An Information Retrieval System for Medical Records & Documents

Article *in* Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference · February 2008

DOI: 10.1109/IEMBS.2008.4649446 · Source: PubMed

5 authors, including:

Shihchieh Chou
National Central University
16 PUBLICATIONS **229** CITATIONS

SEE PROFILE

Chin-Yi Cheng
The University of Tokyo
6 PUBLICATIONS **17** CITATIONS

SEE PROFILE

An Information Retrieval System for Medical Records & Documents

Shihchieh Chou, Weiping Chang, Chin-Yi Cheng, Jihn-Chang Jehng and Chenchao Chang

Abstract—**The forms of the medical records are different from one institute to another. Moreover, medical records are always stored in free text. Consequently, medical records almost can not be logically analyzed and understood by machines. In this paper, we have applied the information retrieval (IR) technique in the using of medical records. We have implemented an IR system for the users, such as doctors and patients, to query similar or related medical records to support diagnosis and treatment. Knowledge retrieval for reuse is the key idea of this system.**

I. Introduction

THE growth of information technology has changed the whole life of human including, of course, medical and healthcare behaviors. The application of information technology in medicine and healthcare includes hospital information system (HIS), electronic medical record (EMR), electronic patient record (EPR), medical diagnosis systems, medical image systems, and so on. Although the functions of the systems mentioned above are very different, the main purposes of improving the efficiency and effectiveness in medical behaviors are the same [1].

In the medical diagnosis system, the focus is on the assistance of diagnosis and treatment. It usually related to the techniques of artificial intelligence, such as expert system [2], fuzzy [3], natural languages [4], and neural network [5][1]. According to the methodology of inference, medical diagnosis systems can be divided into rule or similarity based. Both kinds of systems need to apply the medical data existed in HIS, EMR or EPR in the generation of rules or retrieval of similar medical cases. However, there exist problems in the using of the present data for building medical diagnosis systems. First, medical records are stored by different EMR or EPR systems owned by different medical institutes and the forms of the medical records are different from one institute to another. Hence, application of the medical records owned by different medical institutes is not easy. Second, medical records are always stored in free text [6][7]. Consequently, medical records almost can't be logically analyzed and understood by machines [7].

Our proposition for tackling this data using situation is the application of information retrieval (IR) technique which can retrieve similar or related data in free text format from different sites. Through the retrieving of similar or related medical records, the knowledge existed in the medical records could be reused by the doctor or the patient. To implement this proposition, this research has aimed to develop an IR system to demonstrate the applicability of IR techniques in the support of medical diagnosis and treatment. In the system, a self-developed relevance feedback technique and genetic algorithm (GA) will be applied to enhance retrieval effectiveness.

II. Related work

Human diagnostic procedure encompasses the phases of hypothesis formation and evidence gathering [8]. The patient may complain some symptoms first, then, the doctor makes some tests on the patient according to the symptoms, and finally, uses the result and his medical knowledge to make the judgment and offer the treatment [7]. The early medical diagnosis support system well known as MYCIN had its fundamentals based on human diagnostic procedure. MYCIN is a rule based expert system for diagnosis. In the using of the system, the user must input the symptoms of the patient, then, the system will infer the diseases of the patient according to the symptoms and the rules built-in the system. Development of this kind of system is expensive because it requires close interaction and cooperation of the domain expert and the knowledge engineer in the generation of rules. The other shortcoming of this system is its narrowing in the identification of the specific disease.

Nowadays, the development of diagnosis support systems has been trying to generate rules from among the present medical data existed in HIS, EMR or EPR. The techniques like data mining, fuzzy [3] and rough set [2] have been used to analyze the medical records and medical images for extracting rules or relations. These kinds of systems are thought cheaper because it need fewer work than the previous type of expert systems. Also, by the using of the present data and knowledge in hospital, these systems can be kept up-to-date.

