A Cloud Based Framework for Identification of Influential Health Experts from Twitter

Assad Abbas[1], Muhammad U. S. Khan[1], Mazhar Ali[2], Samee U. Khan[1], and Laurence T. Yang[3]
[1]North Dakota State University, Fargo, ND, USA
e-mail: {assad.abbas, ushahid.khan, samee.khan}@ndsu.edu
[2]COMSATS Institute of Information Technology, Abbottabad, Pakistan
e-mail: mazhar@ciit.net.pk
[3]St. Francis Xavier University Antigonish, NS, Canada
e-mail: ltyang@stfx.ca

Abstract— **The ever increasing growth in the health related data has necessitated the development of pervasive tools and technologies to manage the huge data volumes. Likewise, the conventional healthcare services are transforming into patient-centric services to offer ubiquitous access to the health related information. However, there is a need to extend the capabilities of the existing health services and tools so that users could become aware about their health, devise wellness plans, and seek experts' advice at no or low cost using the social media. In this paper, we propose a cloud based framework that uses Twitter data to offer recommendations about the most influential health experts. We employ a variant of the Hyperlink-Induced Topic Search (HITS) approach to identify the candidate health experts based on the health related keywords used in the tweets. Subsequently, we propose an influence metric that calculates the influence of the candidate experts based on various parameters. The proposed approach attained high accuracy when compared to other approaches for expert user identification. Moreover, experimental results exhibit that the approach is highly scalable for workloads of varying sizes.**

Keywords- Expert users, influence, cloud computing, tweets

I. INTRODUCTION

The increased demand for utilizing the electronic healthcare services has resulted in enormous growth of health data on the Internet [1]. The current volumes of health related data include the diagnosis and prescriptions data, laboratory data, pharmacy records, health insurance claims data, gene extraction and sequencing data [2], [3]. Consequently, such high volumes of diverse data give rise to the term health Big-data [2]. Besides the aforementioned sources of health related data, online health communities and social media platforms, such as Twitter and Facebook have also appeared as the rich sources of health data. Users of social media websites discuss various topics of common interests including the health issues. Twitter, for example is being used by both the doctors and the patients to exchange their experiences and feelings with others. Moreover, Twitter contains various health communities that are meant to answer and resolve the concerns of the users or patients of different diseases. There are also other online health communities, for example PatientsLikeMe [4] that people use to exchange their experiences against different diseases and to seek support from the other patients. According to the Pew Internet survey of 2013, a key benefit of such Internet based health communities is to help people seek advice from the current or past patients and health experts at no cost [5]. The increased trends of finding online health information are due to growth in the numbers of smartphone users. The Pew Internet survey of 2014 reveals that around 90% of the U.S. adults own mobile phones and approximately 72% of the users have searched online for the information pertaining to health related issues [6]. Considering the widespread use of computing and mobile devices for searching the health related information from the online health communities and social media networks, it is the appropriate time to enhance the potential of the online health communities. Therefore, developing methodologies that enable the interaction between patients and health experts through social media will help users seek advice at low or no cost.

The enormous amount of health data requires scalable solutions to efficiently process and store large amounts of data. The scalability issues of the traditional Web based systems not only result in inefficient processing but also affect the accuracy [7]. Therefore, using the cloud computing based scalable and elastic services to manage large amounts of health data is important to offer efficient and scalable services [8]. Besides the performance benefits of the cloud services in healthcare, financial advantages are also of paramount significance that can help in reduction of healthcare costs [9]. A 2013 survey conducted by McKinsey shows that the healthcare expenses of the U.S are approximately 17.6% of the total GDP [10]. Therefore, the cloud computing services are an inexpensive alternative to deliver the quality healthcare services. Moreover, it is expected that in near future more and more computing and mobile devices will generate gigantic volumes of health data that calls for the use of Big-data analytics in the healthcare domain.

