Internet Medical Consultant – A Knowledge-Sharing System

Some of the authors of this publication are also working on these related projects:

Project   PedAst - Clinical Decision Support System for managing asthma in school-age children   View project

Internet Medical Consultant – A Knowledge-Sharing System

Draško Nakić

Computer Science and Informatics
Faculty of Electrical Engineering and Information Technologies – Skopje, Macedonia
drasko_21@yahoo.com

Suzana Loškovska

Computer Science and Informatics
Faculty of Electrical Engineering and Information Technologies – Skopje, Macedonia
suze@feit.ukim.edu.mk

Abstract. *Medicine is a very complex science organized in a variety of disciplines. It is very difficult for a medical professional to rely only on his knowledge gained through school and practice. Aside of already known PDA medical assistants, intra-hospital consultations, web-forums, etc., our goal is to leverage the power of modern ICT to provide a system dedicated especially to doctors, for the purpose of knowledge sharing. The system simulates the following ideal scenario: All doctors and all patients in the world are in the same room, with all the logistics they need at hand. To achieve this goal our system consists of an efficient IR subsystem for fetching the desired information as quickly as possible with great relevance, a precise expert locator to find the appropriate expert to address the question, and synchronous communication system that provides remote collaboration.*

Keywords. knowledge-sharing, information retrieval, document indexing and clustering, remote collaboration, social network

1. Introduction

With the emergence of the Internet, medical information became more available to humans. Medical experts became interconnected with people in need of medical help and other experts as well. On-line consultations became very popular way of gathering valuable information [22, 2].

It is common for a hospital to provide a way to ask its experts a consultation question. The problem is that in many cases there are plenty limitations: number of words, no option for attaching files, no option for choosing the expert to send the consultation to, and no option for second opinion aside from hospital bounds. Hospitals are more concentrated in providing on-line services to potential patients and referring physicians, repositioning expert knowledge-sharing in the lower priority on-line services, if implemented at all [15].

Second source of information are web-forums. Although these popular knowledge-sharing networks involve experts from around the globe, drawbacks still exist: anyone can answer your question, you do not have insight in ones expertise or agility, you may receive a large number of questions, you do not have privacy in the communication, you can not attach files to your question, etc [5].

There are sites that provide expert consultation, enabling file attachment to the question, but that service is commonly charged [6].

Our goal was to design a system that provides consultations between experts with practically no word limits and certain number of attached files, free of charge. All of the users are evaluated and ranked accordingly and so can be located by their abilities for answering a question. The communication is secure and consultations stored in organized archive with efficient search engine. Additional features of remote collaboration are also available.

A system that is very similar to ours is the asynchronous remote medical consultant for Ghana [14]. It is designed to enhance the hierarchical communication between the medical institutions, engaging small number of physicians (around 2600). The system faces poor traffic and communication infrastructure across the country. Our, system, on the contrary, is intended to be global with not only intensive vertical, but horizontal communication as well.

The paper is organized as follows. In Section 2 we describe the problem and set the fundaments of the system. In Section 3 we describe

the system and its subsystems. We see the deployment of the system in Section 4, and finally, in Section 5, we look through further enhancements of the system.

2. Definition of the problem and system fundaments

The general framework of the system design was deducted from the analysis of the information obtained from an interview of 8 physicians at the Surgery Department in the Military Health Center in Skopje. A special consultant from that department was involved in the analysis and in modeling of the desired functionalities of the system.

