World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:1, No:3, 2007

Cross-search technique and its visualization of peer-to-peer distributed clinical documents

Yong Jun Choi, Juman Byun, and Simon Berkovich

Abstract— one of the ubiquitous routines in medical practice is searching through voluminous piles of clinical documents. In this paper we introduce a distributed system to search and exchange clinical documents. Clinical documents are distributed peer-to-peer. Relevant information is found in multiple iterations of cross-searches between the clinical text and its domain encyclopedia.

Keywords—**Clinical documents, Cross-search, Document exchange, Information retrieval, Peer-to-peer**

1. Introduction

A lot of medical information such as clinical documents resides within unstructured free-text documents. Even though it is convenient to create documents in free text by doctors, it is clear that information what they are looking for is difficult to find in the documents. Moreover, reading and going through numerous clinical documents in its entirety is neither efficient nor feasible. Reading only the documents with a key word is found has been traditionally accepted as a way of acquiring information without going through each one of them. Relevant information can be left unread when expressed without using the key word.

Cross-search technique allows doctors to find appropriate documents even if search-words or phrases are mistyped or expressed non-technical ways.

Medicine is well known for numerous compound words that derived from several Greco-Latin roots and such words are often misspelled [1]. Clinical documents within a hospital record system contains many more misspellings and incorrect punctuation than written text intended for publication. Spelling errors in medical text tend to be especially numerous in collections in which a significant fraction of the documents consists of brief notes that are reminders to oneself rather than meant for public review. A fast look over of the contents of such records is rather difficult as long as they are handwritten and poorly structured.

Manuscript received November 30, 2004.

Yong Jun Choi is a Dsc. student in Computer Science at the George Washington University, Washington DC, 20052 USA (phone: 202-994-8248; fax: 202-994-8248; e-mail: yongj@gwu.edu).

Juman Byun is a Dsc. student in Computer Science at the George Washington University, Washington DC, 20052 USA (phone: 202-994-8248; fax: 202-994-8248; e-mail: juman@gwu.edu).

Simon Berkovich is a professor in Computer Science at the George Washington University, Washington DC, 20052 USA (phone: 202-994-8248; fax: 202-994-8248; e-mail: berkov@gwu.edu).

For the demand of doctors such as finding clinical documents or fragments of the document, we propose specially developed technique that allows more convenient handling of clinical documents. The operational principle of the presented technique is based on using free text searching facilities to highlight certain appropriate parts of an inspected clinical document. However, the conventional methods of informational retrieval [2] are not suitable for this task. For this design, we have applied the associative access method (ASSA) [3]. The ASSA method is based on constructing of bit-attribute matrix by transforming a free text into a set of trigrams [4]. This gives rise to possibilities for approximate matching of pieces of distorted text and fuzzily formulated queries.

A clinical document is a doctor's journals entry written at each of a patient's visit. Each entry consists of patient ID, diagnoses and the note. There are two types of clinical documents depending on the access right to it. A local clinical document is a clinical document with full access. If a doctor kept the clinical documents on a remote server and he can see all the clinical documents, they are considered local.

A clinical document is referred as remote when full access to it is not given outside the owner. Note that remote clinical documents are remote because of the operating policy but not the physical location or affiliation. For example, clinical documents in the orthopedics department in a hospital are considered remote for pediatricians if the full access is restricted.

Clinical documents must be private, but sometimes they can be disclosed or shared between doctors for the medical treatment or research purpose [5]. For example, they can be provided to a doctor to whom patients have been referred to ensure that the doctor has the necessary information to diagnose or treat them. In addition, doctors may disclose patients' record to another doctor or health care provider (e.g., a specialist or laboratory) who, at the request of the doctor, becomes involved in the care by providing assistance with the health care diagnosis or treatment to the doctor.

The peer-to-peer distribution scheme is now one of the most prevalent Internet distributed applications due to their scalability, fault-tolerance, and self-organizing nature [6]-[8]. The current move toward peer to peer systems for document distribution affords us the opportunity to improve search substantially. This trend was triggered in 1999 by Napster [9], a centralized architecture, where a central directory server offers an index to locate data items. Where web searching is

centralized, peer to peer searching can be entirely distributed and thus more scalable and reliable. As an example for a peer to peer system, the Gnutella protocol implements fully distributed searching [10]. Queries are broadcasted to all peers by a multi-hop flood algorithm that transmits them from device to device. This article depicts GUI user interfaces instead of web interface for clarity.

2. Methodology

This section describes the details of the methodologies for information retrieval and document exchange.

