Med-Tree: A User Knowledge Graph Framework for Medical Applications

Maunendra Sankar Desarkar, Sandip Bhaumik, Sailesh K Sathish, Satnam Singh, Ranga V Narayanan
Samsung R&D Institute India - Bangalore, Bangalore, India-560037
{m.desarkar, sandip.bh, sailesh.sk, satnam.singh, ranga.n}@samsung.com

***Abstract*— The field of context and intelligence, as a topic of pervasive computing, has been gaining considerable momentum. Typically, context-aware intelligence is applied to understand the situation of the users and their behavior with the objective of providing adaptive services that are closely associated with that context. In this work, we have taken an orthogonal approach wherein we attempt to aggregate knowledge and cognition, of the user, on a given topic to build models out of them. The model thus created is analyzed to derive inferences about the user, where the analysis is performed on a graph model comprising topics based information obtained by mining domain specific personal data sources and from certain facts on which the user has expressed fair level of belief. We have explored the possibility of deriving beneficial information by provisioning an appropriate representation of knowledge as belief-graph with specific orientation in healthcare and call this model as Med-Tree. Subject to privacy conditions, we open up the belief-graph model to establish objective based social connections that gets contextually bound. As a next step, such contextually bound ad-hoc networks are subjected to advanced querying process resulting in useful information extraction and inferences. Leveraging on user's knowledge or the belief-graph, the proposed Med-Tree could help derive benefits towards better personal healthcare and disease management.**

I. INTRODUCTION

The field of context and intelligence has been gaining huge interest in recent times. The primary focus of these efforts are towards understanding the situation of the users, their behavior, and for providing context information for enabling adaptive services [1]. In medical domain, context-aware healthcare frameworks are being designed for providing improved healthcare solutions through intelligent use of patient health data, and for continuous monitoring of patient vital signs [2]. Importance is being given to wearable computing systems that are worn on the body to have better behavioral modeling of the patients [3]. Apart from this behavioral and situational modeling of the patients, one may also take an orthogonal approach and try to model the users by capturing the *knowledge* that they might be having on different medical subjects.

Knowledge can be typically divided into three categories: personal, procedural, and propositional [4]. *Personal knowledge* is also called knowledge by acquaintance and can be obtained by being acquainted with something or observing some event. *Procedural knowledge* can be gathered by performing certain task and thereby being familiar with the "know-how" for the task. Propositional knowledge, on the other hand, is the knowledge of facts. It focuses on *what* rather than *how*. *Propositional knowledge* can be gained by reading articles, documents etc. It is often referred to as *justified true belief* [5]. Capturing the belief or propositional knowledge of the users on different medical topics provides another way of modeling or profiling the users. There are some studies that propose to build knowledge profile of users from email conversations [6], measure user expertise based on the answers received in response to different questions [7], measure student knowledge in different educational data mining problems [8], [9]. However, to the best of our knowledge, construction of user profiles based on propositional knowledge on different medical topics and use of the same for healthcare applications is a novel approach that has not been discussed much in literature.

In this work, we propose a framework for capturing the users' propositional knowledge on different medical topics and subtopics, and use that in several applications related to healthcare domain.

II. ASSUMPTIONS AND THE PROPOSED SYSTEM

We begin with the assumption that the user's online access history, subscriptions, blogging/micro-blogging related to medical topics or his medical conditions provide insights into the user's propositional knowledge or belief about a particular set of topics. From these sets of documents, we extract a set of topics that are distributed across the content. The strengths of these topics are recorded in a user *knowledge graph model*, which results in an instance of a well-defined medical ontology. Google knowledge graph [10] is a popular knowledge graph, which captures concepts and their relations from knowledge ontologies like FreeBase [11]. The proposed Med-Tree is not only a connected knowledge graph, but it also represents the level of belief and knowledge of users towards particular topics.