Science evolves with great speed in this modern era, so it becomes harder to humans to rely only on their personal knowledge. In medicine as well, doctors face this problem every day. The problems that arise from informational perspective include:
- Medical science has a huge knowledge base
- Medical science is highly time dependant
- Information must be precise and correct

The first problem is the huge knowledge base. Our goal is to significantly shrink that ocean of information to a small pool of related information to our requirement [4]. Getting the right information, or getting the answer to a problem, is not always finding a simple statement from a knowledge base. On the contrary, it often requires reading some or sometimes much related information, so that one can induce the required information, or answer. That is why we concentrate our system to human knowledge sources. Suppose that:
- The entire knowledge base in medicine is represented with a very large set of statements $M_{kb}= (s_1, s_2, \ldots, s_n)$, where n is the total number of statements and n is a very big number. The set is not constant and it increases its size as the medical science develops
- All doctors in the world are represented with the set $D=(d_1, d_2, \ldots d_m)$
- There is a surjective function $f : D \mapsto M_{kb}$, that is interpreted as "for each statement in M_{kb} there is at least one doctor from D that knows that particular statement"

Following the previous assumptions we note DS_i is the subset of statements from M_{kb} ($DS_i \subset M_{kb}$), that i-th doctor knows. DS_i defines doctor's specialty. If we observe the doctors as knowledge nodes, each supplied with appropriate DS_i and a mechanism for deriving new knowledge, we can construct a knowledge based network. Each network node can ask other node(s) to provide it with required knowledge.

Thus we get a social-network [12, 17] based on knowledge sharing [22]. The network poses one fundamental characteristic. Each requirement for new knowledge is initiated by the patient. This means that it is the patient (more likely the patient's condition) who is the initiator of the enlargement of a certain DS_i and implicitly the enlargement of M_{kb}. We can make the following supposetion to complete the picture of the system:
- All patients in the world are represented with the set $P=(p_1, p_2, \ldots, p_r)$
- There is a m:m function $g : D \mapsto P$, that is interpreted as "for each patient in P there is a doctor from D"

DP_i is the subset of patients from P, which the i-th doctor is responsible for. Let the parameters describing patient condition be $CP(P_i)=\{cp_1, cp_2, \ldots, cp_q\}$. Having this in mind, demand for new knowledge (from now on Consultation Question - CQ) is created as follows:

$$CQ(D_i) = Q(CP(P_j), DS_i),$$

$$\text{where } i \in \{1, 2, \ldots, m\} \text{ and } P_j \in DP_i \tag{1}$$

The problem that arises from this relation is how to define $CP(P_j)$ as precise as possible. So, recalling to the third statement the precision and the correctness of the information is very important, since mistakes are not allowed. The simplest and most common to people is description by words. It is often not satisfactory and we require additional information that will complete the image of a patient. This includes laboratory tests' results and findings, MRI scans and findings, x-rays etc... Therefore we have to provide the option to formulate the CQ as text supported by additional files.

With this we define the asynchronous knowledge sharing. The system implements various ways of synchronous knowledge sharing, too. This includes the well known communication with text messages – chatting, file transfer, VoIP, video communication, as well as remote collaboration where the users of the system can discuss upon a common file, modify the same file or even co-create files. Our system also includes the option for teleconferencing to provide multi-expert discussion on some interesting and intriguing topics.

The system archives the knowledge traffic that will later set the basis for searching certain information before sending a consultation question. For this purposes we have to develop an efficient search engine that will return results with great relevance in shortest time.

Provided that we do not find the appropriate information in the archive, the system needs to locate an expert that we believe will give us the most correct answer in the shortest time interval. An expert search engine is developed, providing the user to customize the search.

Our system is a knowledge-sharing based social network because it integrates an informal layer of communication over the more formal. The informal communication is exactly the same as in well know social networks like Facebook or MySpace.

3. System description

The Internet Medical Consultant system is presented in Fig. 1. The main process is demand for consultation. A user can perform that action asynchronously, through message communication (very similar to e-mailing), or synchronously by remote collaboration, teleconferencing, triangular consultation, etc. Other features in the system are designed to support and enhance the quality of knowledge sharing.

Although this system is primarily dedicated to sharing knowledge over the Internet, it provides a tier for informal communication. This way we provide two tiers of communication, grouping the messages in separate inboxes, providing better organization, greater visibility and easier navigation with no overlapping of unrelated data in an inbox.