In this paper, we propose a cloud based scalable framework that supports both the desktop and mobile users to seek advice related to health matters from the health experts who frequently use Twitter. The framework analyzes the tweets related to different diseases by various doctors and determines the most suitable health experts for a particular disease in that geographical area. Twitter has emerged as vibrant health information source containing more than

784,893,181 health related tweets, around 10,000 doctors and over 6,200 healthcare communities [11]. The aforementioned figures are evidence of the increased use of Twitter for health related issues that enables the quick information exchange without cost. The framework mainly comprises of two modules: **(a)** candidate experts identification module and **(b)** influential user identification module. The candidate experts are identified by using a variant of Hyperlink-Induced Topic Search (HITS) [12]. Subsequently, the candidate experts are further analyzed to determine the influential experts for a disease. The influential users are identified according to the prioritized criteria indicated in the query of the querying user. The users can find the influential health experts based on multiple criteria, such as: **(a)** number of followers of the expert, **(b)** health related tweets by the expert, **(c)** analyzing the followers' sentiments in replies to the tweets by expert, and **(d)** the retweets of the experts' tweets. The rationale for offering multiple selection criteria is that only one criterion cannot be a true characterization of the expertise of an individual. For example, the following relationship on Twitter is slight casual where some individuals might just randomly follow others who in courtesy can follow them back. Therefore, the reciprocity of the following relationship is not a strong indicator of an individual's expertise [13]. Our framework exhibits great potential to turn the Twitter into a collaborative online health community where people can discuss their health matters with the experts without any cost.

The framework performs the identification of multiple influential users simultaneously across different geographical locations. Maintaining large tweet repositories requires scalable infrastructure with massive storage and efficient processing. Therefore, cloud computing services are utilized because of their ability to dynamically scale up and scale down according to the workload characteristics. The framework executes the periodic jobs to update and maintain tweet repositories and to subsequently identify the health experts. The reason to perform the offline processing for identification of candidate experts and the influential users is that it may incur high time overheads if the processing is performed online. Therefore, offline processing avoids the limitations of online processing. The key contributions of the paper are as follows:

- We present a scalable framework that utilizes the cloud computing services to identify the influential health experts from Twitter.
- A variant of HITS approach is employed to identify the candidate health experts based on the health related keywords in their tweets.
- We also propose an influence metric that calculates the influence of the experts in terms of the number of followers, sentiment analysis of the replies to the tweets by followers, health related tweets, and the retweets to the experts' tweets.
- The framework is capable of managing multiple queries simultaneously by executing parallel jobs to identify the experts from different geographical areas.
- We also demonstrate the scalability of the framework for workloads of different sizes.

The paper is organized as follows. Section II discusses the related work. The architecture of the proposed system is presented in Section III. Section IV presents results and discussion whereas Section V concludes the paper.

II. Related Work

There has been plentiful research conducted on expert identification from various online communities and microblog systems. However, identification of the experts from online health communities has not been very significant. An approach to find influential users in online health communities is proposed by Zhao *et al*. [14]. The approach determines the influence of a user through sentiment dynamics in the threaded community discussions and introduces a metric called Influential Responding Replies (IRR) to determine the influence of others in the community. Weng *et al*. [13] proposed an extension of the PageRank algorithm called the TwitterRank that finds the influential users on Twitter. TwitterRank uses link structures and topical similarities to compute ranking for the influential users on a particular topic. The aforementioned approaches come across the scalability issues whereas our approach is capable of finding the influential users by executing parallel jobs from huge tweets corpus.

Another approach that utilizes probabilistic clustering to identify the topical authorities from the microblogs is presented in [15]. Ghosh *et al*. [16] proposed a crowdsourcing based approach called Cognos to identify the influence of the users by using the Twitter lists. However, the approach in [16] is restricted by hourly restrictions for tweets extraction. A link analysis based approach to identify the influential users from an online healthcare social network is proposed in [17]. The approach quantifies the influence of the users in a small social network. On the contrary, our proposed approach is two-fold that first identifies the candidate experts through a variant of HITS methodology and subsequently determines the influence of the experts based on multiple criteria, such as the follower, number of tweets, sentiments, and retweets.

III. Proposed System Architecture

The proposed framework utilizes the cloud computing services to identify the health experts from Twitter that best match the users' queries. The Software as a Service (SaaS) implementation of the framework allows the availability of the health expert recommendation service by means of Internet. The tweets repositories are maintained by periodically executing the jobs to retrieve the tweets from Twitter. To identify the expert users, the following tasks are performed: **(a)** identification of candidate experts and **(b)** calculation of influential users. The architecture of the proposed framework is presented in Fig. 1. The steps to identify the experts are presented in Algorithm 1.

