A Knowledge-based Approach for Scenario-specific Content Correlation in a Medical Digital Library*

Wesley W. Chu and Victor Z. Liu
Computer Science Department, Email: {wwc, vicliu}@cs.ucla.edu
University of California at Los Angeles, Los Angeles, CA 90095

Manually searching the web for medical literature and teaching materials is labor-intensive and time-consuming. Content Correlation automatically creates semantic links among documents from different collections. As a result, navigating from a patient report to online medical documents is much easier for the user. In this paper, we present a knowledge-based (e.g. UMLS) approach for content correlation. First, we index each medical document using phrases (a combination of word stems and concepts). We illustrate how phrase-based indexing greatly ameliorates the problem of vocabulary mismatch among multiple document collections. Second, we use a phrase-indexed patient report to automatically form a query, and expand the query with scenario-specific phrases derived from a knowledge base. Experimental results reveal that the phrase-based indexing and the knowledge-based query expansion together, yield scenario-specific content correlation.

1 INTRODUCTION

A large number of medical information systems have emerged on the Web with comprehensive coverage of medical literature and teaching materials, e.g PubMed,[1] Harrison's Online.[2] However, the search interfaces of these web sites hinder users from fully utilizing them in health care applications. Consider a typical clinical environment in which a physician tries to find diagnosis or therapy options for a patient's disease based on the patient's past clinical reports and major complaints. In order to efficiently locate the most relevant medical literature on the Web, the physician has to manually form a short query made up of a few keywords. The keywords need to be carefully selected to best summarize the patient's past history and symptoms, and to clearly define the physician's specific information needs, e.g regarding "diagnosis," "treatment" or "cause" of the patient's disease. Recent studies reveal that searching the Web for clinical usage is frustrating, labor-intensive and time-consuming [1, 2, 3, 4, 5]. Although about one third of a physician's clinical questions can be answered by online information resources [6], the overall usage of the Web in medical practice is relatively low [1, 7].

Content Correlation is a technique that automatically generates semantic links between a document in one collection to relevant documents in other collections [8]. Figure 1 illustrates this process in a Medical Digital Library. Using content correlation, a physician first forms his query by selecting one patient report (Step ①). The *Content Correlation Engine (CCE)* automatically navigates from the seeding report to online information resources, forms a query using pertinent concepts to the original patient report (Step ②), submits the query to each resource and collects the corresponding answers (Step ③). Compared to manual searching, Content Correlation provides a more convenient method for the user to efficiently navigate among various information sources [8].

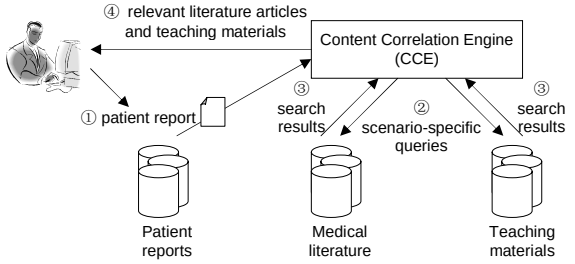

Figure 1: Content Correlation in a Medical Digital Library

Two main challenges exist in designing a CCE. First, different information resources tend to use different expressions for the same concept. This is often referred to as the *vocabulary mismatch* problem. For example, "cancer" is used frequently in patient reports, while its alternative expression "carcinoma" often appears in literature articles. Using the original string forms fails to link the patient report on "cancer" to a literature reference discussing "carcinoma." Concepts have been proposed to replace word stems to index documents, so that different expressions are matched to the same concept. However, because knowledge sources are often incomplete in defining concepts, concept-based indexing has not achieved significant improvement [9, 10, 11]. In this paper, we propose a phrase-based indexing technique that uses both word stems and concepts to index a medical document. This technique compensates for the incompleteness of concepts using word stems, and greatly enhances our system's capability to correlate documents from different collections [12].