Case based reasoning (CBR) [9] which is based on similarity comparison between data records is another type of diagnosis support system. It compares the attributes of cases to determine the similarity between cases. Using this system,

Manuscript received April 7, 2008. This work was supported in part by the National Science Council, Taiwan, under the Grant No. NSC96-2416-H-008-011-MY2.

S. Chou is with the Department of Information Management, National Central University, No.300, Jhongda Rd., Jhongli City, 320,Taiwan (phone: +886-3-4267262; fax:+886-3-4254604; e-mail: scchou@mgt.ncu.edu.tw).

W. Chang is with the Department of Student Affairs, Central Police University, Taiwan (e-mail:una024@mail.cpu.edu.tw).

C. Y. Cheng is with the Department of Information Management, National Central University, Taiwan (e-mail: yoda0612@seed.net.tw).

J. C. Jehng is with the Institute of Human Resource Management, National Central University, Taiwan (jehng@mgt.ncu.edu.tw).

C. Chang is with the Department of Information Management, National Central University, Taiwan (e-mail: scchou@mgt.ncu.edu.tw).

the user can find similar cases for reference by the input of a new case. This is helpful to both doctors and patients. In medicine, CBR systems can be classified according to the purposes as diagnostic systems, classification systems, tutoring systems, planning systems [10][11]. Many CBR type diagnosis support systems have been proposed in recent year. For example, Heindl et al. had developed ICONS which is an antibiotics therapy advice system used in intensive care unit [12][13]; Bakaa et al. had developed FM-Ultranet which is a detecting system of malformations and abnormalities of foetus [14]; Perner had proposed a CBR system to detect Alzheimer disease in CT images [15].

The rule based or similarity based systems require the users to input clear facts or attributes, such as symptoms of patient, to systems for determining the results. This is not easy for the user without medical background. Another disadvantage of these systems is the requirement of clear definition for data attributes. Hence, data transforming before rule extraction and similarity comparison becomes a heavy load of the system [16].

IR is a well known technique applied to text based data. The typical method of IR is to calculate term frequencies in documents for further applications, such as the determining of the similarity between documents or the clustering of documents. Researches in recent years have combined concept model for better semantic representation of the document [6][17]. In the past, many researches have applied the technique of IR to medical documents. For example, Li and Wu [18] present a key phrase identification program (KIP) for identifying topical concepts from medical documents; Chu and Cesnik [6] use conceptual graphs to capture the structure and semantic information contained in the medical documents; Mao and Chu [19] proposed a phrase-vector space model for medical documents retrieving. In our proposition of using IR techniques for diagnosis support, we will treat the medical records as web documents on different servers. Thus, IR system could be used as an efficient and effective tool in the accessing of similar or related medical cases. The free data format requirement and the similarity comparison approach are the two major advantages that make IR techniques applicable to the construction of diagnosis support system in today's medical records situation.

III. METHOD

We have developed an IR system aimed to retrieve similar cases or related documents from whatever data servers to support diagnosis. In the system, the cases or documents are treated as web documents. Figure 1 presents the framework of the system. In the framework, the component of Information Collector is used to collect the user's input and relevance feedback. The two components, Learning Agent and User Profile, are used to learn and maintain the user's searching focus. The two components, Query String Generator and Document Selector, are used to retrieve the web document most relevant to the user's searching focus. The component of Spider is used to collect web documents from servers. The details of each component are described as follows.

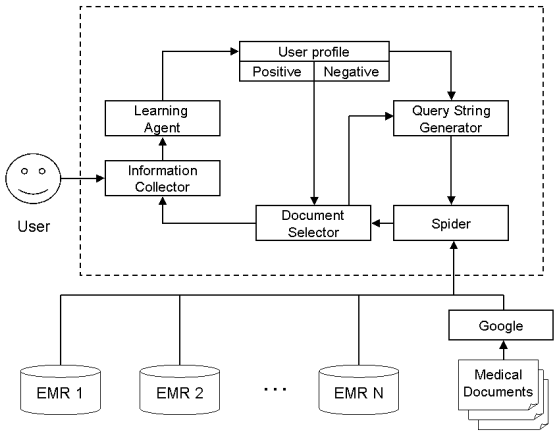

Fig.1. System Framework

Information Collector: This component is used as a user interface for capturing the input of the example document and the user's rating of relevance/non-relevance for the retrieved document. The user's rating for the retrieved document can be very relevant, relevant, not sure, non-relevant and very non-relevant. Except data input, this component also filters out useless information like punctuation and stop words.

