Personalized Healthcare Cloud Services for Disease Risk Assessment and Wellness Management using Social Media

Assad Abbas

North Dakota State University, USA
assad.abbas@ndsu.edu

Mazhar Ali

COMSATS Institute of Information Technology, Abbottabad, Pakistan
mazhar@ciit.net.pk

Muhammad Usman Shahid Khan

COMSATS Institute of Information Technology, Abbottabad, Pakistan
ushahid@ciit.net.pk

Samee U. Khan

North Dakota State University, USA
samee.khan@ndsu.edu

Abstract

We propose a cloud based framework that effectively manages the health related Big-data and benefits from the ubiquity of the Internet and social media. The framework facilitates the mobile and desktop users by offering: (a) disease risk assessment service and (b) consultation service with the health experts on Twitter. The disease risk assessment is performed through a collaborative filtering based approach whereas the hubs and authorities based approach is employed to identify the health experts from Twitter. The framework is implemented as Software as a Service (SaaS) to provide the disease risk assessment and expert user interaction services. Experimental results exhibit that the proposed framework achieves high accuracy as compared to the state-of-the-art approaches in terms of disease risk assessment and expert user recommendation.

Keywords: Cloud computing, expert users, health big-data, risk assessment

1. Introduction

The recent growth in the number of computing and mobile devices has resulted in exponential increase in data volumes over the Internet. Apart from the gigantic data volumes, the complex task of managing the concurrently originating data from multiple sources requires Big-data enabled tools and techniques [BuC10]. Big-data refers to the data with high volumes, high dimensionality and veracity, and greater velocity [BaH13]. The trends in rapid growth of data have also been witnessed in healthcare domain besides the electronic commerce and various scientific domains [AbB15]. Traditionally, Big-data related to healthcare originates from the sources, such as the payer-provider data repositories and the genomic-driven Big-data sources. The payer-provider data comprises of the Electronic Health Records (EHRs), pharmacy prescriptions, insurance data, and patients' feedback, whereas the genomic-driven data consists

of genotyping data, gene extraction data, and sequencing data [MiL12]. The need to exchange and integrate the electronic medical information dispersed across various points-of-care, laboratories, health insurance providers, and medical research centers obligates the efficient, robust, and cost effective storage and communication infrastructure. In this regard, cloud computing paradigm has exhibited tremendous potential and has also drawn the attention of both the academic institutions and research organizations [Kha14]. Above and beyond the performance benefits of cloud computing and Big-data analytics in the healthcare domain, fiscal concern is also among the factors of paramount importance that harnesses the need for Big-data analytics. According to a 2013 survey by McKinsey, the healthcare expenditure of the United States has increased approximately \$600 billion more than the expected benchmark [KaK13]. By embracing the cloud computing services in the healthcare domain, the expenditures for infrastructure development and subsequent management can be reduced that can further help in cutting-down the healthcare costs. Moreover, there is also a need to formulate patient-centered methodologies that involve patients themselves to manage their health affairs and devise wellness plans.

To this end, we propose a framework that facilitates the users or patients in offering personalized healthcare services at no cost using the Internet and social media. The framework primarily offers two services namely, **(a)** disease risk assessment and **(b)** health experts recommendation from Twitter. To accomplish the task of disease risk assessment, we propose an approach called the Collaborative Filtering-based Disease Risk Assessment (CFDRA). The CFDRA approach works by comparing the profiles of enquiring users with the profiles of existing patients. The typical profile attributes that are provided as input to the framework include age, gender, ethnicity, weight, height, family disease history, and other commonly observed symptoms for a disease. Based on the attributes specified in the users' query, the enquiring users' profiles are compared with the existing users and the users are returned a risk assessment score for that disease. Contrary to the various existing approaches used to make disease assessment for only a single disease, our framework is capable of performing simultaneous risk assessments about multiple diseases for several users.

The second module of the proposed framework recommends the health experts to end-users. To identify the health experts for the enquiring users to seek advice at no cost, we utilized the tweets of the users who regularly use Twitter [Twi14]. The users specify the name of the disease in their query and in turn are offered a ranked list of the experts for that disease. The tweets from health professionals are either related to health issues where the experts are mostly speaking about their experiences with patients or the tweets may be to promote health awareness in the public besides other social tweets. Likewise, large numbers of tweets containing health related terms are by another category of users that are not health professionals. Instead the users may be, **(a)** current or past patients of a disease whom they talk about more frequently, **(b)** family members of the individuals suffering from a particular disease, and **(c)** health activists and journalists who are not doctors. Such users are usually knowledgeable enough to guide the other users or patients having no or little exposure about that disease and therefore, we consider such types of users as the expert users in our framework. However, they are not regarded as the doctor experts. Hereafter, the doctors and physicians are termed as the doctor experts, whereas those mentioned above are characterized as the non-doctor experts. However, it is important for the framework to separate doctors from non-doctor experts. The tasks of user segregation and the subsequent ranking are performed by employing the hubs and authority [EaK10] based approach.

To perform the tasks of disease risk assessment about several diseases for multiple users simultaneously and to process the large tweets repositories to identify and rank the experts, parallel task execution mechanisms and enormous amount of storage are required. Therefore, cloud computing based scalable solutions seem apt not only to support the task of parallelization but also to meet enormous data storage and processing requirements for the proposed framework. The tweet repositories are updated and maintained by executing periodic jobs in offline mode to collect and preprocess the tweets to identify disease specific experts in an efficient manner.

1.1 Research Contributions

The main contributions of the paper are as follows:

- We present a cloud based framework that integrates the Collaborative Filtering (CF), social media platform, and social network analysis techniques to manage large volumes of health Big-data.
- We propose an approach for disease risk assessment using the CF. The approach is capable of simultaneously entertaining multiple users' queries to make risk assessments for different types of diseases.
- An expert recommendation module is proposed to help users seek advice from the health experts available on Twitter. The hub and authority based approach is employed to ensure that the users are recommended the most relevant and popular experts (doctors or non-doctors) as specified in the users' queries.
- The experiments for the disease risk assessment are conducted on the National Health and Nutrition Examination Survey (NHANES 2009—2010) dataset whereas the validity of expert user recommendation module is performed on a huge collection of health related tweets. Experimental results testify the effectiveness of the approach in turning the Twitter into a Web based collaborative health community.
- The framework is implemented as a Software as Service (SaaS) to offer scalable processing, storage, and task parallelization.
- The scalability analysis is conducted by increasing the workload and the number of resources for both of the modules.

The remainder of the paper is organized as follows. The motivation for the proposed work is discussed in Section 2. Section 3 discusses the architecture of the proposed cloud based system in detail. Section 4 presents the results and discussion on the performance of the framework in comparison to the state-of-the-art approaches. The related work is presented in Section 5 and Section 6 concludes the paper.

2. Motivation

Since last few years, the use of portable computing devices and smart phones has excessively increased to perform informational searches about health over the Internet. Pew Internet Project survey of year 2013 reported that around 72% of the Internet users consulted the Internet to find health information during the year 2012 [FoD13]. A total of 16% of the online information seekers in the said survey were interested in finding the people having similar concerns, 30% of the users referred to online reviews and treatment services, while 26% of the users looked for the experiences of others on certain health related issues [FoD13]. Moreover, due to the rising healthcare costs, individuals have also started taking initiatives to keep themselves healthy. Considering the importance of patient-centric healthcare services, several online tools for health

risk assessments have been developed. Data mining and machine learning approaches have widely been used for disease risk prediction, prevention, classification, and disease surveillance. Despite the capabilities of the aforementioned models in developing better understanding about the causes of diseases and to learn the appropriate counter measures, they pose realistic challenges concerning the data size, complexity, and data biases. Consequently, the development of more scalable and efficient approaches to discover the meaningful patterns from health data is needed more than ever [ChD13]. In this regard, we propose an approach that uses collaborative filtering to make risk assessment about diseases. Contrary to the several existing methodologies that permit disease risk assessment for only one disease, the proposed CFDRA approach is capable of making risk assessment for several diseases and several patients simultaneously. Moreover, the CFDRA has capability to manage large datasets by reducing their sizes. The influential profile attribute that contributes more than the other attributes in the presence or absence of a disease is selected. Based on the influential attribute, the profiles of all of the existing patients of that disease are retrieved for subsequent comparison with the profile of the enquiring user.

