A Query-based Medical Information Summarization System Using Ontology Knowledge

Ping Chen
Computer and Mathematical Sciences Department
University of Houston-Downtown
1 Main St., Houston, TX 77002
chenp@uhd.edu

Rakesh Verma
Department of Computer Science
University of Houston
4800 Calhoun, Houston, TX 77004
rverma@uh.edu

Abstract

As huge amounts of knowledge are created rapidly, effective information access becomes an important issue. Especially for critical domains, such as medical and financial areas, efficient retrieval of concise and relevant information is highly desired. In this paper we propose a new user query based text summarization technique that makes use of Unified Medical Language System, an ontology knowledge source from National Library of Medicine. We compare our method with keyword-only approach, and our ontology-based method performs clearly better. Our method also shows potential to be used in other information retrieval areas.

KEYWORDS: Biomedical Ontology, UMLS, Text Summarization, Information Retrieval

1 Introduction

Information plays a key role in our society. As huge amounts of knowledge are created and available through WWW, how to efficiently and effectively distribute and access these valuable data becomes critical. A general Web search engine tries to serve as an information access agent. It retrieves and ranks information according to a user's query, and it already makes a huge impact on how we search and organize information. But current search engines only perform shallow string processing due to the lack of deep understanding of natural languages and human intelligence, and users usually have to go through pages before they find something useful or give up. It may not matter much if a user needs information about a pair of shoes, but it will be a serious problem for crucial tasks, such as in medical or financial domains. To solve this problem we are building a document summarization system specialized for medical domain, which will retrieve and summarize up-to-date medical information from trustworthy online sources according to users' queries.

A concise summary will improve productivity since not all documents come with an abstract or summary. Even if some documents do provide abstracts, these abstracts are written by authors to summarize the "main" ideas of an article. However, real world information retrieval often starts with a user's query, which is a set of keywords. These keywords may not match the main ideas of a document. In this case the author written abstracts will not be a good summary for a particular query. Hence, summaries that a user wants need to be generated on the fly based on his query keywords. It is impossible for static author-written abstracts to satisfy such dynamic requirements.

1.1 Related work

There are three existing projects focusing on medical information retrieval, summarization, and management.

1. PERSIVAL

   PERSIVAL is designed to provide personalized access to a distributed patient care digital library. PERSIVAL supports search and summarization of online multimedia information from clinical records to both patients and healthcare providers [10]. PERSIVAL uses context to help the user formulate meaningful queries and extract important information from the clinical record. And it uses patient information to rerank articles, and uses segmentation and domain knowledge to summarize echocardiogram video.

2. HelpfulMed

   HelpfulMed [2] provides access to medical information on the Internet and in medical-related databases for professional and advanced users. Users can locate medical information by extracting noun phrases and

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

determining relationships with other medical terminology through concept-based search support [7].

3. QCS

   QCS indexes documents, retrieves documents relevant to a query, clusters the subset of retrieved documents, and produces a single summary for each of the clusters [3]. QCS system has been evaluated with some news corpus. It is not specially designed for a medical domain, and no ontology knowledge is used in the system.

Although these systems improve the efficiency of medical information access by automatic collection and analysis of medical information, summarization and ranking through limited utilization of ontology, but they fail to take full advantage of existing rich ontology knowledge, such as Unified Medical Language System available from National Library of Medicine. Due to the huge numbers of terms and concepts used in medical domain, analysis of terms and their relationships is key to improve the medical information system performance as shown by [6].

In this paper, we discuss background information about medical ontology knowledge source in Section 2. Our summarization system architecture and algorithm are presented in Section 3. We give detailed process of summarization of one sample medical research paper in Section 4. In Section 5 we conclude and discuss our future work.

2 Medical ontology knowledge

Huge numbers of terms are used in medical domain. To interpret a medical document, understanding of these term and their relationships is very important. An ontology is a description of the concepts and relationships. High-quality ontology knowledge is the key to improve the quality of medical information retrieval and management. In this paper we use Unified Medical Language System (UMLS) from National Library of Medicine (NLM) as our main medical ontology knowledge base.

