically related to the outline of scenarios in the original query. Appending such identified concepts to the original query will result in scenario-based expansion. Retrieving documents that are specifically related to the query's scenario is referred to as scenario-specific retrieval.

To develop the idea in full details, in the following, I first introduce the information structure used in this study and then describe the information-based methods. Figure 2 depicts the components in an information based query expansion and retrieval framework.

Given an original query such as "liver cancer, diagnosis", the information-based query expansion whose scope is marked by the rectangle derives the scenario-based expansion concepts, with the aid of domain knowledge such as UMLS in addition to the cross-search technique marked by the 3D box.

The basic idea of the information-based method is the following. A scenario-based query consists of two parts: a key concept Ck (e.g., "liver cancer") and several scenario concepts Cs's (e.g., "treatment," "diagnosis," etc.) because doctors often pose queries like this to search clinical documents.

Using the Cross-Search expansion, we only can get candidate expansion concepts with the key concept Ck co-occurring with the key concept Ck, e.g., "alcohol," "chemoembolization," etc., for Ck ="liver cancer."

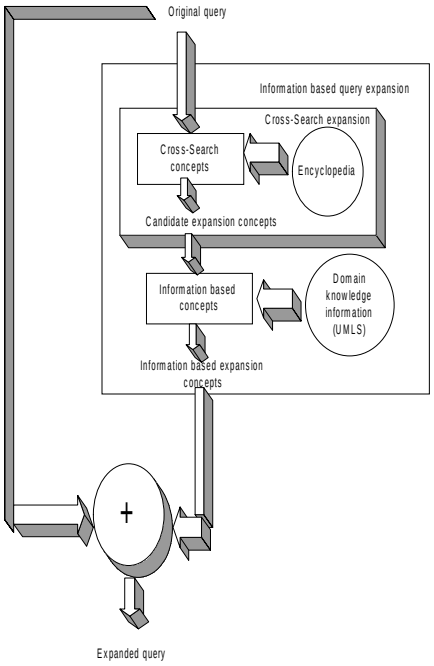
In information-based expansion, we can explore a domain-based information source to identify possible relationships between each candidate expansion concept and Ck. For example, the information source may indicate that "alcohol" is a "risk factor" for "liver cancer," whereas "chemoembolization" is a "treatment" method for this disease. Among these identified relationships, certain relationships are desirable because they match with scenarios of the original query. Thus, our information-based method will keep only the candidate concepts that have a desirable relationship with Ck. Since such concepts should be specifically relevant to the original query's scenarios, appending such concepts should lead to scenario-based expansion.

UMLS has been used for the domain-based information source to retrieve free text in clinical documents. It has the Metathesaurus and semantic network. The Metathesaurus has more than 800,000 medical concepts and a group of concepts in the Metathesaurus belong to a semantic type in the semantic network in UMLS. For instance, "liver cancer" and other disease concepts belong to one semantic type called "disease." Given this structure, here are the procedures to identify the scenario-based expansion concepts.

First, a key concept (the name of a disease) K and a scenario term such as "treatment", "diagnosis", "symptom" are chosen together by a doctor.

K identifies the semantic type it belongs to (e.g. from "liver cancer" to "disease") by referring to UMLS and then it reaches a set of relevant vocabulary sets. [11] Those relevant terms are selected and appended to the original query.

Figure 2. An information-based query expansion and retrieval framework



3. EXPERIMENTAL SETUP

The experiment in this paper uses clinical documents for medical information retrieval using ASSA. ASSA (Associative Access Method) is a fuzzy search engine that can quickly find information on the PC, databases, file servers, the Internet just about anywhere.

The corpus consists of 750 clinical documents and each document contains diagnosis, treatment, symptom, author information, document ID etc.

The query set consists of 12 queries. Each query is short and contains an information request. A key concept is a name of disease. Scenario concept can be diagnosis, treatment, and symptom. All queries are in the form of ("key concept", "scenario concept"). A document is judged by a doctor as relevant, irrelevant for a given query.

3.1 Comparison of the Three Methods (Information Based Expansion vs. Cross-Search Expansion vs. Automatic Query Expansion)

I have performed extensive experimental evaluation of the information-based method by comparing against the cross-search expansion method and automatic query expansion.

This study creates information based query expansion and sees if it presents improvements over the cross-search expansion and automatic query expansion method when handling scenario-based queries. The pseudo-relevance feedback method is used for automatic query expansion. [12]

The query expansion techniques may perform differently for different query scenarios, so I have studied how each expansion technique performs in different scenario. To do this, 12 queries have been grouped according to the scenario such as diagnosis, treatment, symptoms.

I have compared the performance results depending on each group of queries – information-based expansion vs. cross-search expansion vs. automatic query expansion under the same settings.

The 11 point precision average is used to measure the performance of each method.

3.2 Experiments Results and Analysis

In this section, I present the performance of the information based expansion technique compared to that of the cross-search expansion and automatic query expansion technique.