Online health communities and social networking websites, such as Twitter and Facebook have also emerged as the big sources of health related data. Users of the social media networks share and exchange knowledge and experience about various diseases and health related issues. The apparent purpose of expressing the feelings about health on public platforms like Twitter may be to seek out the advice or suggestions from the experts who also use social media to share their experiences. The Pew Internet Project survey [FoD13] also reveals that searching online health support by construing through the health microblogs and Web based health communities proves an inexpensive or mostly free alternative, particularly for the uninsured individuals. Besides convenient conversations with peers, psychological support is a major benefit of the online health communities [ZhY14]. Considering the efficacy of online health communities, the potential of these communities needs to be fully utilized to enhance awareness about health related matters and to offer health consultations at low or no cost. Therefore, this is the appropriate time to develop pervasive tools and methodologies having integrative support to help users make assessments about the health and to seek expert advice from doctors and patients participating in the social media communities. This work also proposes an interaction mechanism between the patients and health experts from Twitter. Twitter is currently a massive data source containing discussions ranging from political affairs to the health related issues. According to Symplur [HeS14], Twitter currently contains 558,624,884 healthcare tweets, around 10,000 doctor profiles, and 5,039 health communities. Besides the names of the diseases for which risk assessment is to be performed, the enquiring users also specify whether they are interested in consulting the doctor experts or non-doctor experts. An important task during expert user recommendation process is to identify the doctors and non-doctors based on tweets instead of the Twitter profiles because not all of the Twitter users mention their profession in the profile. Generally, it has been observed that the tweets by the doctors contain more specialized medical terms related to the disease(s) besides the general disease names, whereas the non-doctors' tweets related to health usually contain names of the commonly known diseases. This observation serves as the basis for the proposed expert user recommendation module to segregate the doctors and non-doctor experts from the huge corpus of tweets.

We are optimistic that the framework would be useful for individuals that are interested in making risk assessment for several diseases and to obtain the health advice at low or no cost. The framework can be accessed from broad range of devices, such as desktop computers,

smartphones, and tablet PCs to utilize the offered services. The framework ensures ubiquitous delivery of health related information to patients and can prove a great tool to make users or patients aware about health affairs so that they could devise their wellness plans accordingly. Moreover, the framework can be useful to avoid doctor visits for consultation because the information about health issues can easily be obtained using the presented Web based services. Furthermore, the users are recommended disease specific experts who may subsequently be contacted via Twitter, email, or through any other communication medium that is agreed upon by both the patients and the experts.

3. Proposed System Architecture

The architecture of the proposed cloud based framework for personalized healthcare services for disease risk assessment and wellness management comprises of the following major modules: (a) disease risk assessment module and (b) expert user recommendation module. The architecture of the proposed framework is depicted in Fig. 1. The framework is capable of managing disease risk assessment queries simultaneously for several querying users. Moreover, the expert user recommendation module utilizes the huge corpus of health related tweets to identify the health experts that are most relevant to the user query. It requires large amount of storage and parallel processing to periodically update the tweet repositories to efficiently answer users' queries. Therefore, the framework is implemented as an interface to the cloud environment because of the key characteristics of the cloud computing, such as the scalability, pervasiveness, and cost effectiveness [KhO14]. The details about the architecture of the proposed framework are presented in Section 3.1, Section 3.2, and Section 3.3.

Fig. 1: Architecture of the proposed cloud based framework

3.1 Disease Risk Assessment Module

To make assessment about the occurrence of diseases that a person may have in future, we employed an approach called Collaborative Filtering-based Disease Risk Assessment (CFDRA) approach. The CFDRA approach determines the similarities between the profiles of enquiring

users and the existing patients or users who have been diagnosed the same disease. The CF is the most popular technique employed in recommender systems to predict the information regarding the tastes and preferences of a certain user from large datasets by computing the similarities with the other users [AbZ15]. In recommender systems, the preferences or tastes of different users are considered to be similar if their assigned ratings/values about different items resemble. However, there are no items and ratings in the case of disease risk prediction. Instead there are different types of attributes, such as the continuous, categorical, and binary attributes. Therefore, the proposed framework uses the normalized weights for each of the profile attributes. Normalizing the attribute values is important because some of the attributes may have significantly high values than the other attributes that eventually affects the overall assessment score. For example, the value of age will always be significantly higher than the attributes having binary values. Therefore, normalizing helps in confining the values between 0 and 1. Fig. 2 presents the working of disease risk assessment module. The symbols used throughout the paper are defined in Table 1.

Fig. 2: Disease risk assessment module

Contrary to various existing approaches, such as [YuL10] and [HeE08] that focus on developing prediction models about any specific disease only, our approach is capable of making predictions for multiple individuals with different disease risk assessment queries. The framework stores the profiles of existing users having similar diseases together. The rationale is to avoid the excessive computations that may have to be performed in case when a single query is matched with the entire database of diseases with millions of dissimilar disease profiles. In other words, to perform the risk assessment about a disease x , only the profiles of patients or users having disease x should be matched, not the entire database of diseases. The users' profiles consist of several attributes, such as the age, gender, ethnicity, height, weight, and several other attributes that are amply specific to a disease. These attributes may have significant impacts on the presence or absence of a disease in an individual. A disease risk assessment system that utilizes multiple attributes for numerous diseases, gives rise to high data volumes that eventually results in the demands for compute-intensive infrastructure. Therefore, to make processing efficient, the CFDR minimizes the dataset search space by applying a reduction approach based on the importance or influence of the attributes. However, it is also ensured that reducing the dataset size does not affect the prediction accuracy. The profile attributes of a diabetic patient

may include the “age”, “gender”, “ethnicity/race”, “height”, “weight”, “diagnosed high blood sugar or pre-diabetes”, “diabetes family history”, “physical activity”, “ever observed high blood

Table 1: Symbols and definitions

Symbol	Definition	Symbol	Definition
Q	Set of querying users	I	set of importance scores of all attributes
e	Existing user	I_a	importance score of attribute a .
d	Disease for which risk assessment is to be done	q_{pt}	Important attribute in user query
q	Enquiring user	R	Risk assessment score
P	Profiles of existing users	U	Set of users
γ	Shortlisted profiles of existing users	K	Set of keywords
δ	profiles of users having a particular attribute	a_d	Authority score for a disease d
$\nexists$	profiles of users without influential attribute	h_d	Hub score for a disease d
A	set of all attributes in the users' profiles	M	Matrix
E	List of expert users	K_d	set of Keywords against disease d
T_d	collection of tweets against disease d	T_k	collection of tweets against keyword k
U_d	set of users collected who had tweeted about disease d	C_{ukd}	number of times user u have used keyword k of disease d in his/her tweets
M_d	user to keyword popularity matrix for disease d	N	number of required expert users

pressure”, “blood cholesterol”, and “smoking”. The selection of user profiles is made on the basis of the attribute that highly affects the presence of that disease. For example, family diabetes history is an important marker for the presence or absence of diabetes in an individual because of the genetic disposition [LiT03]. Therefore, the profiles of users that have a diabetic family history are retrieved for subsequent profile matching. The approach can be generalized to all of the diseases because for every disease such influential attributes exist. Moreover, the CFDR approach observes the value of the influential attribute in the profile of enquiring user. Based on the observed value, only the profiles of the existing users are retrieved to compute the similarities. There are a number of similarity metrics proposed in the literature, such as Pearson Correlation, Cosine Similarity, and Jaccard index. The Pearson Correlation is similar to the Cosine Similarity matrix except that it subtracts the average ratings of all the items given by the users from the value of item rated by that user. The Pearson correlation results perform better if all the ratings are given against similar items, for example the movies. However, in case of medical records, the values of users for the attributes, such as the “age” and “family diabetes history” cannot be correlated to each other as one is continuous and other is a binary attribute. Similarly, the Jaccard Index is used if all the attributes are binary in nature. Therefore, Cosine Similarity measure is appropriate for medical data where attributes are of different types, such as continuous, discrete, and binary. The proposed CFDR approach also uses the Cosine Similarity for similarity computation between the profile of the enquiring user and the existing users or patients. To compute the Cosine Similarity between the profiles of the enquiring user q and each of the existing users' e , both q and e are represented as the vectors and the Cosine of the angle between these two vectors is computed [SaK01]. The following equation is used to compute the Cosine Similarity $sim(U_q, U_e)$:

$$sim(q, e) = \frac{\sum_{i=1}^n (q_i \times e_i)}{\sqrt{\sum_{i=1}^n (q_i)^2} \times \sqrt{\sum_{i=1}^n (e_i)^2}} \quad , \quad (1)$$