2.1. Query protocol

Doctors have confidential clinical documents such as patients' records in their own computer, but sometimes they want to access other doctor's clinical documents for medical or research purposes so that they can treat patients better [5]. The new ASSA search engine available in the Internet will make it happen and facilitate communication between doctors in distance.

Here is the scenario for the document exchange. Doctor A has a website with the ASSA search engine. He also has confidential clinical documents such as patient records or clinical notes in his own computer. In this system, all users are required to join the member first for free to use the ASSA search engine before users search something with it. They have their username and password to log onto the website. A doctor can search for specific information in his local clinical documents in his computer using the ASSA engine.

In addition to searching the local clinical documents, a doctor can also search remote clinical documents. The remote clinical information is downloaded and stored in the local computer. This can be either a web server or a local computer.

The following is an example session of Dr. B searching for some information from Dr. A's clinical documents in a web configuration. (See figure 1)

1. Doctor B visits the doctor A's website and she wants to find clinical documents in which she is interested.
2. Doctor B types keywords to the ASSA search engine. It returns the result from the Dr. A's index in Doctor B's computer. The distribution mechanism is explained in the subsequent section.
3. Doctor B finds out that there are some documents there she may be interested in.
4. Doctor B now contacts doctor A to ask for the documents.
5. Doctor A responds with the documents.

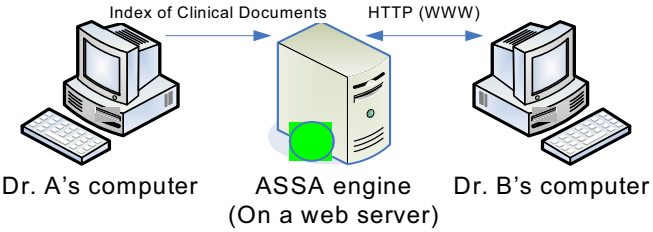

Figure1. Query protocol between doctors in distance

2.2. ASSA search engine interface

Users type keywords and press "Enter" (or click on the Search button) for a list of relevant information to enter a query into ASSA (See figure 2).

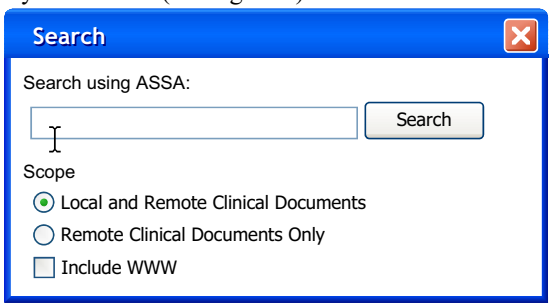

Figure 2. ASSA search engine interface

ASSA only returns information that contains all the words in the query. If a common word is essential to getting the results users want, they can include it by putting a "+" sign in front of it. Users can also exclude a word from the search by putting a minus sign ("-") immediately in front of the word they want to avoid. They also can search for complete phrases by enclosing them in quotation marks. Terms enclosed in double quotes ("like this") will appear together in all results exactly as they have entered them. Users can search for information in WWW, local clinical document and remote clinical documents. (See figure 2)

2.3. Search result

A doctor who is an administrator of the local site has full access to the local clinical documents after he logged on the website. When he searches something with ASSA, he can see the whole documents with highlighted keywords. (See figure 3) In this article underlining is used to increase readability instead of highlighting.

Two different types of access are given to the system users. Those who have full access to the local clinical documents can view them with remote ones right next to it. (See figure 3) One can choose to browse only remote clinical documents. This option can be handy when no information can be found in local clinical documents. It will search only the remote clinical documents. (See figure 4) If one does not have full access to the local documents, "Remote Clinical Documents Only" would be the only option available for her.

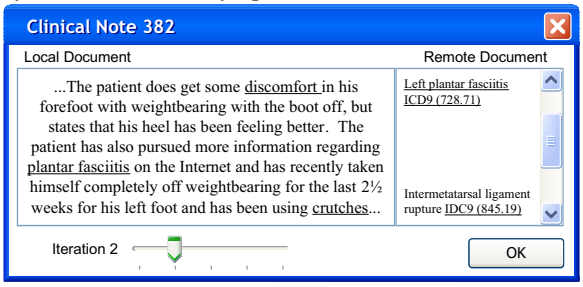

Figure 3. An example of search result for local clinical documents

Therefore, they only see the limited information about the keywords they typed in ASSA (see figure 4). Limited information can be the medical related number or code, so only doctors understand what they are exactly.

There are two different types of search results. One is for those who are searching the local clinical documents and the

other one for the remote clinical documents.