UMLS is designed to help a medical information system "understand" the meanings of the concepts and terms and their relationships in biomedicine and health domain [13]. The UMLS Knowledge Sources are multi-purpose, and it can be used to create, process, retrieve, integrate, and/or aggregate biomedical and health data and information. UMLS divides medical ontology knowledge into three sources: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. SPECIALIST lexicon is designed to provide the lexical information for the SPECIALIST Natural Language Processing System. In our current system we use the Metathesaurus and Semantic network since our focus is on semantic analysis of a medical document. Here is a brief overview about them.

The Metathesaurus is a multi-lingual vocabulary database that contains definitions of biomedical terms, their various names (such as synonyms and abbreviations), and the relationships among them.

The Semantic Network categorizes all concepts contained in the Metathesaurus into organisms, anatomical structures, biological function, chemicals, events, physical objects, and concepts or ideas. The Semantic Network also defines a set of relationships between these concepts. These relationships provide the structure for the network. The primary relationship is the "isa" link, which establishes the hierarchy of types within the Network. There is also a set of non-hierarchical relationships, such as, "physically related to", "spatially related to", "temporally related to", "functionally related to" and "conceptually related to". Here are a few examples,

- C0002871|CHD|C0002891|isa|MSH|MSH||

  Anemia, Neonatal (C0002891) has "CHILD REL" and "isa REL" to Anemia (C0002871)

- C0002871|RB|C0221016||MTH|MTH||

  Red blood cell disorder, NOS (C0221016) has "broader REL" to Anemia (C0002871)

- C0002871|RL|C0002886|mapped to|SNMI|SNMI||

  Anemia, Macrocytic (C0002886) has "like" relationship to Anemia (C0002871)

- C0002871|RO|C0002886|clinically associated with|CCPSS|CCPSS||

  Megaloblastic anemia due to folate deficiency, NOS (C0151482) has "clinically associated with" relationship to Anemia (C0002871)

We use two primary Metathesaurus relationship files, MRREL.RRF and MRCONSO.RRF. MRREL.RRF contains the "distance 1" hierarchical relationships, i.e., immediate parents, immediate child, and immediate sibling relationships, as well as other types of intra-source relationships. MRCONSO.RRF contains medical concept names, their identifiers and key characteristics.

3 A Dynamic Summarization System

3.1 System architecture

Figure 1 shows the architecture for our medical information summarization system. Following is an overview of our approach:

1. Select trustworthy online medical repositories, such as FDA, NLM and other online medical document repositories.

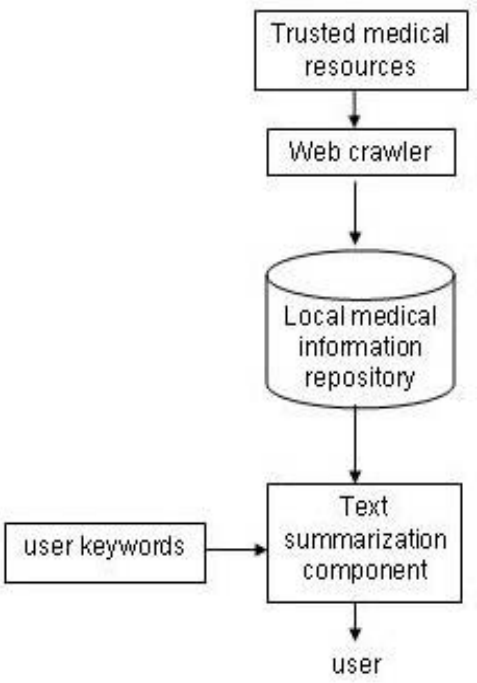

Figure 1. Medical Information Summarization System Architecture

2. Run a web crawler regularly on the selected repositories to extract documents, which are saved in a local document repository.

3. Perform preprocessing tasks on the local repository, data cleaning and indexing based on ontology knowledge obtained from the UMLS Metathesaurus.

4. When a user inputs a query and wants to retrieve relevant documents, provide a query driven summary to reduce redundancy and assist the user in rapidly locating the essential information.

3.2 Summarization Algorithm

Text summarization has become an important area in text mining and generated a lot of research interest recently. There are two types of approaches [8]:

1. The knowledge-based approaches build a semantic representation for the summarization task, such as a set of logical forms [11], using ontology knowledge [5], or a template describing some key concept [1], etc.

2. The surface features-based approaches select summary material from the source based on position information, specific terms or cue phrases [4, 12].