After the similarities are computed, the following equation is used to compute the risk prediction $P(q, d)$ for disease d , for a given user:

$$P(q, d) = \bar{r}_q + \frac{\sum_{e \in U} sim(q, e)(v_{e,d} - \bar{v}_e)}{\sum_{e \in U} sim(q, e)} \quad , \quad (2)$$

where $\bar{r}_q$ is the mean of each of the attributes of q , $v_{e,d}$ represents the predicted value of disease d for the existing user e , and $\bar{v}_e$ represents the mean for particular attribute of the existing user. The algorithm for disease risk prediction is presented as Algorithm 1.

Algorithm 1: Disease Risk Assessment

Input: Set of querying users Q for disease $\bar{d}$

Output: Disease risk assessment score R for all querying users Q

Definitions: $\bar{d}$ = disease profile, q = enquiring user, P =profiles of existing users, γ =shortlisted profiles of existing users, δ =profiles of users having a particular attribute,

$\forall$ = profiles of users that do not have a particular attribute, A = set of all attributes in the users' profiles, I = set of importance scores of all attributes, I_a = importance score of attribute a .

```

1: PARFOR attribute  $a \in A$  do
2:    $I_a \leftarrow getImportance(\bar{d})$ 
3: end PARFOR
4:  $\mu \leftarrow getImpProfileAttributes(I)$ 
5:  $P \leftarrow retrieveProfiles()$ 
6: PARFOR querying user  $q \in Q$  do
7:   if ( $q_\mu == true$ ) then
8:      $\gamma \leftarrow \{\mu \in P | \mu \notin \delta\}$ 
9:   else
10:     $\gamma \leftarrow \{\mu \in P | \mu \notin \forall\}$ 
11:   end if
12:   PARFOR user  $e \in \gamma$  do
13:      $S_{qe} \leftarrow sim(q, e)$ 
14:   end PARFOR
15:  $R_q \leftarrow getAssesmentScore(S, \gamma)$ 
16: end PARFOR
17: Return  $R$ 

```

In Line 1—Line 4, for each attribute in the set of existing profile attributes, Algorithm 1 identifies the important or influential attribute with the high count for a particular value of attributes that may play significant role in the presence or absence of a disease. This is the attribute that is present in most of the profiles having the enquired disease. The *PARFOR* statements in the algorithm show that the tasks are being performed in parallel. The profiles of all of the existing users are retrieved in Line 5. Line 6—line 16 compare the profiles of each of the enquiring users with the existing users and the task is executed in parallel for multiple users and multiple diseases. In Line 7—Line 11, it is determined whether the attribute identified in Line 4 is present in the query of the enquiring user. In case the attribute is found in the profile of the

enquiring user with the value equal to “YES”, the profiles of existing users having the corresponding value of that attribute are retrieved in Line 8. Otherwise the profiles of the users having value “NO” for that attribute are retrieved in Line 10. Line 12—Line 14, compute the similarities between the profile of the enquiring user and the existing users as presented in Eq.1. The disease risk assessment score is computed in Line 15 using Eq. 2 and the calculated score is returned in Line 17.

3.2 Expert User Recommendation Module

The expert user recommendation module finds the expert users who frequently tweet on Twitter particular to the health activities. The proposed framework considers two types of users as the expert users namely: **(a)** the doctors and **(b)** non-doctor experts. Fig. 3 depicts the working of expert user recommendation module.

Fig. 3: Expert user recommendation module

The expert user recommendation module works by evaluating the tweets to segregate the doctor and non-doctor experts based on the health related keywords used in tweets. Separating doctors from non-doctors on the basis of tweets is important because not all the Twitter users mention their professions in the Twitter profile that makes it difficult to determine that whether a user is a doctor or a non-doctor. To separate the doctors from non-doctors, we made an observation that the doctors tweet differently from the non-doctors or patients. The doctors' tweets contain not only the generic health terms but also the specialized medical terminologies pertaining to a disease. For example, for diabetes, the relevant terms, such as “insulin”, “blood sugar”, “metformin”, “pre-diabetes”, “mellitus”, “Type 1”, “Type 2”, “glucose”, “metabolic”, “polygenic”, “ketogenic” etc. are commonly found in doctors' tweets. On the other hand, the tweets by non-doctors usually contain generic keywords including the disease names and symptoms, such as “feeling sick”, “suffering”, “my doctor”, “blood pressure”, “aching”, “muscles”, “pain” etc. Although the non-doctors may also use specialized medical terms in their tweets but it rarely happens. Therefore, to identify the health experts based on the use of health related terms and keywords in tweets, the hubs and authorities based approach is employed. We used WordNet to retrieve domain-specific health and medical terminologies. WordNet is a

lexical database for English language comprising of Sets of Synonyms (Synsets), nouns, and verbs [Mil95]. The benefit of using the Synsets is that they contain the synonymous words and can represent the correlation among the words such that the semantic relationship based on the hypernym, hyponym, meronym, and holonym, and derivationally related terms [Mil95] become more obvious. Consequently, the WordNet serves as ontology to derive the semantic associations from the health related terms. An example of the WordNet semantic representation of diabetes disease is presented in Fig. 4.

Fig. 4: Example of related terminologies for the term Diabetes in WordNet

The framework maintains the tweet repositories comprising of the general health related terms by executing the periodic jobs offline to extract tweets from Twitter. The advantage of the offline processing is that it avoids the limitations of online processing in terms of time efficiency. Based on the user query requesting the services of the health experts of a particular disease d , the disease specific terms, such as the hypernym, hyponym, meronym, holonym, sister terms, and derivationally related terms are used to create disease specific tweet repositories. The profiles of all of the users of the disease specific repository are searched to determine the occurrences of the health related keywords. On the basis of the keywords used by each user, a user-keyword popularity matrix is constructed. The user-keyword popularity matrix identifies the candidate experts with high number of keywords and is constructed on the basis of following equation.

$$U_i^d = \sum_{j \in J} K_{ij}^d, \quad (3)$$

where U_i is set of users and K_{ij} represents the keyword j used by a user i specific to any disease d . The experts identified using the keyword popularity may or may not be the actual health experts as desired by the user because it is quite probable that despite of the high keyword count and frequent use of archetypal health terms, the identified candidate expert is a non-doctor (a patient, family member of the patient, health activists, and health journalists). Therefore, for the enquiring users interested in interaction with the non-doctor experts, the keyword popularity based approach works reasonably well. However, when the interaction with the doctors is requested, the approach based on keyword popularity does not seem effective because it determines popularity on the basis of the total number of keywords by a user. This leads to the assumption that the users repeating only a few archetypal keywords in their tweets may possibly be non-doctor experts whereas the doctors use specialized medical terminologies that are less known to the common people. Therefore, the keyword popularity is not a true characterization of

the capabilities of experts, particularly for the doctors. It is more important for the framework to accurately identify the experts as the potential doctors and non-doctor experts for a disease by providing a ranking score for each of them.