In the current work, we develop Med-Tree using a combination of supervised and unsupervised techniques for knowledge acquisition. Medical domain has large authenticated online contents which are expanding at a rapid pace. We use these knowledge sources to mark the medical concepts present in the documents. A set of latent topics are mined from the documents using unsupervised topic modeling techniques. The supervised learning provides the ground truth information needed to map between latent topics and medical concepts. The users, the documents, medical concepts and extracted topics are stored in the Med-Tree through tree nodes and their relations. This information is analyzed to gather insights into the users' beliefs or knowledge quotients on different medical subjects and can be used in several medical informatics applications such as medical query processing, modeling user expertise etc.

III. MED-TREE USE CASE

The user scenario perceived in Med-Tree is associated with personal healthcare management. We list a few possible use cases where the proposed architecture can be useful.

1) Measuring users' knowledge quotient or expertise in different domains

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

2) Providing better answers to user queries by combining medical knowledge and mined knowledge from the knowledge base
3) Understanding query intent and directing queries to topic experts
4) Integration with electronic patient records will enable the system to use the users' medical details and provide personalized services (e.g. query answers) to the users.

A use-case scenario similar to Case 3 above exists with a popular web based healthcare portal *www.patientslikeme.com*. However, unlike this web service, Med-Tree connects with people having higher knowledge quotient on the queried topics. This is because Med-Tree is not just a graph building mechanism based on patient profiles, it actually identifies and uses the knowledge profile.

IV. IMPLEMENTATION

A. Generating a Medical Corpus

We selected a list of 21 medical topics (disease names) from the website *www.patientslikeme.com*. The reason was that it is a website where patients themselves post and discuss about different diseases. Hence, these diseases might be some important medical topics that the users would be interested to query about. The selected medical topics include: (1) Asthma, (2) Breast-cancer, (3) IBS, (4) Prostate-cancer, (5) Lung-cancer, (6) Multiple-sclerosis etc.

To get data for these topics we consulted *www.webmd.com* which is one of the most popular websites for accessing medical information. We crawled pages for the selected medical topics from *www.webmd.com* to form our corpus.

B. Extracting Knowledge from the Corpus

For each document in the corpus we know the associated medical topic. Next, we extracted the medical terms present in the documents. We used cTAKES (Apache clinical Text Analysis and Knowledge Extraction System) [12] and UMLS (Unified Medical Language System) [13] for annotating the medical terms present in the documents. Following is a sample text from a document related to asthma:

> ...Other symptoms of an asthma attack include: Severe wheezing when breathing both in and out Coughing ...

When we ran cTAKES on the document, we could mark the following UMLS concepts in the text:

> ...conceptType="PROBLEM" conceptText="asthma"
> ...conceptType="PROBLEM" conceptText="attack"
> ...conceptType="PROBLEM" conceptText="wheezing" ...

This concept extraction process uses supervised knowledge existing in the medical ontologies provided by UMLS. We may also use unsupervised techniques to group the documents into some clusters, so that each cluster caters to some unsupervised topic. We ran an LDA (Latent Dirichlet Allocation) [14] based topic modeling tool (MALLET [15]) on our corpus and mined 25 latent topics from the corpus.

For each latent topic (referred to as LDA topic) mined by the algorithm, we also have the keywords that occur frequently within the topic. Table I shows most common keywords from some selected topics that we mined from the data. Looking at the keywords, we may say that Topic 1 is about migraine and headache. Similarly, topics 7 is possibly talking about asthma.

TABLE I
TOP KEYWORDS FROM SOME OF THE LDA TOPICS

Topic 1: headaches headache migraine pain migraines tension caffeine cluster sinus mg vision nausea
Topic 7: asthma symptoms cough attack exercise allergies breathing airways breath wheezing webmd induced
Topic 9: women pregnancy menstrual hormones diet affect pregnant performance marijuana birth doctors estrogen
Topic 23: tests doctor diagnosis test blood diagnose disease procedure history make doctors diagnosing

Topic modeling may also bring out some other interesting categories from the data. A close look at the last two topics from Table I would indicate that Topic 9 is mainly associated with women. Similarly, Topic 23 discusses different medical tests and diagnoses for the diseases. A document may discuss about multiple such topics. The *proportions* or the extents to which a document contains these *latent* topics, are measured statistically by the topic modeling algorithm [14].