3.1 Asynchronous knowledge-sharing

The basic mode of knowledge sharing is via concept of messaging. The doctor who requires help creates a CQ and sends it either to the system, or to a list of specified doctors. When a CQ is sent to the system, the location of an expert is determined by an algorithm in the system. Otherwise messages are simply delivered to the doctors listed in the recipient list. The CQ is bound to its answers and formulates the basic knowledge structure in the system. The whole knowledge traffic is archived as such, so that later, users can search through it.
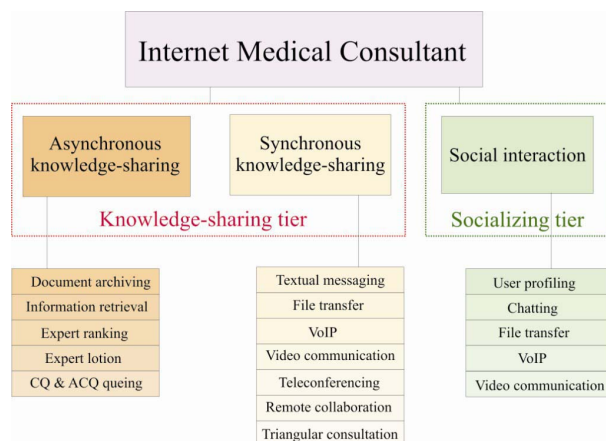

Figure 1. Logical and functional diagram of the IMC system

3.1.1 Consultation Question - CQ

Consultation Question is the moving part of the asynchronous communication. We can represent the consultation in the following manner:

$$CQ = RI \cup SI \qquad (2)$$

where RI is the set of information pointed directly to the recipient and SI is the set of information directed to the system and can be regarded as metadata for RI. We define RI and SI, and mark the elements that are obligatory when constructing the CQ as follows.

$$RI = \{Heading^*, Body^*\} \cup A$$
$$SI = \{Speciality, Subspeciality,$$
$$\quad EmergencyLevel^*, ExpiryDate^*, \qquad (3)$$
$$\quad UnknownRcp\} \cup RL$$

where A is the set of attached files to the consultation, and RL is the list of recipients. The element $UnknownRcp$ is a number that is specified only when we want to add unknown recipient to the recipient list.

If a doctor who received a CQ does not know the answer, he can forward the CQ to a list of users who he thinks may know it. This in-depth travel of a CQ is controlled by the system and there are two numbers b_f (branches) and d_f (depth) which are the top margins for the number of users one may forward a CQ to, and the number of steps a CQ can be forwarded, respectively.

The system provides an option for establishing so called Consultation Dialog (CD) for thorough discussion upon a CQ. A CD is a collection of textual Question-Answer pairs, where each of them files can be attached to. We represent a CD as:

$$CD = \{QA_1, QA_2, \ldots, QA_p\}$$

$$QA_i = \{Q, ANS\} \; ,$$

$$Q = Body \cup A \quad \text{and} \quad ANS = Body \cup A \quad (4)$$

where *Body* is the text portion and *A* is collection of attached files.

The dialog may contain valuable information and therefore it is considered to be an integral part of a CQ.

3.1.2 Archiving and Archive Search Engine (ASE)

Archiving is performed on each portion of knowledge that passes through the system [1]. This portion of knowledge is represented by a CQ, its answers (ACQ, is a collection of answers, where an answer to a consultation denoted as Consultation Answer – CA is represented as $QA = Body \cup A$) and the eventual CDs established for it. The main characteristics of this archive are: a) it is large; b) it has great variety of "documents"; c) it is constantly growing.

The archive is primarily divided in two tiers. The first is Personal Archive, where a user stores the consultations he thinks could be useful in the future, loading-off his inbox. The second is the Global Archive, where all the knowledge traffic is stored regardless of the users' perspective.