A more appropriate way to avoid the experts identification biased towards the keyword frequency is to take into account multiple keywords that are related to a specific disease and then generate the ranking scores. Therefore, we used the Hyperlink-Induced Topic Search (HITS) [EaK10] algorithm to identify and rank the experts that are adequately knowledgeable about the health matters. The HITS algorithm uses the concepts of hubs and authorities to accomplish the ranking task by performing repeated improvements. The HITS was originally proposed as the solution to the Web search problem where a page that points to many other pages is considered as a hub whereas an authority is the page pointed by many other pages [EaK10]. In other words, a page pointed by the other pages having high hub scores is assigned the higher authority weights. Likewise, for the pages pointing to multiple high authority pages, a high hub weight is assigned.

In the proposed framework, the purpose of using hubs and authorities is to identify the users that use a set of keywords with varying frequencies. Similarly, a set of keywords that is being used by the experienced users is also identified to make the ranking process more explicit. The expert users are considered as the hubs, whereas the keywords used by the expert users are considered as the authorities. The hubs (users) that use good authorities (keywords) are assigned higher weights. Similarly, the popular keywords used by the good hubs (expert users) are assigned higher weights that significantly affect the ranking process. In fact, the importance of both the keywords and the users of keywords are helpful in identifying the experts. To produce the ranking of the expert users based on the hubs and authority scores for a particular disease d , a matrix M with U rows and V columns is created. Suppose $[h_d]$ and $[a_d]$ be the matrices for hub and authority scores. The authority and hub scores are calculated using Eq. 4 and Eq. 5, respectively.

$$a_d = M_d^T \times h_d , \quad (4)$$

$$h_d = M_d \times a_d , \quad (5)$$

Similarly, the authority and hub scores at any i -th iteration are given by Eq. 6 and Eq.7, respectively.

$$a_d^i = (M_d^T \times M_d) \times a_d^{i-1} , \quad (6)$$

$$h_d^i = (M_d \times M_d^T) \times h_d^{i-1} , \quad (7)$$

The approach works recursively by assigning all the hubs and authorities as the initial score of 1 followed by the authority update rule to the current score. On the resulting scores, the hub update rule is applied. Algorithm 2 presents the steps for the expert user recommendation module.

In Line 2 of Algorithm 2, the keywords related to the disease d are obtained from the WordNet. From Line 3–Line 6, the tweets repository is searched against each of the keywords to identify the disease specific tweets. From Line 7–Line 10, the users that frequently tweet for a

Algorithm 2: Expert User Identification

Output: List of expert users E

Definitions: D = set of Diseases, K_d = set of Keywords against disease d , T_d = collection of tweets against disease d , T_k = collection of tweets against keyword k , U_d = set of users collected who had tweeted about disease d , C_{ukd} = number of times user u have used keyword k of disease d in his tweets, M_d = user to keyword popularity matrix for disease d , N = number of required expert users

```
1: PARFOR disease  $d \in D$  do
2:    $k_d \leftarrow \text{keyWordsSearch}(d)$ 
3:   PARFOR keyword  $k \in K_d$  do
4:      $T_k \leftarrow \text{searchTweetRepository}(k)$ 
5:      $T_d \leftarrow T_d \cup T_k$ 
6:   end PARFOR
7:   PARFOR tweet  $t \in T_d$  do
8:      $u \leftarrow \text{extractUser}(t)$ 
9:      $U_d \leftarrow U_d \cup u$ 
10:  end PARFOR
11:  PARFOR user  $u \in U_d$  do
12:     $ut \leftarrow \text{tokenize}(u)$ 
13:    PARFOR keyword  $k \in k_d$  do
14:       $C_{ukd} \leftarrow \text{getKeywordCountInProfile}(ut, k)$ 
15:    end PARFOR
16:  end PARFOR
17:   $M_d \leftarrow \text{generateMatrix}(U_d, K_d, C_d)$ 
18:   $\hat{C}_d \leftarrow \text{getTopCandidateExperts}(M_d)$ 
19:   $R_d \leftarrow \text{getRankedExperts}(M_d)$ 
20:   $E_d \leftarrow \text{getTopRankedExperts}(R_d, N)$ 
21: end PARFOR
22: Update  $E$ 
```

particular disease d are identified. From Line 11–Line 16, the tweets are tokenized and it is identified that how many times a user uses disease specific keywords in his/her tweets. Based on the results from Line 11–Line 16, the user keyword matrix is generated in line 17. Line 18 identifies the top candidate experts and line 19 identifies the top experts using the hubs and authorities method. Line 20 selects and returns the required number of top N experts. Line 22 updates the experts list for each disease to respond to the future queries.

An Example of Expert User Identification using the Hubs and Authorities

The task of expert user identification and ranking using the hubs and authorities is explained with the help of an example by capturing the tweets related to diabetes. Suppose $U = \{U_1, U_2, \dots, U_n\}$ and $K = \{K_1, K_2, \dots, K_n\}$ be the sets of candidate expert users and the keywords, respectively. Table 2 presents the candidate expert users based on the frequency of diabetes related keywords in the tweets. The experts were identified on the basis of use of following set of keywords: $\{K_1=\text{Diabetes mellitus}, K_2=\text{Polyuria}, K_3=\text{Polygenic}, K_4=\text{Diabetes}, K_5=\text{Blood glucose}, K_6=\text{Juvenile}\}$. As can be observed from Table 2 that the tweets by users U_2 and U_4

contain only a few keywords and one keyword used by both of the users has high frequency. Despite not using all of the keywords, the row sum values for the keywords used by U_2 and U_4 are sufficiently large. Therefore, according to our supposition users U_2 and U_4 can be considered as the non- doctor experts who only repeat one or a few keywords in the tweets. However, the users U_1 and U_3 are using several keywords pertaining to one disease. To determine the popularity of an expert, the hub and authority based approach instead of only considering the total count of keywords used by an expert relies on both the popularity of the keyword and popularity of the expert. Suppose the initial hubs and authority scores be, $h_d^0 = [1,1,1,1]^T$ and $a_d^0 = [1,1,1,1,1,1]^T$, respectively. By recursively applying the HITS algorithm, the hub and authority scores are updated in each iteration. Table 3 and Table 4 present the hub and authority score, respectively. The algorithm converges at 38-th iteration for the hub score and at the 39-th iteration for the authority score. The hub and authority scores for the first and last iteration are shown in Table 3 and Table 4, respectively.

Table 2: User-keyword matrix

	K ₁	K ₂	K ₃	K ₄	K ₅	K ₆
U ₁	6	1	2	2	6	1
U ₂	-	3	-	10	2	-
U ₃	3	1	2	4	7	-
U ₄	3	-	-	-	-	12

Table 3: Hub score

Iteration No.	U ₁	U ₂	U ₃	U ₄
1	0.281	0.218	0.265	0.234
38	0.275	0.249	0.278	0.196

Table 4: Authority Score

Iteration No.	K ₁	K ₂	K ₃	K ₄	K ₅	K ₆
1	0.197	0.060	0.067	0.235	0.246	0.191
39	0.190	0.065	0.068	0.258	0.254	0.163