User Profile: The main purpose of this component is to maintain the user's searching focus that has been learned by the component of Learning Agent. It encompasses two sub-profiles, the positive and the negative user profiles, to keep the system's learning of the user's interest/disinterest respectively. The positive user profile is a database table with the following four attributes: 'Term', 'Frequency', 'Sensitivity' and 'Adjusted Frequency'. The negative user profile is a database table with the following two attributes: 'N-Term' and 'N-Frequency'.

Learning Agent: This component is used to learn the user's interest/disinterest from his querying of the medical documents. The learning has been based on the user's input, including the initial documents, the retrieved documents and the user's ratings of relevance/non-relevance for the documents. With the input as mentioned, the learning agent will maintain the values for the positive and the negative user profiles to represent the user's searching focus.

In the positive user profile, the values for the attribute 'Term' are the terms that appear in the documents rated by the user as relevant/very relevant. The value for the attribute 'Frequency' by each term is the weighted term frequency of the term belonging to the documents rated as relevant/very relevant. The weighting value has been set as 1 for relevant and 1.2 for very relevant after pretests. The value for the attribute 'Sensitivity' by each term is a constant set to weight the value of the attribute 'Frequency' by the term. It is given according to the classification of the term on term appearance situations. Two term appearance situations have been identified: (1) a term can appear in relevant documents only and never appear in non-relevant documents; (2) a term can appear both in relevant and non-relevant documents. Again, after pretest, the

sensitivity value 1 has been given to terms belonging to situation (1) and the sensitivity value 1.2 has been given to terms belonging to situation (2).The value for the attribute 'Adjusted Frequency' by each term is derived by weighting the value of the attribute 'Frequency' by the value of sensitivity. After the set of the values for the positive user profile, the rows of the positive user profile can be sorted by the values of 'Frequency' or 'Adjusted Frequency' in descending order. A certain number of top ranked terms together with the values of 'Frequency' or 'Adjusted Frequency' are then used to represent the user's interest.

In the negative user profile, the values for the attribute 'N-Term' are the terms that appear only in the documents rated by the user as non-relevant/very non-relevant. The value for the attribute 'N-Frequency' by each term is the weighted term frequency of the term. The weighting value has been set as 1 for non-relevant and 1.2 for very non-relevant after pretests. After the set of the values for the negative user profile, the rows of the negative user profile can be sorted by the values of 'N-Frequency' in descending order. A certain number of top ranked terms are then used to represent the user's disinterest.

Query String Generator: The main purpose of this component is to generate a query string from the terms kept in the user profile and use the query string to retrieve web documents. Genetic algorithm (GA) has been applied to optimize the using of the terms in the forming of the query string. The purpose of the application of GA is to make the system perform well and stably. In our application of GA, each chromosome is represented by N bits. Among the N bits, M bits (M < N) are used to represent positive keywords connected by AND operator and M-N bits are used to represent negative key words connected by NOT operator. The positive keywords are selected from the top M ranked terms in the positive user profile. The negative keywords are selected from the top M-N ranked terms in the negative user profile. In the chromosome, the value of the bit (1 or 0) represents the selection or not selection of one keyword.

Document Selector: The main purpose of this component is to determine the matching of the retrieved document with the user's searching focus represented by the user profile and to cooperate with GA in the Query String Generator to retrieve the most relevant documents to the user. It will compare the similarity between the retrieved document with the positive user profile first, then pass the similarity value to GA in the Query String Generator, and finally generate K most relevant documents to the user.