Second, users' information needs are scenario-specific. Ely et al [1, 13] and Haynes et al [7] discover that physicians' clinical questions largely focus on certain specific aspects of a particular disease, e.g therapy options, drug usage or diagnostic work up. This requires the CCE to generate scenario-specific queries that focus on a specific aspect of the disease concept. Simply using the disease concept to form a query

*This research is supported by NIC/NIH Grant #4442511-33780
[1] www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
[2] harrisons.accessmedicine.com/

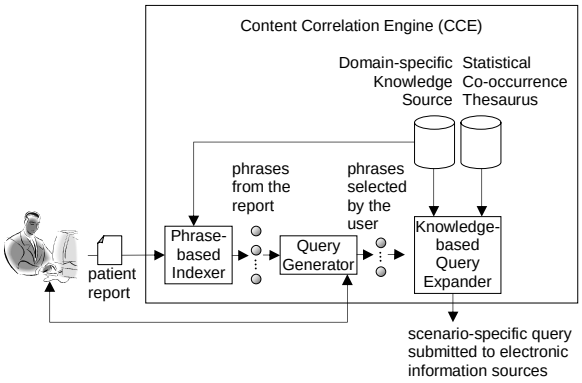

Figure 2: The internal structure of the CCE

cannot serve this purpose, because this method ends up retrieving related documents with all possible topics, e.g general review, treatment-related, diagnose-related, etc. To remedy this shortcoming, we propose a knowledge-based query expansion technique that automatically appends query terms in a specific topic area. As a result, only documents related to the user's application scenario are correlated and retrieved.

The paper is organized as follows. Section 2 presents the internal structure of a CCE which consists of a *Phrase-based Indexer*, a *Query Generator* and a *Knowledge-based Query Expander*. The two techniques of phrase-based indexing and knowledge-based query expansion are discussed in Section 3 and Section 4, respectively. Section 5 provides validation of the proposed techniques. Section 6 discusses related work and Section 7 concludes the paper.

2 CONTENT-CORRELATION ENGINE (CCE)

In this section, we present the internal structure of the CCE. As shown in Figure 2, the user first selects a patient report and inputs it into the CCE. The *Phrase-based Indexer* parses the report and identifies all the phrases in the report. The concepts in each phrase come from a domain-specific knowledge source. The *Query Generator* selects the phrases that are most representative of the original report to form a query. The user can interact with the Query Generator to further narrow down the phrase selection. The *Knowledge-based Query Expander* expands the original query with phrases that belong to a particular topic area, according to the current user's operating scenario, e.g. seeking treatment-related or diagnosis-related information. The expanded query is submitted to electronic information resources to collect relevant medical documents.

3 PHRASE-BASED INDEXING TO RESOLVE VOCABULARY MISMATCH

As we have discussed in the previous section, the concepts in each phrase are defined by a domain-specific knowledge source. In our study, we use the UMLS Metathesaurus [14] as the knowledge source.

In our definition, a phrase is represented as a [concept ID, word stems] pair. To index a medical document using phrases, we scan through the document string, identify each concept that appears in the text and save the concept together with its word stems as one phrase. For example, the phrases detected for a text fragment "60 year old with lung cancer" will be [242379,[3] ("lung," "cancer")].

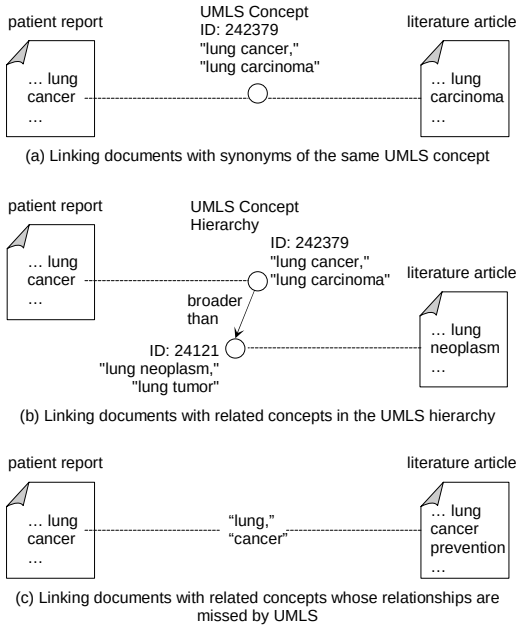

Figure 3: Mapping from concepts to semantic types and the relationships among semantic types

Indexing medical documents using phrases greatly enhances our capability to correlate patient reports with medical literature articles or teaching materials. With simple word stem indexing, a patient report using "lung cancer" cannot be linked to a research article on "lung carcinoma," because they have different string forms. However, these two expressions are synonyms in UMLS and are assigned the same concept ID. As a result, the two documents are linked together in our phrase-based indexing, because they contain phrases with the same concept ID. (Figure 3(a))