C. Finding associations between knowledge components

As mentioned earlier, some of the LDA topics may have high associations with some medical topics. We wanted to automatically find out these correspondences or associations. The problem of finding these associations is defined below:

Definition 1: M is the set of medical categories. L is the set of LDA categories. We wish to find ordered tuples of the form (l, m) such that LDA category $l \in L$ is associated with the medical category $m \in M$.

We have some additional information regarding the documents, keywords, LDA topics and Medical topics. D is the set of documents. The set of keywords is W. We also know the *proportion of belongingness* $\rho(d, l)$ of the document $d \in D$ to the LDA category $l \in L$. We denote by $\mu(m)$ the set of documents that belong to the medical category m. Similarly, set of documents (fully or partially) belonging to the LDA topic l is denoted as $\lambda(l)$. We compute the *association strength* $\alpha(l, m)$ between l and m as:

$$\alpha(l,m) = \sum_{d \in \mu(m)} \sum_{w \in W} \left(\mathrm{tf}_l(w,l) \log \frac{N}{\mathrm{df}_l(w,l)} \rho(d,l) \frac{\mathrm{tf}(w,d)}{\sum_{d'=1}^{|D|} \mathrm{tf}(w,d')} \right) \quad (1)$$

- $\mathrm{tf}(w, d)$: term frequency of the word w in document d.
- $\mathrm{tf}_l(w, l)$: number of times the word w is present in $l \in L$. i.e. $\mathrm{tf}_l(w,l) = \sum_{d \in \lambda(l)} \mathrm{tf}(w,d)$.
- N: Number of documents in the corpus.
- $\mathrm{df}_l(w, l)$: number of documents in $\lambda(l)$ that contain w.
- $\rho(d, l)$: belongingness of the document d in l.

It suggests that if a keyword w is frequent in l (i.e. $tf_l(w,l)$ is high) but does not occur frequently in other LDA topics (i.e. $df_l(w,l)$ is low), and is present in many documents that have high belongingness in l (i.e. $\rho(d,l)$ is high), then w has higher chance of indicating the correspondence of l with the medical categories. If such keywords are frequent in the documents $d \in \mu(m)$ (i.e. $tf(w,d)$ is high) but not so frequent across the entire collection (i.e. $tf_c(w)$ is low), then the association is stronger. High values of $\alpha(l,m)$ indicate possible correspondence between l and m.

We can define correspondence between the concept texts (as mentioned in Section IV-B) and the medical topics also. Concept texts are the problems/symptoms, medical tests, or medications for different diseases. If we have access to

the patient's health records containing the symptoms he is showing, the medical tests etc., then we can match that against the concept texts. Having a mapping from the concept texts to the medical topics might help us in identifying the topics on which the user would possibly be interested in. We formalize the problem of finding associations between the concept texts and the medical topics below:

Definition 2: Let C be the set of concept texts. M is the set of medical topics. We want to find ordered tuples of the form $(c, m) \in C \times M$ which indicates that the concept text $c \in C$ is associated with the medical topic $m \in M$.