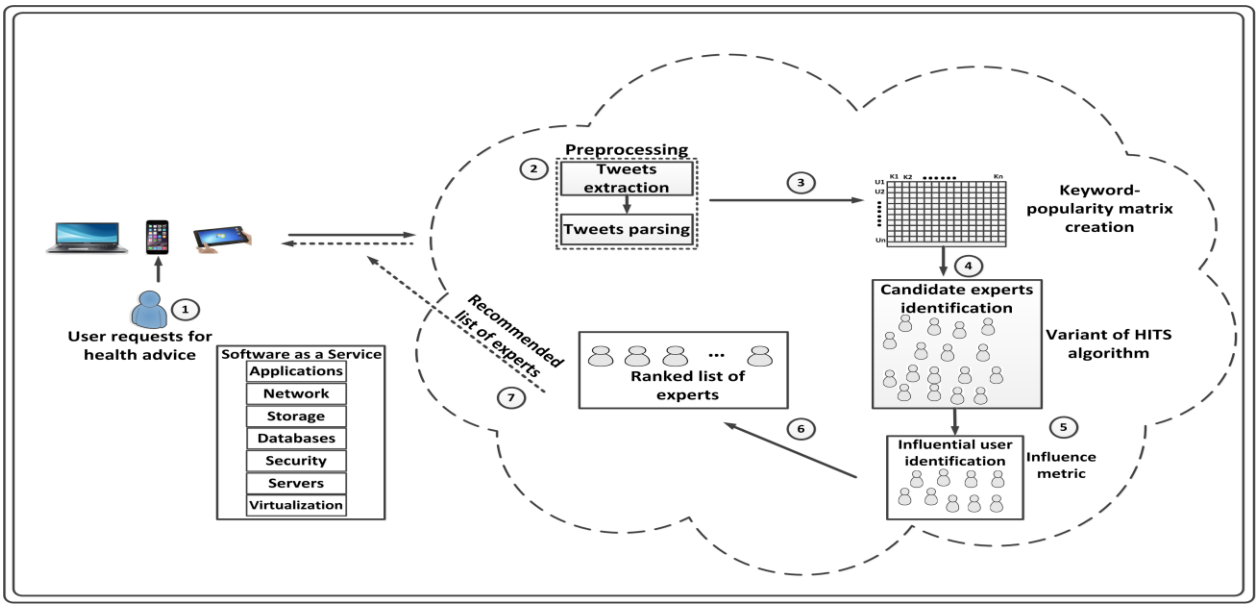

Fig. 1. Architecture of proposed cloud based framework

A. Identification of Candidate Experts

Based on a user query, the tweets from the health experts are analyzed and parsed to extract the disease specific keywords. For the disease specific terminologies to analyze the tweets, we used the WordNet database [18]. The benefit of using WordNet is that it is capable of identifying the relationships between different keywords by using the hypernym, hyponym, meronym, holonym, and derivationally related terms [19]. Interested readers are encouraged to consult [18] and [19] for more details on the hypernym, hyponym, meronym, holonym, and derivationally related terms. Based on the frequency of health related keywords by the health experts in their tweets, a keyword popularity matrix is generated. The set of users U for a particular disease d is represented as below:

$$U_i{}^d = \sum_{j \in J} K_{ij}{}^d, \qquad (1)$$

where K_{ij} is the j-th keyword used by the user i for a particular disease d. However, the popularity of the health experts based on the keywords count is not a true depiction of the real health experts because it only considers the total number of keywords in the tweets used by a particular user. Consequently, the users who frequently repeat a few keywords in tweets may emerge as the top experts. Therefore, to accurately identify the health experts, it is essential to consider the frequency of keywords, importance of keywords, and the importance of the particular experts who use the keywords. To this end, we use a variant of the hubs and authorities based approach to identify the candidate expert users. The concept of hubs and authorities is based on a Hyperlink-Induced Topic Search (HITS) approach that has been used in Web search such that the page that points to several other pages is called hub whereas the pages that are pointed to by several other pages are called authorities [12]. The proposed framework considers the health experts as the hubs and the keywords as the authorities. An issue with the HITS approach is that the good hubs point mostly to the good authorities. Therefore, the ranking decisions using the HITS for experts are mostly based on the frequency of keywords used by important experts. However, there are multiple parameters that contribute for identification of good hubs. The parameters include the usage of multiple different keywords by an expert, importance (frequency) of the particular keywords, and the importance of the hubs using those keywords. Therefore, we modify the HITS approach by multiplying the hub scores with the number of distinctive authorities pointed by the hubs. Consequently, the final ranking score for the hubs is more balanced and is not dependent merely on the frequency of keywords. To identify the candidate experts for a particular disease d, we construct a matrix A with U rows and V columns. We calculate the authority and hub scores using Eq. 2 and Eq. 3, respectively.