Besides grouping synonyms together into one concept, the UMLS Metathesaurus also organizes related concepts in a hierarchy where upper level concepts have broader meanings, and lower level concepts are more specific. For example, "lung cancer" is defined as a broader concept than "lung neoplasm." Using the hierarchical relationships, we are able to link patient reports and research articles whose concepts have "general/specific" or "broader/narrower" relationships. As a result, a patient report using "lung cancer" is linked to an article using "lung neoplasm." (Figure 3(b))

Although UMLS incorporates quite a number of medical thesauri, it is still incomplete regarding the relationships between certain concepts. For example, the concept "lung cancer" is totally unrelated with "lung cancer prevention"[4] in UMLS. Our phrase-based indexing solves this problem by preserving the word stems in each phrase, and linking phrases with each other using both concepts as well as word stems. In this example, although concept 242379 and 281194 are not related in UMLS, we still consider the two phrases [242379,

[3] 242379 is the UMLS concept ID for "lung cancer"
[4] Concept ID 281194

("lung," "cancer")] and [281194, ("lung," "cancer," "prevention")] as related since they share the word stems "lung" and "cancer." (Figure 3(c)) As a result, we can link patient reports with potentially related literature articles, even when they use concepts uncorrelated in UMLS.

4 KNOWLEDGE-BASED SCENARIO-SPECIFIC CONTENT CORRELATION

In this section, we present the technique that generates scenario-specific correlation from phrase-indexed patient reports.

Recent studies reveal that physicians' clinical questions represent different scenarios [1, 13, 7]. For example, in [1, 13], the most frequently asked questions are "What is the cause of symptom X," "How should I manage disease Y," etc. These questions typically consist of two parts: a key concept, c_k, e.g. "symptom X" or "disease Y," and a scenario concept, c_s, e.g. "cause" or "management." Currently reviewed patient reports only provide the key concept c_k, e.g. "lung cancer." Therefore, to correlate a report to online medical documents according to a particular scenario, the physician has to fill in the scenario concept c_s, e.g., "treatment."[5] We can then form a scenario-specific query using the c_k provided by patient reports and c_s indicated by the user, and issue the query to online medical resources. The *Content Correlation* problem is thus converted into *Query Answering*.

In practice, however, directly using scenario concepts like "cause" or "treatment options" to form a query fails to retrieve relevant literature articles, because relevant documents tend to use different scenario concepts. For example, literature articles that discuss "treatment options" for "lung cancer" seldom mention the terms of "treatment" or "option" directly, but rather use "chemotherapy" or "radiotherapy." To solve this mismatch problem, we need to replace the scenario concepts given by the user, with the c_s that are actually used in medical literature. Asking the user to replace the concepts is labor-intensive and time-consuming. Further, a knowledge source, e.g. UMLS, does not indicate whether a concrete concept "chemotheraphy" "treats" "lung cancer."[6] In the following, we propose a knowledge-based query expansion method that solves this problem.

Knowledge-based Query Expansion Our method is leveraged on the UMLS Metathesaurus and the UMLS Semantic Network. As indicated in Figure 4, a group of concepts in the Metathesaurus is abstracted into one semantic type in the Semantic Network. Although UMLS does not specify the potential relationships among the Metathesaurus concepts, it indicates the relationships between semantic types in the Semantic Network level. For example, UMLS does not provide a "treats" link between "radiotherapy" and "lung cancer." Nevertheless, "radiotherapy" belongs to "Therapeutic or Preventive Procedure" which treats "Disease or Syndrome," the semantic type that "lung cancer" belongs to. Using this knowledge structure, we derive the following knowledge-based query expansion method to expand scenario concepts c_s for a given c_k:

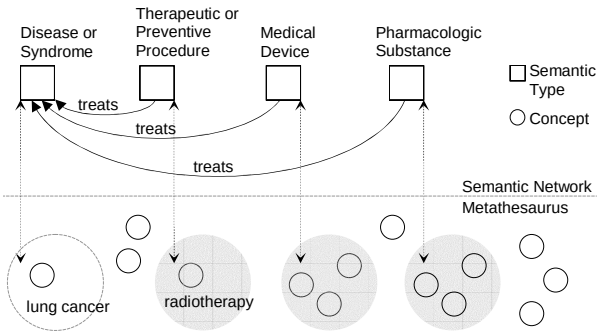

Figure 4: Mapping from concepts to semantic types and the relationships among semantic types