It can be observed from Table 3 that the hub score for U_1 in the 1-st iteration has the highest value whereas the users U_3 , U_4 , and U_2 are at 2-nd, 3-rd, and 4-th positions, respectively. However, as we iterate through the HITS algorithm and apply the hub update and authority update rules, the hub scores change in each of the iterations. In 38-th iteration, the hub score of U_4 turns out to be the lowest that actually was 2-nd lowest in the 1-st iteration. The user U_3 having the second highest hub score in 1-st iteration emerges as the user with the highest hub score in 38-th iteration. Similarly, the hubs scores at 38-th iteration for users U_1 and U_2 are the second and third highest, respectively. Table 4 presents the authority score for each of the keywords. It can be observed that K_4 and K_5 gain the position of two keywords having the highest and second highest authority score. It means that both K_4 and K_5 are the most important keywords at the convergence iteration. The hub and authority scores presented in Table 3 and Table 4 sufficiently validate our statement that for being the most popular and the most expert user it is not necessary to use or repeat the popular words only. Instead, it depends on both the importance of the keyword as well as the importance of the users of that keyword. It can be noted from Table 2 that the keywords K_4 and K_5 are among the most popular keywords because

of their high frequencies. On the other hand, Table 3 shows the highest authority scores for K_4 and K_5 ; whereas the authority score for K_1 is the third highest that had low count even than K_6 . The keyword K_4 besides having the higher frequency is also being used by U_1 and U_3 that results in high authority score for K_4 . Interestingly, K_6 that was used twelve times by U_4 has the lowest authority score and the reason is that it is being used by the user with the low hub score. As a whole, the hubs that use good authorities (keywords) and the use of good keywords by the experienced hubs affects the overall ranking score. The expert users that gain high hub scores at the convergence iteration are considered as the doctor experts while the others with low hub scores are identified as the non-doctor experts. In the above example, U_1 and U_3 are accurately identified as the doctor experts whereas U_2 and U_4 are correctly identified as the non-doctor experts. Therefore, depending on whether the query of the enquiring user demands for consultation with the doctor or non-doctor expert, the list of users identified as the hubs can be sorted accordingly to offer the recommendation. In conclusion, the hubs and authority based popularity ranking shows that to derive the importance of the users, merely the excessive use of only one or a few keywords is not necessary. Instead the importance of the keywords and users and the use of several disease specific keywords with reasonably large frequency also affects the overall hub and authority scores.

Moreover, the framework uses caching mechanism to reduce the time consumption for queries requiring expert user identification for the same diseases by multiple users. In other words, the time required for duplicate searches to identify experts is reduced by temporarily storing the results of users' queries in cache. For each user query, the results are cached for a small time and if within that time a user query is received requesting the experts for the same disease, then that query is also responded by selecting the expert from the cached list. This reduces the query response time and also can allow the system to scale better. However, it is also important to mention that overly caching and storing the results for a quite longer period of time may degrade the accuracy and can result in increased demand of resources, such as memory.

3.3 Prototype Implementation

The prototype of the framework is implemented as Software as a Service (SaaS). The SaaS model of cloud permits to host the software as the service that is made available to the customers via Internet [AbB15]. A key benefit of the SaaS model is the significant reduction in Information Technology (IT) costs at the customers' end. The users are relieved of the tasks of infrastructure development and maintenance [AbK14]. Instead the users are charged according to the pay-as-you-go model to access the services. Based on the user query for risk assessment of a particular disease, the framework performs the profile matching of one user with multiple existing users or patients having the similar disease through the collaborative filtering. We conducted the experiments on Ubuntu cloud computing setup comprising of Supermicro SuperServer SYS-7047GR-TRF systems. The end users can access the framework to specify their queries using computers, smartphone, and other handheld devices. The mapping of the proposed framework to the cloud environment is presented in Fig. 5.

It is important to mention that the patients having similar disease profiles are stored together in the framework. Consequently, a particular user query requesting assessment for any disease is only mapped to the patients having profiles similar to that of the enquiring user. For multiple users, the process can be applied simultaneously to multiple user profiles in a parallel manner. The framework also offers a service to help users interact with the disease experts on the Twitter.

To access the tweets from Twitter, we used the twitterR package of R [TwR14]. The framework contains a general tweets repository that is further subdivided into disease specific tweet repositories by matching with the disease specific keywords obtained from the WordNet semantic ontology. The expert users as specified in the query of enquiring user are segregated from the tweet repositories based on the use of disease specific keywords and ranked using the hubs and authorities based approach. All of the above mentioned tasks related to expert user recommendation are preprocessed and are performed in offline mode by executing parallel jobs to avoid the overhead occurring due to real-time processing for time consuming tasks, such as the extraction of tweets from Twitter, processing the tweets to maintain disease specific tweet repositories, and segregation of the expert users. Based on a user query, the preprocessed list of disease specific experts is retrieved and provided to the user. This helps in efficiently responding to the user queries in real-time. Moreover, to ensure that the users are provided the updated information, the task of offline preprocessing is performed periodically to update both the tweet repositories and the lists of experts.

Fig. 5: Cloud service mapping of the proposed framework

4. Results and Discussion

To determine the efficacy of the framework presented in this paper, we conducted experiments to evaluate the disease risk assessment module and the expert user recommendation module. The evaluate results for the two modules are discussed in detail in the proceeding subsections.

4.1 Evaluation of Disease Risk Assessment Module

The performance of the proposed CFDR module was evaluated through comparison with various popular approaches and classifiers, such as the CART, logistic regression, Naïve Bayes classifier, BF decision tree, MLP, Bayesian Network, RF, RoF, and the approach presented in [YuL10]. The brief description of each of the related techniques used for comparison is presented below.

Classification and Regression Tree (CART): The CART is a tree based model for classification that uses the cross-validation for the selection of appropriate tree [HeE08]. The method works by recursively partitioning the data space where each partition can be represented

as a decision tree. The CART based approaches have been applied on various clinical and demographics variables for classification purposes.

Logistic Regression: Logistic regression is a standard classification method widely used for disease risk prediction. The outcomes in logistic regression are the class labels based on multiple features or predictors [LiC07].

Naïve Bayes: Naïve Bayes uses the strong attribute independence assumption and is used to develop models with high predictive capabilities [PaA08]. The conditional independence of attributes greatly minimizes the computation cost. The posterior probability of occurring of X given each C_i is calculated as in Eq. 8.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (8)$$

Best First (BF) tree: The BF tree expands the nodes in best-first order. The node that maximally minimizes the impurity is considered as the best node and is included in the decision tree [Shi14]. An attribute from all the context attributes is selected and the branches are made based on some predefined criteria. The training object pairs extending from the root node are split into subsets. The aforementioned process is repeated for a chosen branch of tree till a specific number of expansions of the tree.

Bayes Net: The Bayesian Network classifier is a probabilistic model that characterizes a set of random variables and their conditional dependence upon each other through a Directed Acyclic Graph (DAG) [Bou08]. The Bayesian Networks are used to represent the relationship between the symptoms and diseases that are subsequently used to compute the probability of occurrence of a disease.

Multilayer Perceptron (MLP): The MLP is class of supervised neural networks that is frequently used in medical decision support systems for diagnoses. The multilayer perceptron comprises of at least three or more layers of nodes, namely the input layer, hidden layer, and the output layer [YaJ06]. For the input received at the input layer, processing is performed at the successive layers till the output is received at the output layers.

Random Forest (RF): The RF is an ensemble learner capable of generating several classifiers and then integrating their results. The RF creates multiple CART trees and each of them is trained on a bootstrap sample of the original training dataset [KhC11]. Each of the trees in RF casts the vote for certain input and the classifier output is subsequently computed by majority voting.

Rotation Forest (RoF): The RoF is a relatively new ensemble classifier for feature extraction and is capable of transforming the dataset while preserving all of the information using the Principle Component Analysis (PCA) [RoK06]. By rotating the subspaces of the original dataset, the classifiers with features are constructed. In addition we, also compared the result of the proposed CFDRA approach with the approach Support Vector Machine (SVM) based approach presented in [YuL10].

The NHANES (2009-2010) [NaH14] survey data was used for comparison of the CFDRA with the above mentioned approaches. The NHANES is a program of study for health and nutrition status assessment of children and adults in the United States. The reason to use NHANES 2009-2010 dataset is that it encompasses the life styles of the population of the U.S. with sufficiently large amounts of data on demographics, diet, examination, and laboratory reports. We conducted experiments to make risk assessment for diabetes. The variables that we

used to perform the risk assessment for diabetes include “age”, “gender”, “ethnicity/race”, “height”, “weight”, “ever diagnosed high blood sugar or pre-diabetes”, “diabetes family history”, “physical activity”, “ever observed high blood pressure”, “blood cholesterol”, “smoking”, and “ever diagnosed diabetes”.