I average the performance of three expansion techniques within each group of queries and show the results in Figure 3. Each bar shows the performance of information-based expansion averaged over the cross-search expansion and automatic query expansion under the same settings.

The results of applying the query expansion techniques are summarized below (and in Figure 3). All values are given in terms of 11-point average precision.

A. Diagnosis scenario
Automatic query expansion: 2.8

Cross-search query expansion: 3.7

Information-based expansion: 4.5

B. Treatment scenario
Automatic query expansion: 3.0

Cross-search query expansion: 3.8

Information-based expansion: 5.1

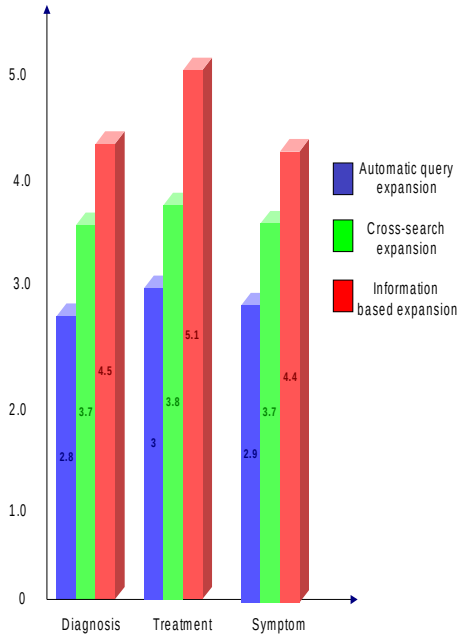
C. Symptom scenario
Automatic query expansion: 2.9

Cross-search query expansion: 3.7

Information-based expansion: 4.4

The results show that the information-based technique can create scenario-based query expansion and produces improvements over cross-search and automatic query expansion technique when handling scenario-based queries.

Figure3. Comparison of retrieval results in different scenarios



The results also suggest that information-based expansion performs differently for queries with different scenarios. This may happen because the knowledge structures defined for these scenarios exhibit different characteristics.

The information-based technique produces more improvements in the "treatment" than "diagnosis" or "symptom" scenario. A possible explanation lies in the different information structures for these three scenarios. For example, there are more relevant semantic types than those in the "diagnosis" or "symptom" scenario in the treatment scenario.

4. CONCLUSION

I have proposed an information-based query expansion method to improve the retrieval performance when handling scenario queries for which doctors are often searching.

The previous studies have not tried to take advantage of a domain-based information source to reformulate the query expansion results and provide scenario-based expansion.

This research focuses on a type of medical queries, namely scenario-based queries, which have been shown to be predominant among medical users' search requests.

An information-based query expansion method is presented to improve the retrieval performance for such queries and it is a method that automatically takes advantage of the knowledge structures in the semantic network and the UMLS to identify concepts that are specifically related to the scenarios in the original query. Adding such identified concepts to the original query results in scenario-based expansion and improves the search performance.

The experiments reported in this paper have examined the performance of retrieval results of automatic query expansion, cross-search expansion, and information-based query expansion. The results suggest that information-based query expansion provides more consistent increases in retrieval effectiveness.

I conclude that an information-based query expansion along with UMLS is an effective method of enhancing retrieval effectiveness when handling medical scenario queries.

REFERENCES

J.W. Ely, J.A. Osheroff, M.H. Ebell, G.R. Bergus, B.T. Levy, M.L. Chambliss, and E.R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7); 1999 211-220,

E.N. Efthimiadis. Query expansion, American Society for Information Retrieval by Information Today, Inc. Medford, NJ. *Annual Review of Information Science and Technology*, 1996 31: 121–187.

Y. Qiu and H.P. Frei. Concept-based query expansion. In *Proceedings of ACM SIGIR '93*, 1993. 160-169

Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO '94*, 1994. 146-160

J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proceedings of ACM SIGIR '96*, 1996. 4-11

M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of ACM SIGIR '98*, 1998. 206-214

W. B. Frakes and R. Baeza-Yates, (eds.), *Information Retrieval - Data Structures & Algorithms, Prentice Hall PTR, Saddle River*, NJ, 1993.

G. M. Lapir, Use of Associative Access Method for Information Retrieval Systems, *Proceedings of the 23rd Annual Pittsburgh Conference on Modeling and Simulation*, vol. 23, part 2, 1992 951-958

S. Berkovich, E. El-Qawasameh, G. M. Lapir, M. Mack, C. Zincke, Organization of Near Matching in Bit Attribute Matrix Applied to Associative Access Methods in Information Retrieval, *16th IASTED International Conference on Applied Informatics, IASTED*, 1998 62-64.

Y. Choi , J. Byun, S. Berkovich, Cross-search technique and its visualization of peer-to-peer distributed clinical documents, *(ICIT) International Conference on Information Technology*, Turkey, 2004 49-53,

A.T. McCray, S.J. Nelson., The representation of meaning in the UMLS. *Methods Inf Med*;34(1-2): 1995 193-201

C. Burkely, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART:TREC-3. *In proceedings of the third text retrieval conference(TREC-3)*, 1994 69-80