Concept ID	Concept String
34618	Radiotherapy
8838	Cisplatin
3393	Antineoplastic Agents, Combined
9429	Combined Modality Therapy
13216	Chemotherapy
79172	Cranial Irradiation
15133	Etoposide
42679	Vincristine
38903	Surgery, Lung
58928	ECHO protocol

Figure 5: Scenario concepts related to the "treatment" of "lung cancer," automatically derived by our knowledge-based method for query expansion

1. Navigate from c_k to its semantic type (e.g. from "lung cancer" to "Disease or Syndrome").
2. Starting from c_k's semantic type, follow the relationships as indicated by the query's original c_s, and reach a set of relevant semantic types (e.g.starting from "Disease or Syndrome," following "treats" if c_s is "treatment options," and reaching "Therapeutic or Preventive Procedure," "Medical Device," and "Pharmacologic Substance").
3. Include all concepts that belong to the relevant semantic types as candidate c_s. For a sample query "lung cancer, treatment options," we reach the shaded circular areas in Figure 4 as the set of candidate c_s.
4. Assign weights to each derived c_s according to how it co-occurs with c_k in a sample corpus. c_s is assigned a higher weight if it highly co-occurs with c_k. The weights distinguish c_s that are truly semantically related to c_k (since they co-occur more often) from those that are only marginally related. For example, although both concepts belong to "Therapeutic or Preventive Procedure" (the leftmost shaded circle in Figure 4), "radiotherapy" co-occurs with "lung cancer" more often than "heart surgery." As a result, "radiotherapy" receives a much higher weight than "heart surgery" when appended to the query "lung cancer, treatment." Details of weight computation can be found in [15].

Following the above procedure, we can derive the scenario concepts c_s for the query "lung cancer, treatment op-

[5] In our prototype, the physician selects one or more scenario concepts from a pre-compiled scenario concept list

[6] There are certain relationships defined in the latest version of UMLS and SNOWMED, yet they are far from complete

recall	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	Average
precision (stem baseline, no expansion)	0.80	0.66	0.56	0.46	0.41	0.37	0.33	0.28	0.24	0.19	0.08	0.40
precision (statistical expansion)	0.84	0.70	0.62	0.54	0.50	0.46	0.41	0.37	0.32	0.23	0.08	0.46
precision (knowledge-based expansion)	0.92	0.73	0.63	0.58	0.53	0.50	0.45	0.41	0.36	0.28	0.12	0.50

Figure 6: Retrieval effectiveness, knowledge-based expansion vs. statistical expansion

Query ID	Original Query Form
14	**PANCYTOPENIA IN AIDS,** *workup and etiology*
15	**THROMBOCYTOSIS,** *treatment and diagnosis*
64	*prevention, risk factors, pathophysiology of* **HYPOTHERMIA**

Figure 7: Sample OHSUMED queries. Key concepts are shown in capital letters and scenario concepts are in italics

Document ID	Document Title
90313852	Optic disk elevation in Down's syndrome
88131226	Ischemic optic neuropathy in sickle cell disease
90075918	Small cell lung cancer
91069110	Radiotherapy for lung cancer
88191922	Optic nerve swelling secondary to the obstructive sleep apnea syndrome

(a) Stem-based document retrieval

Document ID	Document Title
90172011	Ten-year survival of patients with small-cell lung cancer treated with combination chemotherapy with or without irradiation
91132695	Preoperative chemotherapy (cisplatin and fluorouracil) and radiation therapy in stage III non-small-cell lung cancer: a phase II study of the Lung Cancer Study Group
90075918	Small cell lung cancer
87320019	Preoperative and adjuvant chemotherapy in locally advanced non-small cell lung cancer
89289246	How should thoracic radiotherapy be given in limited small cell lung cancer?

(b) Knowledge-based query expansion

Figure 8: Comparing top 5 documents for query "lung cancer, *treatment options*," retrieved by the stem-based retrieval method and the knowledge-based query expansion method

tions." We list ten such c_s with the highest weights in Figure 5. By appending the original query with these c_s, we obtain a query that matches better with online medical resources. Issuing this query to online resources results in scenario-specific (i.e. "treatment" scenario) correlation from a patient report on "lung cancer" to related literature articles.