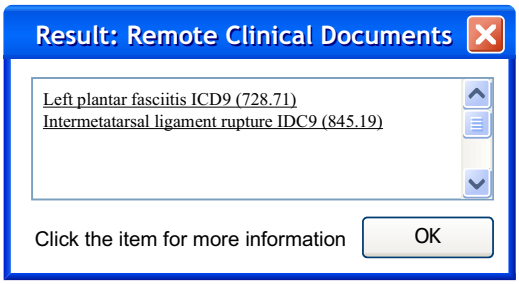

Figure 4. An example of search result for remote clinical documents

2.4. Cross-search algorithm for clinical documents

Cross-search in this article is defined as a search method to find related information by searching for the additional vocabulary. 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.

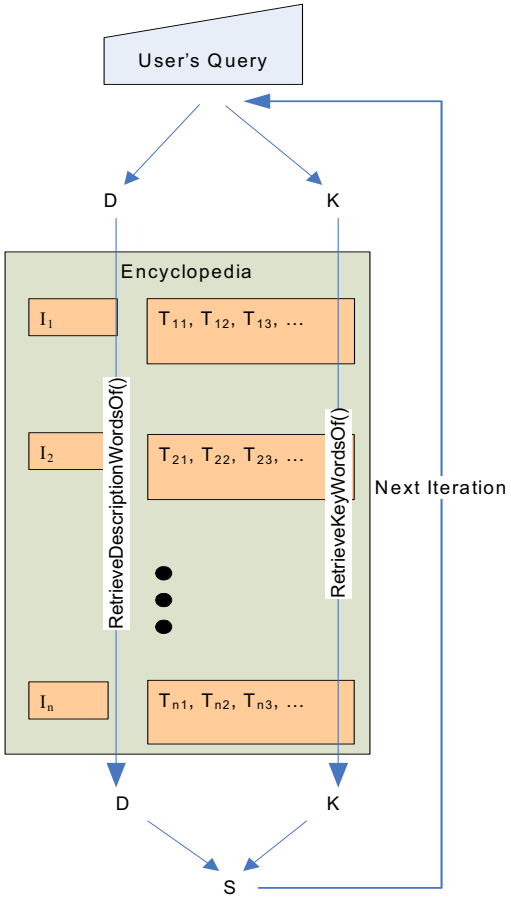

Figure 5. Cross-searches for clinical documents.

When users search for fragments of clinical documents with the ASSA search engine, they type keywords into the engine. The cross-search algorithm we have developed is working to find fragments of clinical documents.

Figure 5 shows the cross-search mechanism for clinical 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, $\{T_{n1}, T_{n2}, \ldots\}$. 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, $\{K_p, K_q, \ldots\}$

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 ASSA 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 and highlighted. The less part is highlighted, the less time it would take one to go through all the highlights.

Reading time can be significantly reduced with differential highlighting. In differential highlighting, only the differences between two iterations are highlighted. For example, if S_1 is a set of highlighted vocabulary on the initial iteration and S_2 for the second one, only S_2-S_1 is highlighted instead of S_2 on the second iteration. Which iteration is being displayed can be adjusted using the iteration bar in Figure 3. For instance, when the iteration bar indicates 2, S_2-S_1 is highlighted. 3 for S_3-S_2 and so forth.

2.5. Just-In-Place Visualization of Remote Clinical documents

Just-in-place visualization allows remote clinical documents to be referenced with ease. Remote clinical documents are displayed right by local clinical documents whose vocabulary

is matched with remote ones. (See figure 3) Only the remote notes with different diagnoses are displayed because the same diagnoses are redundant with the local notes.

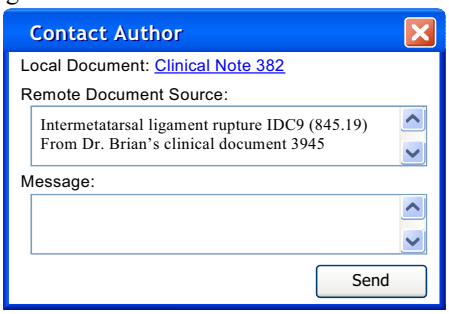

Figure 6 Contact the author of a remote clinical document

If there is a different diagnosis in a similar observation, one is given a chance to pay close attention to it in this system. When a different diagnosis is displayed on the side, one may choose to click it and contact the owner depending on the severity as depicted in figure 6. Thus the information can be used to prevent misdiagnosis of symptoms that look benign.

In addition to the benefits above, the familiarity of a doctor to the local clinical documents speeds up skimming process. No extra time would be spent going through remote clinical documents separately since related information is displayed in close proximity within the context of local clinical documents.