To exceed the three issues listed above we create System Defined Clusters – SDC [16, 1]. These clusters specify different fields in medicine and the documents are classified accordingly. A document is a portion of knowledge and is consisted of: CQ, ACQ and CDs established for the CQ. We note:

$$ARC = \{Doc_1, Doc_2, \ldots, Doc_N\}$$

$$Doc_i = CQ_i \cup ACQ(CQ_i) \cup DCS(CQ_i) \quad (5)$$

where *N* is the total number of documents in the archive. With this constellation, we perform clustering based only on text content.

The SDC are the primary clustering tier. Since we suppose that number of documents increases rapidly over time we should provide the following features for the clustering: a) control the size of the clusters; b) allow sub-clustering within the SDC; c) eliminating duplicates; d) reconstruct the classifying tree; and perform them altogether periodically.

The decision to which cluster a document belongs to is made by a classification tree – CT [11, 16]. To represent the document [9], we observe it as a collection of several mini-documents or independent text modules. Each mini-document

has its own weight in formulating the document representation [10]. According to this we note the document representation as:

$$DocR_i = \rho(Heading(CQ_i), Body(CQ_i),$$
$$Body_c(ACQ_i), Body_c(Q_c(CDS_i)),$$
$$Body_c(A_c(CDS_i))) \quad (6)$$

where functions *Heading* and *Body* extract the heading and the body of a single structured text item, whereas $Body_c$ separates the bodies of a collection of structured text items into a single text item. Functions Q_c and A_c extract the questions and the answers from a collection of CDs. As a result we can reformulate the representation of the document as follows:

$$DocR_i = \rho(CQ - Heading, CQ - Body,$$
$$CQ - Answers, CQ - DQuestions,$$
$$CQ - DAnswers) \quad (7)$$

Each of the arguments of the function ρ is pure text and has its own weight when formulating the index of the document Doc_i. The first step of the classifying is eliminating QA that do not contain medical information. This action is performed in the root of the CT and these answers are not archived at all. Further, classifying is performed by computing the similarity of the document and the classifiers in the nodes of the CT and traversing the tree hierarchy downwards by selecting the node with the most similar classifier to the document until we reach a leaf node which will point us to a unique cluster.

The returned results must strongly adhere to the query and be ordered in the most appropriate way possible [18, 20]. We consider the query to be a small portion of text. We index the short text, and apply it to the CT. This will lead us to a certain cluster where the system performs the similarity comparison between the query and the documents in that specific cluster. We define the similarity as:

$$S = SIM(q, Doc_i) = Sim_1(\rho_q(q), \rho_d(Doc_i)) \quad (8)$$

where the cumulative representation of the document through function ρ_d is defined as:

$$S = SIM(q, Doc_i) = Sim_2(sim(\rho_q(q),$$
$$\rho_q(CQ - Heading)),$$
$$sim(\rho_q(q), \rho_q(CQ - Body)),$$
$$sim(\rho_q(q), \rho_q(CQ - Answers)),$$
$$sim(\rho_q(q), \rho_q(CQ - DQuestions)),$$
$$sim(\rho_q(q), \rho_q(CQ - DAnswers))) \qquad (9)$$

The system calculates the similarity between the query and each textual module from the logical structure of the document. Each *sim* parameter from the function Sim_2 is weighted differently in descending order. The documents are collections of short portions of text, some of them semantically dependant, thus containing implicit information that are not represented by text. Therefore we should carefully select document representation. We suggest using the Latent Semantical Indexing – LSI [16, 1]. When using SDI we observe a document with increased granularity, meaning that we tend to separate the answers to the CQ, divide dialogs' Q-A pairs in separate question and answers in independent units when comparing them to a query. This means that the system computes the micro-similarity between the query and each of those independent units first, and then computes the macro-similarity as a weighted function of the micro-similarities.

Knowing how to evaluate the similarity between a query and a document the system returns the *r* topmost ranked results, or the results that satisfy a certain degree of similarity [21]. The system can widen the search to neighboring clusters if ASE doesn't return enough results. It is good idea to suggest similar documents to the ones returned by the ASE. We can search such documents in the same cluster or widen the search to neighboring clusters, if ASE lacks documents to return.