Let $D(c)$ be the set of documents containing the concept text c. Let $D_m(c)$ be the set of documents that belong to m and contain the concept text c. Given this information, we find the association weight between c and m as:

$$\beta(c, m) = |D_m(c)|/|D(c)| \tag{2}$$

$\beta(c, m)$ is high if many of the documents that contain c belong to m. High values of $\beta(c, m)$ indicate possible correspondence between c and m. It can be seen that $\beta(c, m)$ has a probabilistic interpretation. $\beta(c, m)$ is the answer to the following question: *Given a document d that contains the concept text c, what is the probability that d belongs to the medical topic m?*

D. Storing the knowledge in a database

We store the extracted supervised and unsupervised knowledge in a graph database. This graph structure is a part of Med-Tree. The graph has different types of nodes and relations. The graph has nodes of the following types: Document, Medical Topic, LDA Topic etc. Some of the relationship types are: CONTAINS (between Document and Keyword), IN_MED_CATEGORY (between Document and Medical Topic), IS_IN (between Document and LDA Topic) etc. The corresponding data model including all node and relationship types is shown in Figure 1.

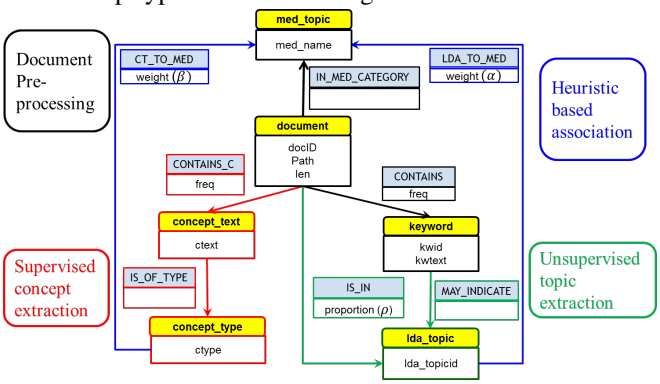

Fig. 1. The data model: Names of the nodes are in small letters. Names of the relations are in capital letters.

V. Finding Knowledge Quotient of the Users

We can extend the proposed knowledge graph architecture to determine the knowledge quotient of users. If user u has read the document d and d is in medical category m (or LDA category l), then we assume that u has gathered some knowledge about the category m (or l). Relation type KNOWS_ABOUT captures this information (see Figure 2). *Knowledge quotient* (e) is computed using Equation 3.

$$e(u, c) = \sum_{u \text{ accesses } d} \text{belongingness } (d, c) \tag{3}$$

In Equation 3, c denotes the topic (medical or LDA). If c is a medical topic, then belongingness(d, c) is 1 if d is about the medical topic c and 0 otherwise.

If c is an LDA topic, then belongingness(d, c) is same as $\rho(d, c)$.

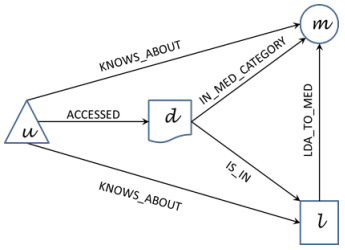

Fig. 2. Adding user nodes in the knowledge graph structure

VI. Query Processing

In this section, we describe the methods and the results for two of the use cases mentioned in Section III, namely, query processing and personalized query processing. The processing of a query involves the following steps:

1) Determine the medical topic (M) of the query (q)
2) Determine the LDA topics $L(q) = \{L_1, L_2, .., L_k\}$ of q
3) If M is non-empty, filter out a candidate set (S) of documents for the medical topic M
4) Assign scores to the documents in S depending on their belongingness to the LDA topics $L(q)$
5) Form the result set

To determine the medical topic of the query q in Step 1, we first see if q contains the name of any medical topic. However, there might be cases where the user specifies a generic name in the query (e.g. joint pain instead of arthritis) or enters the symptoms (e.g. coughing and wheezing instead of asthma). Hence, we see if any of the LDA topics in $L(q) = \{L_1, L_2, \ldots, L_k\}$ corresponds to a medical topic.

To assign scores to the documents in Step 4, we do the following. For each document (d) in the candidate set (S), we consider their *proportion of belongingness* (ρ) to the LDA topics in $L(q)$. The score of d is computed as:

$$score(d) = \sum_{i=1}^{|L(q)|} \left(\sum_{w \in q} tf(w, d) idf(w) \right) \rho(d, L_i) \tag{4}$$

$idf(w)$ is the inverse document frequency of w. Documents with highest scores are returned are results.