Spider: the function of this component is to retrieve the medical cases or documents from whatever servers providing the data.

IV. EVALUATION

We have conducted some formal tests to study the system. Since the system is still on development, it is not appropriate at this stage to cooperate with hospital to evaluate the performance of the system on querying medical records which are confidential. Considering the requirement of free text format, the web pages on different servers have been used as the data source for performance evaluation. Before the performance test, we have done some pretests first to detect the appropriate values for the system variables. Table I summarizes the result.

TABLE I
APPROPRIATE VALUES FOR SYSTEM VARIABLES

SYSTEM VARIABLE	INITIAL VALUE
Amount of example documents provided by the user	1
Amount of terms of the user profile	100
Amount of key words represented by a chromosome	20
Amount of negative terms	4
Amount of the retrieved documents provided for the user	10

We have invited fifteen persons possessing a minimum of a bachelor's degree and five years of web search experiences as the subjects. Each subject is asked to perform web search on a self-selected topic. The subject's job was just to give relevance rating to the documents retrieved by the system. All our systems are implemented in C++ and performed on Windows XP. The process of the test is as follows:

1) The user inputs one initial document by Information Collector.
2) The Learning Agent uses the terms and the terms' frequencies residing in the initial document to construct part of the positive user profile and apply it to retrieve related documents from web servers.
3) The user browses and rates the 10 retrieved documents and feedback the retrieved documents together with the relevance ratings to system by Information Collector.
4) The Learning Agent uses the input of retrieved documents and relevance ratings to construct and maintain the complete positive user profile by modifying the values of the attributes 'Term', 'Frequency', 'Sensitivity' and 'Adjusted Frequency' and the negative user profile by modifying the values of the attributes 'N-Term' and 'N-Frequency'.
5) The Learning Agent sorts the positive user profile by 'Adjusted Frequency' and the negative user profile by 'N-Frequency' in descending order. Then, selects the 16 top ranked terms from positive user profile and 4 top ranked terms from negative user profile and passes the selected positive and negative terms to the Query String Generator. The Query String Generator and the Document Selector work together to produce optimal query string for webpage retrieval.
6) The Document Selector output the 10 documents most close to the positive user profile to the user for relevance rating.
7) Repeat steps 3-6.

8) The user browses the retrieved documents and rates each document as relevant, not sure, non-relevant. This is the final relevance rating for the retrieved documents and is used to measure the performance of relevance retrieval.

Figure 2 presents the original data of percentages of relevant recommendations rated by the fifteen subjects. It has obtained an average of 76% relevant recommendations. This preliminary result shows that the IR system that we have developed could perform fairly well.

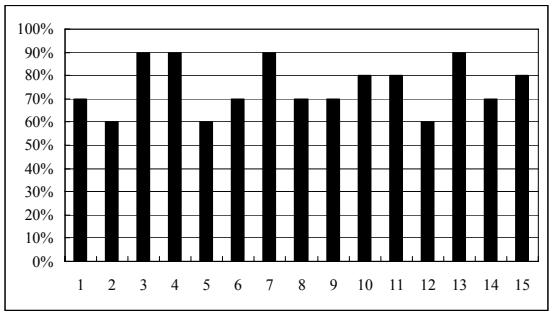

Fig.2. The original data of percentages of relevant recommendations of the 15 subjects.

V. CONCLUSION

The major contribution of this research is the demonstration of a method which is different from traditional approach in diagnosis and treatment support. This method could conquer the difficulty of system implementation requirement on the specification of data format/data structure which usually exists as free format/unstructured in the real world. Another advantage of this method is that the knowledge of diagnosis and treatment to be applied could be kept up-to-date. Therefore, the proposed method could be one of the alternatives in the many efforts endeavored to diagnosis and treatment support.

The limitation of our method is that it retrieves medical cases and documents without the examination on words meaning. It might result in misunderstanding or confusion. Future research could consider the combination of concept and ontology studies [6][18][19] to deal with the meaning of words.