$$a_d = A_d{}^T \times h_d \qquad (2)$$

$$h_d = A_d \times a_d \times \mathcal{P} \qquad (3)$$

where $\mathcal{P}$ is the number of distinct authorities pointed by each of the hubs. The approach recursively works by assigning the hubs and authorities scores initially equal to 1. In each iteration, the hub and authority score are updated and the scores at the converging iteration are considered as the final hub and authority scores.

An Illustrative Example of Candidate Experts Identification

Suppose U and K be the two sets such that $U = \{U_1, U_2, \dots, U_n\}$ and $K = \{K_1, K_2, \dots, K_n\}$ represent Twitter based expert users and the keywords used by each expert, respectively. The initial scores for hubs and authority are assumed as ${h_a}^0 = [1,1,1,1]^T$ and ${a_a}^0 = [1,1,1,1,1,1]^T$, respectively. Table I shows a matrix comprising of four

Table I: User-keyword matrix

	K_1	K_2	K_3	K_4	K_5	K_6	K_{total}
U_1	3	2	-	5	6	-	16
U_2	2	-	-	6	7	-	15
U_3	3	3	2	4	2	-	14
U_4	3	-	-	-	-	15	18

users and six different keywords. Apparently it seems that user U_4 is the most popular user among all of the four users because it uses 18 keywords in total. Likewise, U_1 is at 2-nd position with 16 keywords, and U_3 is the lowest in terms of keywords count. The algorithm for finding the candidate experts is recursively applied such that in each iteration the hub and authority scores are updated. Table II and Table III respectively show the hub and authority score at the first and the convergence iteration. Table II shows that the hub score for U_3 after first iteration is highest whereas U_1 is at 2-nd position. However, after 41-st iteration, U_1 emerges as the hub with the highest score whereas U_3 is at 2-nd position. Table III presents the authority scores at the first and converging (37-th) iteration. It can be observed from Table III that the keyword count for each of K_4, K_5, and K_6 is equal to 15 whereas K_1 has 2-nd highest keyword count.

Table II: Hub score

Iteration No.	U_1	U_2	U_3	U_4
1	0.914	0.643	1	0.514
41	1	0.790	0.839	0.178

Table III: Authority score

Iteration No.	K_1	K_2	K_3	K_4	K_5	K_6
1	0.689	0.389	0.160	1	0.964	0.621
37	0.578	0.342	0.127	0.991	1	0.202

After 1-st iteration the keyword K_4 emerges as the keyword with the highest authority score whereas the authority score for K_6 drops down to even K_1 that had fairly less count as compared to K_6. More obvious differences in authority scores can be observed at converging (37-th) iteration where K_5, K_4, and K_1 evolve as the authorities with the highest scores. Interestingly, K_6 that has the highest keyword count turned extremely low in terms of authority score. Likewise, the hub score for U_4 that actually used the highest keywords turned lowest at converging iteration. The reason is that U_4 only used two keywords and one of them was repeatedly used. On the other hand, U_1 that fairly used distinctive keywords with different frequencies evolved as the hub with the highest score despite of low disease specific keywords as compared to U_4. Another interesting observation is about U_3 that used even more distinct keywords with varying frequencies attained the 2-nd highest hub score. The reason that U_3 also used K_2 and K_3 that were not as important as used K_4 and K_5. Consequently, it can be concluded that there is no single factor that individually helps in accurate identification of the candidate experts. Instead each of the keyword frequency, use of distinctive keywords, and the importance of the hubs pointing to the authorities contribute in accurately identifying the candidate health experts.