VII. Results

We compare the quality of the results retrieved by **Med-Tree** system against two baseline systems: (a) **MedOnly**: system that uses medical topics but no LDA topic and uses tf-idf scoring, and (b) **VS**: system that uses vector space approach and uses neither medical nor LDA topics.

A. Answering medical queries

Several queries were used for experimentation. Due to space limitation, here we discuss the following three queries.

Query q_1 is "*joint pain therapy*". However, in the corpus, there is no medical topic with name "joint pain". Using the method described in Section VI, we resolve the query's association with the medical topics osteoarthritis and rheumatoid-arthritis. The proposed system gives importance to these medical topics. q_1 is also mapped to the LDA topic that predominantly talks about treatment, cure or medication. Hence documents that are about treatment are retrieved by the system. The top-3 retrieved results are shown in Table

TABLE II

RESULTS RETURNED BY DIFFERENT SYSTEMS FOR DIFFERENT QUERIES

Query	Rank	Med-Tree: Using Medical and LDA Topics	MedOnly: Using Medical Topics	VS: Using vector Space
joint pain therapy	1	osteoarthritis-treatment-care.html	foot-ankle-osteoarthritis.html	fibromyalgia-pain.html
	2	treatment-care-rheumatoid-arthritis	shoulder-osteoarthritis-degenerative-arthritis-shoulder.html	osteoporosis-pain.html
	3	shoulder-osteoarthritis-degenerative-arthritis-shoulder.html	rheumatoid-arthritis-basics.html	foot-ankle-osteoarthritis.html
avoid constipation diet	1	controlling-ibs-with-diet.html	controlling-ibs-with-diet.html	controlling-ibs-with-diet.html
	2	ibs-triggers-prevention-strategies.html	ibs-other-treatment.html	fatty-liver-fatty-liverguide.html
	3	ibs-when-to-call-a-doctor.html	ibs-medications.html	fatty-liver-liverdoctor.html
diabetes effect on pregnancy	1	gestational_diabetes.html	diabetes_symptoms_types.html	diabetes_symptoms_types.html
	2	diabetes_diagnosis_tests.html	preventing-type-2-diabetes.html	preventing-type-2-diabetes.html
	3	gestagenic-diabetes-insipidus-symptoms-causes-treatments.html	gestational_diabetes.html	gestational_diabetes.html

TABLE III

PERSONALIZED RESULTS FOR USER QUERIES

Query	Rank	Without personalization	With personalization
cancer treatment	1	breast-cancer/breast-cancer-treatment-by-stage.html	breast-cancer/breast-cancer-treatment-by-stage.html
	2	prostate-cancer/prostate-cancer-treatment-care.html	breast-cancer/breast-cancer-treatment-care.html
	3	lung-cancer/lung-cancer-clinical-trials.html	breast-cancer/breast-cancer-clinical-trials.html
control fatigue weakness	1	fibromyalgia/fibromyalgia-work-and-disability.html	multiple-sclerosis/ms-related-fatigue.html
	2	fibromyalgia/fibromyalgia-and-fatigue.html	multiple-sclerosis/treating-multiple-sclerosis-pain-page2.html
	3	fibromyalgia/fibromyalgia-treatments.html	multiple-sclerosis/multiple-sclerosis-treatment-care.html

II. Manual comparison of the results for this query from the table indicates that the proposed system returns better responses to the query than the baseline systems.

Query q_2 is "*avoid constipation diet*". From topic modeling, we infer that constipation relates to "ibs" (irritable bowel syndrome) category in the corpus. Also, the user is seeking information on suggested food habits to *avoid* or prevent ibs. It can be seen in Table II that the first two documents returned by MedTree match this query intent of the user. MedOnly returns results from ibs but it contains links for medication too. VS brings documents from fatty-liver category also.