2.6 Indexing local clinical documents for distribution

The text and diagnoses of clinical documents are indexed. The index is updated whenever a new clinical document is added or an existing one is updated.

The text of a clinical document is indexed for full-text search. In other words any word in the text can be found without scanning every word of the text. The index is packaged into a single file. Each site generates an index of its local clinical documents. Then the index is distributed in peer-to-peer fashion among participating users.

2.7. Exchanging clinical documents between doctors using the Peer-to-peer

Clinical information is distributed peer-to-peer without special network or equipment. The indexes of clinical documents are downloaded at request.

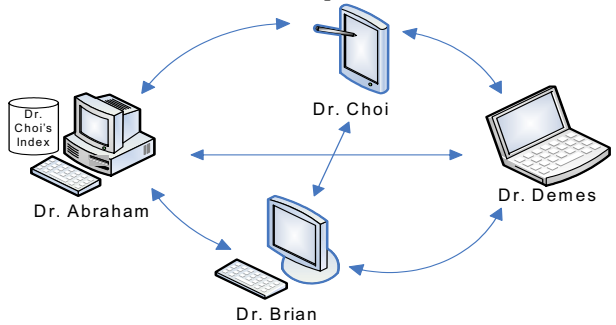

Figure 7. Peer-to-Peer Distribution of Clinical Information``

Peer-to-peer distribution makes clinical information to be downloaded from anyone in the network and eliminates the need for centralized server systems for distribution. For example, Dr. Demes would like to download Dr. Choi's index and it is not available, the system allows him to download it from Dr. Abraham's computer.

Remote clinical documents indexes are be propagated using peer-to-peer networks. In peer-to-peer distribution network, the clinical documents are distributed in a star topology.

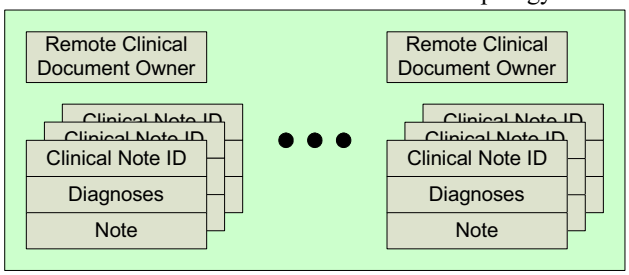

Figure 8. Index for remote clinical documents.

Just-in-place visualization of remote clinical documents requires only the index of remote clinical documents and their diagnoses. In addition, an option is provided to contact the owner of clinical document to allow further collaboration.

The indexes of clinical documents are downloaded. Then they are merged or updated into the index in the local computer. The local computer keeps track of the owner of the clinical documents in the merged index to make the next partial update possible. Thus the system uses one local index for local clinical documents and another one for remote clinical documents.

3. Conclusion

We have proposed the new system for doctors to find clinical documents or fragments of them with the ASSA search engine which has the approximate matching technique. Cross-search algorithm is used to find information in the clinical documents and we take advantage of the peer to peer for the document exchange. We believe that our information retrieval system will allow doctors to find what they are looking for effectively and efficiently in the free text documents without possessing expertise in the technical aspects of the system.

The beauty of the system is simplicity. Users of the system would only need to focus on the domain knowledge but not the internal algorithms. In the medical application we proposed, doctors would need to focus medical knowledge. Cross-search and its visualization technique were originally developed to help doctors to find useful information both in local and remote documents.

The technique is generic enough to be applicable in other domains as well as medical fields.

4. References

[1] J.M. Fisk, Pradeep Mutalik, Forrest W. Levin, Joseph Erdos, Caroline Taylor, Prakash Nadkarni, Integrating query of relational and textual data in clinical databases: A case study. Journal of the American medical informatics association Vol.10 No.1 Jan/Feb pp21-38, 2003

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

[3] 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, 951-958 (1992).

[4] 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, 62-64 (1998).
[5] Notice of privacy practices http://www.ehealthconnection.com/regions/cincinnati/content/images/HIPAA_English.pdf
[6] Tylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In proceedings of ACM SIGCOMM'01, 2001.
[7] Antony Rowstron and Peter Druschel. Storage management and caching in PAST, a large-scale, persistent peer to peer storage utility. In ACM Symposium on Operating Systems, 2001.
[8] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek and Hari Balakrishnam. Chord: A scalable peer to peer lookup service for Internet applicatioins. In proceedings of ACM SIGCOMM'01, 2001.
[9] Napster www.napster.com
[10] Clip2, The Gnutella Protocol Specification v0.4, 2001. www9.limewire.com/developer/gnutella_protocol_0.4.pdf