Our summarization technique is knowledge-rich and user query-based. We represent the original document with a semantically connected concept network. We will choose a subset of sentences from the original document as its summary. Our approach is totally term-based, i.e., we are going to recognize and process only terms defined in UMLS and ignore all other words. Here is the summarization procedure:

1. Revise the query with UMLS ontology knowledge. We will add relevant keywords, delete redundant keywords. We return the revised query and let the user finalize it.

2. Calculate distance of each sentence in the document to the finalized query. Distance function used will be metrics (satisfying $d(x,x) = 0$, symmetry, and triangle inequality). If the distance is smaller than a threshold, the sentence will be a candidate to be included in the summary.

3. Calculate pair-wise distances among the candidate sentences (metrics can reduce the number of computations required). Then, divide candidate sentences into groups based on a threshold and select highest-ranked one from each group.

Summarization evaluation is an important and hard topic [9]. In our case study, for evaluation purpose we choose an article containing both an abstract and a set of keywords. Since both the abstract and keywords are written by authors, it is reasonable to assume that both of them reflect the main ideas discussed in the article. Hence if we use the keywords as user query words, a perfect summarization method should choose all and only the sentences in the abstract as the summary. In our case study we found that some sentences in abstract are very similar or even totally identical to some sentences in the article. Since our summarization method can calculate distance among candidate summary sentences, we will choose only one from these close sentences.

4 Experiment

In this section we discuss detailed processing of one medical research paper using our summarization method.

4.1 Document preprocessing

We downloaded the paper titled "Aqueous levels of macrophage migration inhibitory factor and monocyte chemotactic protein-1 in patients with diabetic retinopathy" at www.blackwell-synergy.com/toc/dme/21/12. The

paper provides four keywords, "aqueous humor", "diabetic retinopathy", "macrophage migration inhibitory factor", and "monocyte chemotactic protein-1". We use them as the original query keywords in our experiment. We performed some basic cleaning steps, such as removal of references and special characters. Stemming is not performed since with the use of UMLS we are able to find most variances of terms. We did not remove stop words since our method is medical term-based and ignores any words not included in UMLS. The whole paper is divided into 118 sentences. Sentence 1 is the paper title. Sentence 2 is the author information. Sentences 3 to 13 belong to the abstract in the original paper. Sentences 14 to 114 are regular content, and 115 to 118 are conclusion. Combining abstract and conclusion together, we have totally 15 sentences in the original summary.

In our experiment we choose a set of sentences as summary of the article. The summary is evaluated by how many generated summary sentences are from the original abstract and conclusion.

4.2 Keyword expansion

The four original keywords are expanded using MRCONSO.rrf and MRREL.rrf files from UMLS Metathesaurus. The expanded set totally includes 556 keywords, and can be classified as:

1. Simple variants of the original keywords, such as "Humor aqueous", "Aqueous Humors", "Diabetic Retinopathies".
2. Abbreviations, such as "MCP1" or "MCP-1" for "monocyte chemotactic protein-1", "MIF" for "Migration-Inhibitory Factor".
3. Semantically similar or medically related terms, such as "Monocyte Chemoattractant Protein 2", "Monocyte Chemoattractant Protein 3", "Monocyte Chemoattractant Protein 4", "metabolic aspects", "Aqueous humor of eyeball", "blood", "MCP1 protein, human", "Monocyte-Derived Macrophages", "monocyte chemoattractant protein 1 receptor", etc.

4.3 Selecting sentences for summary

With the expanded keyword set we perform a simple string matching step for the paper, and count the number of matched keywords in each sentence. When we determine which sentences will be included in the summary, for comparison reason we generate three different "scores" for each sentence:

1. Simply count the number of matched original keywords and select the sentences with many matching keywords.
2. If a sentence contains an original keyword, assign weight 1 to it. If a sentence contains an expanded keyword, assign weight 0.5 to this keyword. Add all the weights together, and we can get a score for each sentence. We select sentences with high scores.
3. This approach adds one step to the approach 2. After we get a score for each sentence, we will normalize the score with the length of the sentence. And we select sentences with high normalized scores.

4.4 Evaluation

Table 1 gives the experiment results using precision and recall measures. Precision is defined as,

$$P = \frac{a}{b} \tag{1}$$

where a is the number of matched sentences in the generated and original summaries, and b is the number of sentences in the generated summary.