3.1.3 Expert location

The system provides two basic approaches for locating an expert [3, 8, 23, 13]. The first is *targeted search*, when a sender knows who the recipients for his CQ are, and the second is *blind search*, when the user lets the system find the most appropriate expert(s) for the CQ. There is also a *combined search* which, as obvious, is combination of targeted and blind search.

In a targeted search a user specifies a recipients list for the CQ, and the system delivers it to each of the entries in the list. In a blind search

given a CQ, the system should find a set of the most appropriate experts to deliver it to. A user is represented as a vector. Each dimension of the vector represents certain expertise evaluating characteristic of a user (Table 1).

Table 1. Groups of characteristics, their dimensions, default priority and meaning. Priorities marked with * are unchangeable.

Group	Default Priority	Dimension	Characteristic
System Activity Char.	1^*	c_1	Specialty
	2^*	c_2	Subspecialty
	3	c_3	Altro-activity
	4	c_4	Response Time (Sent)
	10	c_5	Ego-activity
	11	c_6	Response Time (Received)
User Feedback Char.	5	c_7	Average Response Mark
	6	c_8	Altro-activity (Saved)
	7	c_9	Altro-activity Efficiency (Read)
	8	c_{10}	Search Presence (Saved)
	9	c_{11}	Search Presence (Read)

The main purpose is to find the cluster by inspectting the Specialty and Subspecialty fields of CQ. If there is no information there, we classify the CQ by its textual content as described in *Clustering* section. The cluster of doctors the consultation is supposed to be sent is now the set of doctors who are listed in the documents either as askers or answerers [19].

$$C = Cluster_{\sigma = Specialty(CQ) \wedge Subspecialty(CQ)}(D)$$

$$RL = Sel_{\sigma = r_1}(C) \cap Sel_{\sigma = r_2}(C) \cap ... \cap Sel_{\sigma = r_c}(C)$$

where

$$r_i = best\_value(i - th\_element(Pc)) \leq$$
$$\leq i - th\_element(Pc) \leq$$
$$\leq best\_value(i - th\_element(Pc)) +$$
$$+ \Delta(i - th\_element(Pc)) \qquad (10)$$

and r_c is the number of elements in P_c. P_c is a permutation, made by the user, of the priorities of the characteristics (the first element in the permutation has the highest priority).

The system performs intersections by the sets, formed by selecting the elements with values that belong to a small best value interval, following the selection characteristics in the given order, introducing one characteristic at a time (shown in eq. 10). It stops when the set reaches a

number equal or smaller then the number of recipients. When RL is finally obtained we in fact perform targeted search.

3.1.4 CQ and Answers Queuing

Suppose a doctor receives a number of CQ that can not be represented in a single view. The CQ that are not in that view, are handicapped compared to those in it. The same stands for the received answers. Regarding these issues, the system ranks CQ to be answered, and answers received. This way we stimulate the users to improve their activity in the system, as well as their expertise, since better users get greater priority in CQ and answers ordering.

3.2 Synchronous knowledge-sharing

This is another way of sharing knowledge and it happens in real time. A user inspects a list of currently on-line users. Chatting, VoIP, video communication and teleconferencing are popular well established notions of Internet communication and they are implemented as a way of synchronous communication in the IMC system, too.

4. System deployment

The system is represented by two applications, one for personal computers and another for mobile devices (Fig. 2). The first application is designed to be used when more thorough consumption of the systems features is required, like creating a CQ answer, creating files to be attached to the CQ, searching for a certain document, organizing new knowledge gained from answers to a CQ, performing remote collaboration, triangular consultation or teleconferencing.

The mobile version of the personal computer application is dedicated to users in movement. It is used in short time intervals, usually to check for new answers, read answers to remind of something, sending very brief answers to CQs. This application implements all the features of communication, except triangular consultation.