REFERENCES

[1] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis," *Expert Systems with Applications*, vol. 30, pp. 272-281, 2006.

[2] S. Tsumoto, "Automated extraction of medical expert system rules from clinical databases based on rough se theory," *Information Sciences*, vol. 112, pp. 67-84, 1998.

[3] E. Sanchez, "Truth-qualification and fuzzy relations in natural languages, application to medical diagnosis," *Fuzzy Sets and Systems*, vol. 84, pp. 155-167, 1996.

[4] S. Meystre, and P. J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," *Journal of Biomedical Informatics*, vol. 39, pp. 589-599, 2005.

[5] C. F. Bassøe, "Automated diagnosis form clinical narratives: A medical system based on computerized medical records, natural language processing, and neural network technology," *Neural Networks*, vol. 8, no. 2, pp. 313-139, 1995.

[6] S. Chu, and B. Cesnik, "Knowledge representation and retrieval using conceptual graphs and free text document self-organisation techniques," *International Journal of Medical Informatics*, vol. 62, pp. 121-133, 2001.

[7] A. Segev, M. Leshno, and M. Zviran, "Internet as a knowledge base for medical diagnostic assistance," *Expert Systems with Applications*, vol. 30, pp. 251-255, 2007.

[8] C. Cole, "Intelligent information retrieval: diagnosing information need. Part I. The theoretical framework for developing an intelligent IR tool," *Information processing & Management*, vol. 34, no. 6, pp. 709-720, 1998.

[9] L, Beatriz, and E., Plaza, "Case-based learning of plans and goal states in medical diagnosis," *Artificial Intelligence in Medicine*, vol. 9, pp. 29-60, 1997.

[10] A. Holt, I. Bichindaritz, R. Schmidt, and P. Perner, "Medical applications in case-based reasoning," *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 289-292, 2006.

[11] M. Nilsson, and M. Sollenborn, "Advancements and treads in medical case-based reasoning: an overview of systems and system development," *Proceedings of the 17th International FLAIRS Conference*, 2004.

[12] B. Heindl, R. Schmid, G. Schmid, M. Haller, P. Pfaller, L. Gierl, and B. Pollwein, "A case-based consiliarius for therapy recommendation (ICONS): computer-based advice for calculated antibiotic therapy in intensive care medicine," *Computer Methods and Programs in Biomedicine*, vol. 52, pp. 117-127, 1997.

[13] R. Schmidt, and L. Gierl, "Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes," *Artificial Intelligence in Medicine*, vol. 23, pp. 171-186, 2001.

[14] Z. E. Bakaa, A. Strauss, P. Uziel, K. Maximini, and R. Traphner, "Fm-ultranet: a decision support system using case-based reasoning, applied to ultrasonography," in *Workshop on CBR in the Health Sciences*, 2003, pp.34-44.

[15] P. Perner, "An architecture for a cbr image segmentation system," *Journal on Engineering Application in Artificial Intelligence*, vol. 12, no. 6 ,pp. 749-75, 2000.

[16] S.S. Abidi, and S. Manickam, "Transforming XML-based electronic patient records for use in medical case based reasoning systems," in *Medical Infobahn for Europe*, A. Hasman, B. Bolbel, J. Dudeck, R. Engelbrecht, G. Gell, and H. Prokosch, Ed. Amsterdam: IOS Press, 2000.

[17] A. L. Houston, H. Chen, B. R. Schatz, S. M. Hubbard, R. R. Sewell, and T. D. Ng, "Exploring the use of concept spaces to improve medical information retrieval," *Decision Support Systems*, vol. 30, pp. 171-186, 2000.

[18] Q. Li, and Y. B. Wu, "Identifying important concepts from medical documents," *Journal of Biomedical Informatics*, vol. 39, pp. 668-679, 2006.

[19] W. Mao, and W. W. Chu, "The phrase-based vector space model for automatic retrieval of free-text medical documents," *Data and Knowledge Engineering*, vol. 61, pp. 76-92, 2007.