The data of over 5,000 users with the ages ranging from 18-years to 80-years was collected. The dataset was evaluated using the *k-fold* cross validation with $k=10$. The cross validation is typically a method used to estimate the predictive capability of a model [PoL12]. The dataset is divided into *k-folds*, where one fold is used as the testing fold while the remaining $k-1$ folds are used as the training folds. Repeating the process k -times ensures that all of the examples both from the training and testing data are used for analysis. To evaluate the performance of the CFDRA approach with the other approaches, the common model evaluation metrics, such as the precision, recall, and F-measure [BeS12] were used.

Precision is the ratio of correct (True Positives) disease predictions regarding the presence or absence of a disease to the total number of occurrences of disease (True Positive (TP) + False Positive (FP)), given as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad , \quad (9)$$

Recall is defined as the ratio of correctly identified patients to the total size of testing set. In other words, recall is the probability of identification of a randomly selected user profile in the set and is given as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad , \quad (10)$$

where FN stands for False Negative.

F-measure uses both the precision and recall and is the harmonic mean of precision and recall values and is given as:

$$F - \text{measure} = \frac{2TP}{2TP + TP + FN} \quad , \quad (11)$$

The approach was evaluated by testing the accuracy against the values of the attribute “ever diagnosed diabetes” (YES or NO) in the dataset. The “YES” and “NO” respectively represent that the person is either a diabetic patient or not a patient. Fig. 6 presents the comparison results for the case “YES” when the test patients had diabetes, whereas the comparison results for the case “NO” are presented in in Fig.7. The SVM based approach presented in [YuL10] is depicted as “SVM” in Fig. 6 and Fig. 7.

The reason to evaluate the algorithms for both types of aforementioned data is that estimating the algorithm on only one type of examples (YES or NO) does not accurately predict the presence or absence of a disease. A good prediction technique should identify both the patients and healthy individuals with higher accuracy. As can be observed from Fig. 6 and Fig. 7 that the CFDRA approach achieved high precision, recall, and F-measure as compared to other approaches. The other approaches, such as the BF tree, RoF, SVM, Naïve Bayes, and MLP also exhibited reasonably good results. However, logistic regression and the RF turned low in terms of accuracy. The results by the logistic regression, Naïve Bayes, RF, and RoF were more dependent on the attribute “disease family history” while the attributes “height” and “weight” did not have any significant effect on the prediction. On the other hand, in CFDRA, the attribute “ever diagnosed high blood sugar or pre-diabetes” was observed as the most important attribute due to the high count of negative (No) responses by the users. In conclusion, our approach of identifying one important attribute first and then retrieving the profiles on the basis of that

attribute not only achieves high accuracy but also is computationally efficient because of the smaller datasets.

Fig. 6: Comparison of the proposed CFDR approach with the related approaches for case (YES)

Figure 7: Comparison of the proposed CFDR approach with the related approaches for case (NO)

4.2 Evaluation of Expert User Recommendation Module

To evaluate the performance of the expert user recommendation module, we collected around 3,363 profiles (doctors and non-doctors) of Twitter users containing a total of 396,655 tweets by using the keywords related to the disease “diabetes”. Downloading the tweets using Twitter API is restricted by the rate limits that eventually requires large amount of time to collect the tweets. Therefore, the task was performed offline by executing periodic jobs. The keywords presented in Table 5 were used by using the WordNet dictionary to retrieve the tweets.

Table 5: WordNet keywords used to retrieve tweets

Diabetes Specific Terms Used				
Diabetes	Pre-diabetes	Insulin	Blood sugar	Blood glucose
Metformin	Diabetes mellitus	Type 1 diabetes	Type 2 diabetes	Metabolic disorder
Polygenic disorder	Ketogenic	Insulin dependent diabetes	Insulin independent diabetes	Polydipsia
Polyuria	Adult onset diabetes	Diabetes insipidus	Ketosis resistant diabetes	Hypoglycemic agents
Nephrogenic diabetes insipidus	Juvenile diabetes	Ketoacidosis-prone diabetes	Episodic ketoacidosis	Autoimmune diabetes

Around 3% of the user profiles were manually identified and flagged as medical doctors due to their self-claim as doctor on their Twitter profile. The recommended lists of doctors as a results of applying the hubs and authority based approach were compared with the profiles that were manually collected. The total number of TP, TN, FP, and FN were determined and on the basis of those the precision, recall, and F-measure scores were calculated. Moreover, the hub and authority based approach to identify and rank the experts was compared with the popularity based approach using the row sum method and the approaches presented in [PaC11] and [ChC14]. The approach presented in [PaC11] identifies the topical authorities in microblogs by using the features, such as the topical signals and mention impacts of the users for calculating the ranked lists. The approach presented in [ChC14] identifies the expert users by calculating their topical expertise. Each technique is executed 20 times and their average results about precision, recall, and F-measure are shown in Fig. 8, Fig.9, and Fig. 10. It can be observed that the values for precision, recall, and F-measure for the proposed approach termed as Expert User Recommendation Module (EUR) in Fig. 8, Fig. 9, and Fig. 10 are higher than the compared approaches for *Top-k* experts, where $k = (5, 10, 15, 20)$.

Moreover, the results for precision, recall, and F-measure for the proposed approach are significantly higher than the compared approaches even for large values of k (for example, $k=15$ and $k=20$). Among the three compared approaches, the approach proposed in [ChC14] performed substantially better than the other two approaches. However, the accuracy of the popularity based approach using the row sum method was significantly low. This testifies the efficacy of the proposed hubs and authorities based approach that segregates the expert users based on the use of several important keywords by the popular experts.

4.3 Complexity Analysis

The complexity analysis of the algorithms for the disease risk assessment and expert user recommendation are presented in this section. Algorithm 1 presents the steps used for disease risk assessment. Line 1–Line 4 of algorithm 1 takes $O(n \times a)$, where n represents the number of profiles and a is the number of profile attributes. The operation at Line 5 takes $O(n)$ to execute. Execution of either of Line 8 and Line 10 takes $O(n)$. Each of the Line 12–Line 14 executes in

$O(n)$. Line 15 calculates the risk assessment score and also has complexity $O(n)$. The overall complexity from Line 6–Line 16 will be $O(Q \times n)$, where Q is the set of enquiring users.

Fig. 8: Comparison of the Precision of the proposed EUR approach with related approaches

Fig. 9: Comparison of the Recall of the proposed EUR approach with related approaches

The total complexity becomes $O((n \times a) + (Q \times n))$. Because a is very small as compared to n , therefore, the complexity in worst case is considered as $O(Q \times n)$. Moreover, the parallel execution of algorithm further results in the decrease in complexity, which is given as $O((Q \times n)/p)$, where p represents the number of processors used for computations.

Algorithm 2 presents the steps to identify and rank the expert users from the Twitter using the hubs and authorities based method. Line 2 of Algorithm 2 executes in $O(k)$, where k is the number of keywords. Line 3–Line 6 search the repositories and have complexity $O(T \times k)$, where T represents the tweets. The operations in Line 7–Line 10 extract the users based on the use of keywords and have combined complexity of $O(\partial \times T \times k) = O(U)$. In other words, it is the worst case complexity for extracting all the possible users from the database. Line 11–Line 16 execute in $O(U \times x \times k)$, where x be the number of tokens. Line 17 and Line 18 execute in

$O(U \times k)$ and $O(U)$, respectively. Line 19 takes $O(n \times (U^2 + k^2))$ to identify and subsequently rank the users. The total complexity of Algorithm 2 for d diseases becomes $O(d \times ((T \times k) + (U \times x \times k) + (n \times (U^2 + k^2))))$.