5 VALIDATION OF SCENARIO-SPECIFIC CONTENT CORRELATION

Experimental Setup To illustrate how we achieve scenario-specific content correlation using knowledge-based query expansion, we evaluate our technique on a standard testset OHSUMED [16]. The testset consists of all MEDLINE bibliography records from 1988 to 1992. It also provides a collection of medical queries and all the experts' judgements indicating which documents are relevant to each query. Each query is generated from a real health care scenario and contains two sections. The first section, a patient description, describes the real situation of a clinical patient regarding which the query is asked. The second section, an information request, explicitly states the physician's information needs in treating this patient. Therefore, each query in this testset can be considered

Document ID	Document Title
89100938	Prospective evaluation of fi ne needle aspiration in the diagnosis of lung cancer
87070550	Bronchial brushing and bronchial biopsy: comparison of diagnostic accuracy and cell typing reliability in lung cancer
91335375	Value of washings and brushings at fi breoptic bronchoscopy in the diagnosis of lung cancer
90247220	Pitfalls in the radiologic diagnosis of lung cancer
87320022	Diagnostic and therapeutic uses of pleuroscopy (thoracoscopy) in lung cancer

Figure 9: Top 5 documents retrieved from OHSUMED for query "lung cancer, *diagnostic options*," using the knowledge-based query expansion method

as a miniature of a real patient report, and can be used to validate our technique that correlates patient reports with medical literature articles.

We focus our study on 41 OHSUMED queries that are scenario-specific, i.e. inquiring about the "treatment," "diagnosis," "prevention" and "cause" of a particular disease. Some of the sample queries are shown in Figure 7. For each query, we use the method in Section 4 to identify semantically related concepts and append them into the original query.

We index each of the 41 queries and all the OHSUMED documents using phrase-based indexing. Query-document similarity is computed using the standard Vector Space Model (VSM) [17]. The most similar documents to each query are retrieved.

Evaluation Metrics and Baseline We use the standard precision-recall measurement to compare our method with traditional ones. When a certain number of documents are retrieved, *precision* is the percentage of the retrieved documents that are relevant, and *recall* is the percentage of the relevant documents that have been retrieved so far. We evaluate the retrieval accuracy by interpolating the precision values at eleven recall points.

In this experiment, we compare with two existing information retrieval methods. In the first method, stem-based document retrieval, both queries and documents are indexed by stems and documents are retrieved using the standard VSM. No expansion is made to the queries.

The second method, statistical-query-expansion-based document retrieval [18, 19, 20], appends all terms that statistically co-occur with the original query terms in a sample corpus, e.g. OHSUMED. No knowledge source is consulted to filter out terms that are not pertinent to the query topic. For example, this method may append a diagnose-related term into the query "lung cancer, treatment options," simply because this diagnose term highly co-occurs with "lung cancer" in a sample corpus.

Results and Discussions In this experiment, we focus on scenario-specific queries that inquire about the "diagnosis" or "treatment" of a disease. Our method yields the high-

est retrieval precision, which means that we are able to retrieve more documents relevant to the "diagnosis" or "treatment" of the queried disease. Figure 6 shows the average 11-point precision-recall over the 41 queries studied, generated by our knowledge-based expansion method and the two existing methods. In terms of average retrieval precision, our method improves over the stem baseline by 25%, and improves over the statistical expansion method by 8.7%. As argued in [21], if a new retrieval method improves the precision-recall by more than 5% over traditional methods, on the basis of no less than 25 queries, then the new method is preferable for future usage. Therefore, our approach is significantly more effective in terms of scenario-specific content correlation. To better understand the improvement, we retrieve the top 5 documents from OHSUMED for query "lung cancer, treatment options," using the stem-based retrieval and our knowledge-based expansion (Figure 8).[7] Comparing Figure 8(b) with Figure 8(a), the knowledge-based query expansion method retrieves more relevant documents to the query. Also, to illustrate how the retrieved documents vary according to different scenarios, we use the knowledge-based method to select the top 5 documents for "lung cancer, diagnostic options" (Figure 9). Comparing Figure 9 with Figure 8(b), it can be seen that our method yields scenario-specific correlation, e.g. treatment or diagnosis.

6 RELATED WORK

Content Correlation is related to the research of mediating distributed and heterogeneous databases and searching for similar answers for a given query [22, 23, 24]. These works mainly focused on structured data sources. In this paper, we have studied the interoperability among highly unstructured data sources of free text documents.