Query q_3 is "*diabetes in pregnancy*". The name of the disease "diabetes" is included in the query. So all the systems are able to return documents from this category. However, the user is interested in a subset of diabetes documents that also discuss about pregnancy. In the corpus, there are many documents on gestational diabetes which is a condition where women without previously diagnosed diabetes exhibit high blood glucose levels during pregnancy. Topic modeling helps to associate the query with these documents. As a result, at the top positions, MedTree returns more documents that are relevant to the user's query intent.

B. Personalized responses for medical queries in the presence of Patient Health Records

In next experiment, we wanted to give personalized query responses to patients for whom we know the medical conditions. For example, suppose we identify from medical records that a patient is suffering from breast-cancer. When she gives the query "*cancer treatment*", the system returns documents on treating breast-cancer but not on lung or prostate cancer. If a user suffering from multiple-sclerosis wants to know about means to control fatigue and weakness, the system mainly returns documents from multiple-sclerosis category. If we do not give importance to this category, then majority of the results come from the topic fibromyalgia, where also patients suffer from muscle pain and fatigue. The results for these queries are shown in Table III.

VIII. CONCLUSION

Our proposed system involves a topic extraction model from multiple data sources that identifies the users' knowledge across a range of topics. Using this, we construct a knowledge graph for healthcare domain, called Med-Tree, that depicts the knowledge quotient of each user. The application of Med-Tree includes personalized query processing, connecting users/patients based on their medical belief or similarity in medical history (disease /suffering) and medical parameters (symptoms/treatment). Our experimentation has shown significant improvement in medical query performance using Med-Tree. Our future vision involves bringing further contextual information into such system enabling context specific queries and dynamic groupings based on user requirements. We understand that this will enable many more intelligent medical applications in the future.

REFERENCES

[1] Jongyi Hong, Eui-Ho Suh, Junyoung Kim, and SuYeon Kim. Context-aware system for proactive personalized service based on context history. *Expert Syst. Appl.*, 36(4):7448–7457, May 2009.

[2] Bingchuan Yuan and John Herbert. Web-based real-time remote monitoring for pervasive healthcare. In *PerCom Workshops*, pages 625–629, 2011.

[3] Dario Farina, Ernestina Cianca, Nicola Marchetti, and Simone Frattasi. Special issue: Wearable computing and communication for e-health. *Med. Biol. Engineering and Computing*, 50(11):1117–1118, 2012.

[4] Theory of knowledge. `http://www.theoryofknowledge.info/what-is-knowledge/types-of-knowledge/`.

[5] Justified true belief. `http://www.nutters.org/log/jtb`.

[6] Eric Wang David L. Gilmour. Method and apparatus for constructing and maintaining a user knowledge profile, 07 2002.

[7] Sumit Basu, Lucretia H Vanderwende, and Lee Becker. Adaptively presenting content based on user knowledge, 06 2013.

[8] Anna N. Rafferty, Michelle Lamar, and Thomas L. Griffiths. Inferring learners' knowledge from observed actions. In *Educational Data Mining*, pages 226–227, 2012.

[9] Kim Kelly, Neil T. Heffernan, Cristina Heffernan, Susan R. Goldman, James Pellegrino, and Deena Soffer Goldstein. Estimating the effect of web-based homework. In *AIED Workshops*, 2013.

[10] Google knowledge graph. `http://www.google.co.in/insidesearch/features/search/knowledge.html`.

[11] Freebase. `http://www.freebase.com/`.

[12] Apache cTAKES. `http://www.nlm.nih.gov/research/umls/`.

[13] Unified medical language system. `http://www.nlm.nih.gov/research/umls/`.

[14] Steven P. Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In *Mining Text Data*, pages 129–161. 2012.

[15] Mallet homepage. `http://mallet.cs.umass.edu/`.