Figure 10: Comparison of the F-measure of the proposed EUR approach with related approaches

4.4 Scalability Analysis

We also evaluated the performance of the framework in terms of scalability. An algorithm is scalable if by increasing the resources, such as the number of processors, the efficiency of the algorithm does not decrease significantly [AhA11]. In other words, with the increase in workload the processing time should be maintained within desirable limits by increasing the number of resources, such as the processors. The elasticity or scalability of the cloud permits the on-demand procurement of resources. Amazon Elastic Compute Cloud (EC2) [AmE14], the commercial cloud service provider, also provides the resources, for example, the processors, memory, and storage on the basis of prices that the consumers are willing to pay. Therefore, we observed the effects of varying the number of processors and the data sizes on execution time because it is the most critical factor that determines the efficiency of the proposed framework in terms of query response time. Fig. 11 presents the effects of increasing the number of processors and the number of user profiles on the execution time for the disease risk assessment module. For the disease risk assessment module, increasing one processor results in decrease in the execution time by 12.69 % on an average, whereas doubling the amount of data increases the execution time by 28.97 % on an average.

Fig. 12 also presents the effects of increasing the number of processors and the data size on execution time for the expert user recommendation module. It can be observed from Fig. 12 that the execution time increases significantly with the increase in data size. However, increasing the number of processors results in minimizing the execution time. With the increase of one processor, the execution time decreases by 7.15 % on an average, whereas doubling the amount of data increases the execution time by 9.01 % on an average. For both of the modules, relatively small decrease in time consumption were observed when the number of processors was increased over six. The apparent reason for increase in time consumption is that the overheads, such as the processor startup time, and inter-processor communication time also contribute to the total time consumption [AhA11]. Therefore, increasing the additional number of processors results in

increased overheads that contribute to the increased response time. The offline processing time per query for the expert user recommendation module is still very high. Therefore, our cloud based framework periodically executes the jobs in offline mode to collect the tweets from the Twitter, evaluates the tweets based on the disease specific keywords, updates the disease specific tweet repositories, and identifies and subsequently ranks the experts. A user query requesting a recommendation about the experts is responded by returning the expert users identified during the offline processing. This results in minimal response time against a query because all of the compute-intensive tasks are already preprocessed by the cloud using Algorithm 2.

Figure 11: Relationship between the processing time, no. of processors, and data size for CFDRA

Figure 12: Relationship between the processing time, no. of processors, and data size for EUR

Moreover, to give a better insight about the performance of each of the modules in terms of scalability, we calculated the number of Transactions Per Second (TPS) per processor. This analysis helps in determining the ability of the framework to handle the TPS per processor. For each of the disease risk assessment module and the expert user recommendation module, we define the number of transactions differently. For the disease risk assessment module, the number of existing users' profiles that the framework is able to compare per second is considered

as the TPS. Likewise, the amount of data size in MBs per second is the TPS for the expert user recommendation module. Fig. 13 and Fig. 14 present the analysis according to the number of TPS per processor for workloads of different sizes for risk assessment module and expert recommendation module, respectively.

Figure 13: Transactions per second per processor for the CFDRA approach

Figure 14: Transactions per second per processor for the EUR approach

5. Related work

The proposed framework introduces: **(a)** a disease risk assessment mechanism and **(b)** an approach that finds the health experts available on Twitter. Therefore, in this section, various proposals are discussed that are relevant to the proposed framework with respect to the two aforementioned aspects.

Khalilia *et al.* [KhC11] employed a Random Forest (RF) based approach for disease prediction. The approach takes into account the diagnosis history of the individuals on a highly imbalanced dataset and combines the RF method with the repeated random sub-sampling. The approach claims to be achieving high prediction accuracy in comparison to several other machine learning approaches. However, a limitation of the RF method is that it comes across the issue of overfitting with noisy datasets that degrades the accuracy for different datasets. On the other hand, we used collaborative filtering to perform disease risk assessment by computing the similarities between the profile of the enquiring user and the existing users having similar diseases. Yu *et al.* [YuL10] used Support Vector Machine (SVM) based approach to develop classification models for persons with diagnosed or undiagnosed diabetes. The scheme is claimed to be the first ever used to diagnose the common disease without the laboratory tests. However, the SVM based approaches are uncertain about the selection of kernel function and also require large memory and computational resources. Conversely, our approach reduces the size of dataset by retrieving profiles based on one influential attribute that eventually results in minimizing the computation time. The authors in [KhM10] used fuzzy set theory to make risk assessment for coronary heart disease. However, the fuzzy modeling approaches are limited in handling diversity of medical data. The authors in [WeV11] used Naïve Bayes approach to make risk assessment for Alzheimer disease using genomic driven data. Nonetheless, the conditional

independence assumption of the attributes in Naïve Bayes approach affects the posterior probability estimate for risk assessment. The CFDRA approach on the other hand uses the Cosine Similarity method to compute similarities between the profiles of enquiring users and the existing users. The similarity scores are used to calculate the risk assessment scores for the enquiring users. Moreover, the aforementioned discussed works only make risk assessments for only single disease whereas the proposed CFDRA approach has the capability to make risk assessments for multiple diseases simultaneously and in an efficient manner.

Apart from the disease risk assessment, another important dimension of the proposed work is to find health experts from Twitter. A lot of research has been conducted on identifying the experts in various online communities. However, the studies focusing on finding the expert users from online health communities have been very negligible. Zhao *et al.* [ZhY14] proposed an approach to find influential users in online health communities by estimating the emotional support through text mining and sentiment analysis. The approach utilizes an influence model of social network theories where dynamics of social influence are characterized using a diffusion model. The authors introduced a metric called Influential Responding Replies (IRR) to determine influence of other members. However, the approach is limited in offering interaction with only the patients of the online health community. On the other hand, our proposed approach enables the users to interact with both the doctor and non-doctor experts by using the hub and authority based approach. Moreover, our proposed approach ranks the experts based on the use of health related keywords by experts instead of replies by the users. The authors in [PaC11] proposed an approach to find the topical authorities in microblogs. The authors exhibited the efficacy of the probabilistic clustering for selection of high authority users and also proved the effectiveness of Gaussian-based ranking to rank the users. Ghosh *et al.* [GhS12] used Twitter lists to mine the topical experts. The approach in [GhS12] utilizes the crowdsourced annotation of topical experts and suggests experts that might have knowledge to answer the questions. Moreover, the approach in [GhS12] manually curates the Twitter lists to identify and rank the experts. Our approach on the other hand periodically extracts the tweets from Twitter, preprocesses the tweets, identifies the candidate experts, and then segregates the experts into doctors and non-doctors using the hub and authority based approach. The approach presented in [ChC14] identifies the local experts by calculating their topical expertise based on expertise propagation in geo-tagged social connections on Twitter. The approach considers those individuals as the local experts that are well recognized in a community based on the views of others. However, our approach identifies the experts based on their tweets and the use of disease related terminologies. Moreover, our proposed approach uses cloud computing services to process large repositories of tweets data.

5.1 Significant Differences of the Framework from the Discussed Related Works

The differences of the proposed framework from the approaches discussed in the related work are presented below. Each of the discussed approaches for disease risk assessment makes risk assessments for only one disease. However, the proposed CFDRA approach is capable of making risk assessments for multiple diseases to support multiple users' queries simultaneously. The framework maintains the profiles of patients of different diseases separately and based on the users' queries, retrieves the profiles of the patients of the disease as specified in the user query. In addition, the framework achieves high accuracy because it utilizes the similarity scores to predict the risks. Moreover, the CFDRA approach reduces the dataset size by retrieving the

profiles of existing users for subsequent comparisons based on the influential attributes. The expert user recommendation module uses hub and authority based approach and also identifies both the doctors and non-doctor experts. The module instead of relying only on the popularity of the keywords considers the importance of keywords and the importance of users who use the keywords. Moreover, the tweets are extracted continuously so that tweet repositories are frequently updated. Moreover, to reduce the query response time, we perform the compute-intensive operations offline. The presented framework can be a useful resource to offer healthcare services through Web at no cost.