Past research on Content Correlation largely focused on applying statistical or syntactical information to link documents [8, 25]. No scenario-specific correlation has been provided. In this paper we propose a knowledge-based approach to correlate data contents in various textual information sources. Experimental results reveal that we can provide effective scenario-specific content correlation.

7 CONCLUSION

In this paper, we present a new knowledge-based approach for scenario-specific content correlation. The phrase-based indexing technique helps link documents from different collections that use different concept expressions. The knowledge-based query expansion technique formulates the phrase-indexed patient reports into scenario-specific queries. As a result, the Content Correlation Engine is able to search for relevant documents specifically related to the user's current scenario.

REFERENCES

[1] 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):211–220, 1999.

[2] D.G. Covell, G.C. Uman, and P.R. Manning. Information needs in office practice: are they being met? *Ann Intern Med*, 103:596–599, 1985.

[3] P.N. Gorman and M. Helfand. Information seeking in primary care: How physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making*, 15(2):113–119, 1995.

[4] W.R. Hersh, J. Pentecost, and D.H. Hickam. A task-oriented approach to information retrieval evaluation. *JASIS*, 47(1):50–56, 1996.

[5] J. Marshall. The continuation of end-user on line searching by health professionals: preliminary survey results. In *Proceedings of the Medical Library Association Annual Meeting*, 1990.

[6] P.N. Gorman, J. Ash, and L. Wykoff. Can primary care physicians' questions be answered using the medical literature? *Bull Med Lib Assoc*, 82:140–146, 1994.

[7] R. Haynes, K. McKibbon, C. Walker, N. Ryan, D. Fitzgerald, and M. Ramsden. Online access to medline in clinical settings. *Ann Intern Med*, 112:78–84, 1990.

[8] W.W. Chu, D.B. Johnson, and H. Kangarloo. A medical digital library to support scenario and user-tailored information retrieval. *IEEE Tran. on Information Technology in Biomedicine*, 4(2):97–107, 2000.

[9] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO '97*, 1997.

[10] R. Richardson and A.F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. In *Proceedings of* 17^{th} *BCS-IRSG*, 1995.

[11] E.M. Voorhees. Using WordNet to disambiguate word sense for text retrieval. In *Proceedings of ACM SIGIR '93*, 1993.

[12] W. Mao and W.W. Chu. Free-text medical document retrieval via phrase-based vector space model. In *Proceedings of AMIA Annual Symp 2002*, 2002.

[13] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri. A taxonomy of generic clinical questions: classification study. *BMJ*, 321(12):429–432, 2000.

[14] National Library of Medicine. *UMLS Knowledge Sources*. 12^{th} edition, 2001.

[15] V.Z. Liu and W.W. Chu. Expanding queries with semantically related terms derived from knowledge sources. Technical report, Computer Science Department, UCLA, 2003.

[16] W. Hersh, C. Buckley, T.J. Leone, and D. Hickam. Ohsumed: an insteractive retrival evaluation and new large test collection for research. In *Proceedings of ACM SIGIR '94*, 1994.

[17] G. Salton and M.J. McGill. *Introduction to Mordern Information Retrieval*. McGraw Hill, 1983.

[18] Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO '94*, 1994.

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

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

[21] C. Buckley and E.M. Voorhees. Evaluating evaluation measure stabililty. In *Proceedings of ACM SIGIR 2000*, 2000.

[22] W.W. Chu, A.F. Cardenas, and R.K. Taira. KMeD: A knowledge-based multimedia medical distributed database system. *Inform. Syst.*, 10(2):75–96, 1995.

[23] W.W. Chu, H. Yang, K. Chiang, M. Minock, G. Chow, and C. Larson. CoBase: A scalable and extensible cooperative information system. *J. Intell. Inform. Syst.*, 6(11), 1996.

[24] S. Melnik, H. Garcia-Molina, and A. Paepcke. A mediation infrastructure for digital library services. In *Proceedings of ACM DL 2000*, 2000.

[25] S.L. Price, W.R. Hersh, D.D. Olson, and P.J. Embi. Smartquery: Context-sensitive links to medical knowledge sources from the electronic patient record. In *Proceedings of AMIA Annual Symp 2002*, 2002.

[7] The stem-based method retrieves quite a few articles on "optic"-related topics, because the term "optic" in the documents and the term "options" in the query share the same word stem "opt"