Another feature is inbox updating notification signal whenever a user receives a CQ or an answer.
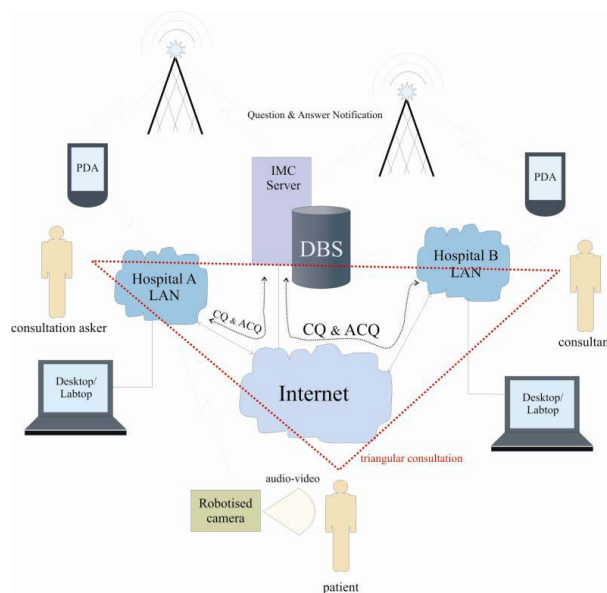

Figure 2. The overall look of the IMC system

5. Further research

Because the system is still work-in-progress, usability study will be made and performance in the real life scenario will be evaluated prior implementation of more advanced functionalities described below. This way we can conform the advanced features layer to the basic features layer, after accommodating the latter to the users' needs in the real life.

5.1 Document clustering

When we establish dialog concerning particular CQ, it sometimes happens that we go of the trail of the primary topic, and after while return to it. This discourse in the line of communication can contribute to noise when representing document content for clustering. We suggest filtering algorithm that will remove the noise, and use the text, that is strongly related to the CQ heading, when representing document for clustering. This doesn't mean that noise will be removed from the document. The filtered out noise is additionally clustered while the process of filtration. Each of the noise clusters is than created a representation which is used for classifying by the CT. If noise cluster representation is above a specified threshold of similarity, then we say that the CQ the noise was filtered out from is related to another cluster with certain strength s_{Doc_i, C_j}. A pointer from that cluster is created to the CQ, but latter does not participate in creation of the cluster represen-

tation. It is only the noise cluster that causes document-cluster soft relation, which is included in the cluster representation. However, the search engine returns the whole CQ, when its noise matches certain query.

5.2 Remote collaboration

The idea is to allow two users to modify a picture file in a way that each of them sees the modification the other performs on the file in real time. This way the users can discuss a problem, having the resource of the discussion, without the risk of ambiguity or extended time consumption.

5.3 Triangular consultation

This subsystem introduces new hardware and a robot in the system. The purpose of this subsystem is to formulate a real-time consultation in which will be involved the three main actors of a consultation: a) *initiator* – the patient; b) *consultation asker* - the responsible doctor; c) *consultant* - the remote doctor asked for consultation.

The picture of the patient is obtained by a robotized camera controlled by the responsible doctor. The robot has a microphone for audio communication and a display where the patient sees his doctor and the consulted doctor. The patient's doctor and the consultant have the other "remaining points" of the triangle on their computers display. This way the users get the illusion that they are all in a same room which is crucial for performing a precise consultation with a remote consultant.

6. Conclusion

The Internet Medical Consultant is a two-module web-application, one module for PCs and the other for PDAs. The application provides the users to gain answers to their problems by searching the archive or selecting the experts to send the question to, with additional files, gathering valuable knowledge in shortest possible time, free of charge. The application provides various ways of communication and is supposed to integrate advanced methods of communication and remote collaboration in the future, for further enhancement of patients' treatment. We also hope that this system will not only help physicians with their professional issues, but will encourage them to establish social connections through informal interaction.