Recall is defined as,

$$R = \frac{a}{c} \tag{2}$$

where a is the number of matched sentences in the generated and original summaries, and c is the number of sentences in the original summary.

With expanded keywords, it is very obvious that we can generate a better summary. Although normalized weight approach seems more reasonable, it performs poorly. We examined the original abstract and conclusion and found that the sentences in the original summary are very long, generally with 40 to 50 words. So many short non-summary sentences are chosen even with fewer matching keywords.

5 Conclusion and future work

In this paper we present our on-going work on user query-based summarization of medical literature. Ontology knowledge is proven to be an effective way to go beyond the mere keyword-based information retrieval methods. With our experiment, we feel that ontology knowledge can be further utilized in other fields of broad information management and knowledge discovery process. Our future work includes:

1. Conduct more experiments with our summarization method.
2. Use thresholds for selecting sentences in the summary using statistical data of sentences in the abstract when it is available.

Method	Score	Number of sentences in generated summary	Number of matched sentences	Precision	Recall
Original keywords	≥ 2	20	6	0.30	0.40
	≥ 1	66	13	0.20	0.87
Expanded keywords	≥ 3	16	10	0.63	0.67
	≥ 2.5	20	12	0.60	0.80
Normalized weight	≥ 0.1	19	6	0.32	0.40
	≥ 0.09	25	8	0.32	0.53

Table 1. Comparison of three summarization approaches

3. Utilize some natural language processing techniques in our method, such as parsing and syntax analysis.

4. Index and organize generated summaries for future access and reuse.

5. Expand to multi-document summarization.

6. Integrate our summarization component into a broad medical information retrieval system, which may include document clustering, ranking and other components.

6 Acknowledgment

This work was supported by NSF grants 0306475 and 0313880. We thank NLM for providing the UMLS DVD used in our system implementation.

References

[1] R. Barzilay, N. Elhadad, and K. McKeown. Inferring strategies for sentence ordering in multi-document news summarization. Journal of Artificial Intelligence Research, 17:3 8, 2002

[2] H. Chen, A. Lally, B. Zhu and M. Chau, HelpfulMed: Intelligent Searching for Medical Information over the Internet, Journal of the American Society for Information Science and Technology (JASIST), pages 683-694, Volume 54, Issue 7, May 2003.

[3] D. Dunlavy, J. Conroy, and D. O'Leary, QCS: A Tool for Querying, Clustering, and Summarizing Documents, Proceedings of the HLT-NAACL Conference, Edmonton, Canada, 2003.

[4] N. Elhadad, K. McKeown. Towards generating patient specific summaries of medical articles. Workshop on Automatic Summarization, NAACL, Pittsburgh, USA, 2001

[5] M. Fiszman, T. Rindflesch, H. Kilicoglu. Abstraction Summarization for Managing the Biomedical Research Literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004.

[6] M. Fiszman, T. Rindflesch, H. Kilicoglu. Summarization of an Online Medical Encyclopedia. MedInfo. 2004 Sept.;2004: 506-510.

[7] G. Leroy, H. Chen, Meeting Medical Terminology Needs - the ontology-enhanced medical concept mapper. IEEE Transactions on Information Technology in Biomedicine 5(4): pp. 261-270, 2001.

[8] M. Maybury, I. Mani, Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999

[9] Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. 1999. The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of the Ninth Conference on European Chapter of the Association For Computational Linguistics (Bergen, Norway, June 08 - 12, 1999). European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 77-85.

[10] K. McKeown, S Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. Jordan, J. Klavans, A. Kushniruk, V. Patel, and S. Teufel. PERSIVAL: A System for Personalized Search and Summarization over Multimedia Healthcare Information. ACM/IEEE Joint Conference On Digital Libraries, Roanoke, VA, 2001, 331–340.

[11] T. Mori, M. Nozawa, and Y. Asada. Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine. In Proceedings of the 20th International Conference on Computational Linguistics, pp. 439–445, 2004.

[12] Y. Sun, S. Park, Generation of Non-redundant Summary Based on Sum of Similarity and Semantic Anal-

ysis, Asia Information Retrieval Symposium, Beijing, China, 2004

[13] Unified Medical Language System, available at www.nlm.nih.gov/research/umls/