6. Conclusions and Future Work

In this paper, we presented a cloud based framework that enables the Web and mobile users to make risk assessments about probable diseases. We employed a collaborative filtering based approach for disease risk assessment that computes similarities between the profiles of enquiring users and the existing users. The results of proposed disease risk assessment approach were compared to various approaches and classifiers, such as the CART, Naive Bayes, logistic regression, MLP, BF-tree, RF, RoF, SVM, and Bayesian Network. The accuracy of the proposed approach was found significantly higher than the approaches used for comparison. Moreover, we also proposed an approach that utilizes Twitter data to offer users an opportunity to interact with the health experts for consultation. By observing the tweets related to health, we identified the health experts and ranked them by using the concept of hubs and authorities. The comparison of the approach with the state-of-the-art approaches shows significant improvements in terms of accuracy. We are optimistic that the proposed framework will prove as the basis for the researchers to combine the predictive modeling approaches and the social media networks to develop collaborative health communities where the patients can obtain health information and seek advice from the experts without any cost.

We plan to extend the existing work by mining the tweets for diseases based on the geographical locations of the users. The geographical locations will help to understand the current spread of diseases and to identify and recommend the doctors based on the diseases in specific area. Recommending the doctors to the users belonging to the same geographical region can help the individuals or patients to contact the doctors physically as well, if required. In addition, another important open issue worth exploring is identification of fake user profiles from Twitter. Several machine learning-based, graph theory-based, and honeypot harvesting approaches have been proposed recently for the said purpose [GuW15], [ThM13], and [LeD11]. The techniques collect the users' behaviors through tweet patterns and classify them as genuine or fake. Likewise identification of fake profiles through analysis of tweet contents, the number of hostages per word, reputation scores, number of duplicate tweets, or number of URLs per tweet has also been performed [Wan10]. Integrating the approaches employed in the above mentioned works to identify fake users with the proposed framework will certainly enhance the reliability and accuracy of the system.

Acknowledgments: The authors are thankful to Ahmad Fayyaz and Muhammad Jawad for the valuable reviews, suggestions, and comments.

References

- [AbB15] A. Abbas, K. Bilal, L. Zhang, and S. U. Khan, "A cloud based health insurance plan recommendation system: A user centered approach," *Future Generation Computer Systems*, vols. 43-44, 2015, pp. 99-109.
- [AbK14] A. Abbas and S. U. Khan, "A Review on the State-of-the-Art Privacy Preserving Approaches in E-Health Clouds," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, 2014, pp. 1431-1441.
- [AbZ15] A. Abbas, L. Zhang, and S. U. Khan, "A Survey on Context-aware Recommender Systems Based on Computational Intelligence Techniques," *Computing*, vol. 97, no. 7, 2015, pp. 667-690.
- [AhA11] M. Ahmed, I. Ahmad, and S. U. Khan, "A Theoretical Analysis of Scalability of the Parallel Genome Assembly Algorithms," in *IEEE/EMB/ESEM/BMES International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS)*, Rome, Italy, January 2011, pp. 234-237.
- [AmE14] "Amazon EC2 Pricing," <http://aws.amazon.com/ec2/pricing/>, accessed on October 12, 2014.
- [BaH13] M. A. Barrett, O. Humblet, R. A. Hiatt, and N. E. Adler, "Big data and disease prevention: From quantified self to quantified communities," *Big Data* 1, no. 3, 2013, pp. 168-175.
- [BeS12] P. Bedi and R. Sharma, "Trust based recommender system using ant colony for trust computation," *Expert Systems with Applications*, vol. 39, no. 1, 2012, pp. 1183-1190.
- [Bou08] R. R. Bouckaert, "Bayesian network classifiers in weka for version 3-5-7," *Artificial Intelligence Tools*, vol. 11, no. 3, 2008, pp. 369-38.
- [BuC10] J. Bughin, M. Chui, and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," *McKinsey Quarterly* 56, no. 1, 2010, pp. 75-86.
- [ChC14] Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, *Who is the Barbecue King of Texas? A Geo-Spatial Approach to Finding Local Experts on Twitter*, http://faculty.cse.tamu.edu/caverlee/pubs/cheng_sigir14.pdf, accessed on October 14, 2014.
- [ChD13] N. V. Chawla, and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of general internal medicine* 28, no. 3, 2013, pp. 660-665.
- [EaK10] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University, Press, 2010.
- [FoD13] S. Fox, M. Duggan, "Health online 2013," http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf, accessed on September 1, 2014.
- [GhS12] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: crowdsourcing search for topic experts in microblogs," In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 575-590.
- [GuW15] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach," In *Proceedings of the 2015 ACM International Conference on Social Media & Society*, 2015, pp. 9.
- [HeS14] "Healthcare Social Media Analytics," <http://www.symplur.com/healthcare-social-media-analytics/>, accessed on September 18, 2014.
- [HeE08] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes Risk Calculator A simple tool for detecting undiagnosed diabetes and pre-diabetes," *Diabetes Care* 31, no. 5, 2008, pp. 1040-1045.
- [KaK13] B. Kayyali, D. Knott, and S. V. Kuiken, "The big-data revolution in US health care: Accelerating value and innovation," *Mc Kinsey & Company*, 2013, pp. 1-13.

- [Kha14] S. U. Khan, "Elements of Cloud Adoption," *IEEE Cloud Computing Magazine*, vol. 1, no. 1, 2014, pp. 71-73.
- [KhC11] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making* 11, no. 1, pp. 2011, pp. 51
- [KhO14] A. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A Survey of Mobile Cloud Computing Application Models," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, 2014, pp. 393-413.
- [KhM10] V. Khatibi, and G. A. Montazer, "A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment," *Expert Systems with Applications* 37, no. 12, 2010, pp. 8536-8542.
- [LeD11] K. Lee, B. David Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," In *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*, 2011, pp. 1-8.
- [LiC07] J. G. Liao, and K. -V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics* 23, no. 15, 2007, pp.1945-1951.
- [LiT03] J. Lindström, Jaana, and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," *Diabetes care* 26, no. 3, 2003, pp. 725-731.
- [Mil95] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM* 38, no. 11, 1995, pp. 39-41.
- [Mil12] K. Mille, "Big Data Analytics in Biomedical Research," *Biomedical Computation Review*, 2012, pp. 14-21.
- [NaH14] "National Health and Nutrition Examination Survey," http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx, accessed on September 29, 2014.
- [PaA08] S. Palaniappan, and R. Awang, "Intelligent heart disease prediction system using data mining techniques," In *IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 108-115.
- [PaC11] A. Pal, and S. Counts, "Identifying topical authorities in microblogs," In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 45-54.
- [PoL12] C. Porcel, A. T. Lorente, M. A. Martínez, and E. H. Viedma, "A hybrid recommender system for the selective dissemination of research resources in a Technology Transfer office," *Information Sciences* 184, no. 1, 2012, pp. 1-19.
- [RoK06] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, no. 10 2006, pp. 1619-1630.
- [SaK01] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," In *Proceedings of the 10th ACM international conference on World Wide Web*, 2001, pp. 285-295.
- [Shi14] H. Shi, "Best-first decision tree learning," <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=966CB8346D86A79ECF45344B4CB76D82?doi=10.1.1.149.2862&rep=rep1&type=pdf>, accessed on October 8, 2014.
- [ThM13] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and Vern Paxson, "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse," In *USENIX Security*, 2013, pp. 195-210.
- [Twi14] "Twitter", <http://www.twitter.com>, accessed on September 13, 2014.

- [TwR14] "twitteR: R based Twitter client," <http://cran.r-project.org/web/packages/twitteR/index.html>, accessed on October1, 2014.
- [Wan10] A.H. Wang, "Don't follow me: Spam detection in Twitter," In *Proceedings of the 2010 IEEE International Conference on Security and Cryptography (SECRYPT)*, 2010, pp. 1-10.
- [WeV11] W. Wei, S. Visweswaran, and G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data," *Journal of the American Medical Informatics Association* 18, no. 4, 2011, pp. 370-375.
- [YaJ06] 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* 30, no. 2, 2006, pp. 272-281.
- [YuL10] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, 2010, pp. 16
- [ZhY14] K. Zhao, J. Yen, G. Greer, B. Qiu, P. Mitra, and K. Portier, "Finding influential users of online health communities: a new metric based on sentiment influence," *Journal of the American Medical Informatics Association*, 2014, pp. 1-7.