7. References:

[1]   Baeza-Yates R., Rberio-Neto B., Modern Information Retrieval, Addison Wesley Longman Limited, 1999

[2]   Bian J., Liu Y., Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. *WWW 2008*, April, 2008, Beijing, China, ACM Press 467-476

[3]   Chung K. K. S., Hossain L., Davis J., Individual Performance in Knowledge Intensive Work through Social Networks. *SIGMIS-CPR'07*, April, 2007, St. Louis, Missouri, USA.

[4]   Detmer M. W., Shortliffe H. E., Using the Internet to Improve Knowledge Diffusion in Medicine. *Communications of the ACM*, 1997, 101-108

[5]   www.docsboard.com

[6]   www.doctorinternet.co.uk

[7]   Dom B., Eiron I., Cozzi A., Zhang Y., GraphBased Ranking Algorithms for Email Expertise Analysis. *DMKD' 03: 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, June, 2003, San Diego, CA, USA, ACM Press, 42-48

[8]   Ehrlich K., Shami N. S., Searching for Expertise. *CHI 2008*, April, 2008, Florence, Italy,ACM Press, 1093-1096

[9]   Hliaoutakis A., Zervanou K., Petrakis G.M.E., Automatic Document Indexing in Large Medical Collections. *HIKM'06,* November, 2006, Arlington, Virginia, USA, ACM Pess 1-8

[10]   Holub M., Semecky J., Divis J., Searching for Topics in a Large Collection of Texts. *ACL 2004 workshop on Student research*, July, 2004

[11]   Holub M., A New Approach to Conceptual Document Indexing: Building a Hierarchical System of Concepts Based on Document Clusters. *ISICT '03: 1st international symposium on Information and communication technologies*, September, 2003, 310-315

[12]   Jamali M., Abolhassani H., Different Aspects of Social Network Analysis. *2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[13]   Liu X., Croft W. B., Koll M., Finding Experts in Community-Based Question-

Answering Services. *CIKM'05*, 2005, Bremen, Germany, ACM Press 315-316

[14] Luk R., Ho M., Aoki M. P., Asynchronous Remote Medical Consultation for Ghana. *CHI 2008 Proceedings · Healthcare in the Developing World*, Florence, Italy, 2008, ACM Press, 743-752

[15] www.mayoclinic.com

[16] Manning D. C., Raghavan P., Schutze H., An Introduction to Information Retrieval, Cambridge University Press, February, 2008.

[17] Mislove A., Marcon M., Gummadi P. K., Druschel P., Bhattacharjee B., Measurement and Analysis of Online Social Networks. *IMC'07*, 2007, San Diego, California, USA, 29-42

[18] Perkio J., Buntine W., Tirri H., A Temporally Adaptive Content-Based Relevance Ranking Algorithm. *SIGIR'05,* August, 2005, Salvador, Brazil, ACM Press 647-648

[19] Reichling T., Schubert K., Wulf V., Matching Human Actors based on their Texts: Design and Evaluation of an Instance of the ExpertFinding Framework. *GROUP'05,* November, 2005, Sanibel Island, Florida, USA, ACM Press 61-70

[20] Shehata S., Karray F., Kamel M., Enhancing Search Engine Quality Using Concept-based Text Retrieval. *2007 IEEE/WIC/ACM International Conference on Web Intelligence*, 2007, 26-32

[21] Wang Z. J., Taylor W., Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method. *2007 IEEE/WIC/ACM International Conference on Web Intelligence*, 2007, 395-401

[22] Zhang, J., Mark S. A., Adamic L., Expertise Networks in Online Communities: Structure and Algorithms. *16th international conference on World Wide Web*, Banff, Alberta, Canada, 2007, ACM Press, 221-230

[23] Zhang J., Ackerman S. M., Searching For Expertise in Social Networks: A Simulation of Potential Strategies. *GROUP'05*, November, 2005, Sanibel Island, Florida, USA, 71-80