B. Influential User Identification

After the candidate experts have been identified through the hubs and authorities based approach, we further refine the process of expert user identification to ensure that the querying users are recommended the most relevant experts. Therefore, we introduce a metric that computes the influence of each of the candidate experts. The influence of a user is calculated based on the: **(a)** number of followers of the expert on Twitter, **(b)** total health related tweets, **(c)** sentiments of the followers in replies to the tweets by experts, and **(d)** retweets. The intuition behind using the aforementioned multiple criteria is that only single criteria, for example the number of followers is not sufficient to determine the influence or popularity of an expert on Twitter. Therefore, it is important to evaluate the influence of an expert based on several different criteria. This will also enable the querying users to evaluate the influence of an expert based on multiple prioritized criteria. The replies of the followers of a health expert are important in determining the influence and reputation of a health expert. The users in their replies to the tweets by the health expert express their sentiments. The sentiments expressed in the tweets may be positive, negative, or neutral. To classify the sentiments from the replies to the tweets as positive, negative, or neutral, we used Stanford CoreNLP library [20]. However, we only used positive sentiments scores for the replies against all of the tweets of a particular health expert. The reason for considering the health related tweets as one of the influence criteria is that a health expert may also tweet about some matters different from the health. Therefore, considering the total number of tweets on all topics by the health experts may significantly affect the total influence calculated for that expert. Likewise, the numbers of retweets by the followers of an expert are also an important factor that can portray the influence and popularity of an expert.

The users that are interested in finding the health experts based on the number of followers assign high importance to that criteria in their queries. The users are returned a ranked list of the health experts that best match their query. The criterion with the high importance or priority indicated by the user is assigned higher weights

whereas those with the low importance are assigned lower weights while ranking the experts. Weight assignment is an important task to rank the experts based on a certain criteria. We used Rank Order Centroid (ROC) method [21] to assign weights to different criteria. In ROC method, the weights to different attributes or decision criteria are assigned according to their relative importance. The weight assignment using the ROC is performed as follows:

Algorithm1: Expert User Identification

Output: List of health experts H_E

Definitions: D = set of diseases, K_d = set of Keywords against disease d, T_d = tweets for disease d, T_k = tweets collection for a keyword k, U_d =set of users who tweet about a particular disease d, C_{ukd} = frequency of a keyword k in the tweets of a user u for disease d in his/her tweets, $м_d$= user to keyword popularity matrix for disease d, N= number of required expert users, , $ʁ_t$=ratio of health related tweets to the total tweets, ṅ=retweets, Ẃ =weight assigned to each decision criteria, Ī = Influence Matrix, and W_m = weighted influence matrix for all possible combinations of weights.

1: **PARFOR** each $d \in D$ **do**
2: $k_d \leftarrow keyWordsSearch(d)$
3: **PARFOR** each $k \in K_d$ **do**
4: $T_k \leftarrow searchTweetRepository(k)$
5: $T_d \leftarrow T_d \cup T_k$
6: **end PARFOR**
7: **PARFOR** tweet $t \in T_d$ **do**
8: $u \leftarrow extractUser(t)$
9: $U_d \leftarrow U_d \cup u$
10: **end PARFOR**
11: **PARFOR** user $u \in U_d$ **do**
12: $ut \leftarrow tokenize(u)$
13: **PARFOR** keyword $k \in k_d$ **do**
14: $C_{ukd} \leftarrow getKeywordCountInProfile(ut, k)$
15: **end PARFOR**
16: **end PARFOR**
17: $м_d \leftarrow generatePopMatrix(U_d, K_d, C_{ukd})$
18: $Ć_d \leftarrow getCandidateExperts(м_d)$
19: **PARFOR** each $c \in Ć_d$ **do**
20: $f \leftarrow getFollowers(U_d)$
21: $Ṡ \leftarrow getSentimentsScore()$
22: $ʁ_t \leftarrow getHealthTweets()$
23: $ṅ \leftarrow getRetweets()$
24: $Īc \leftarrow calculateInfluence(f, Ṡ, ʁ_t, ṅ)$
25: $W_m \leftarrow calculateWeightedMatrix(Īc, W)$
26: **end PARFOR**
27: **PARFOR** each $cw \in setofPossibleCombinations\ of\ weights$ **do**
28: $E_{cw} \leftarrow getTopRankedExperts(W_{m_{cw}})$
29: **end PARFOR**
30: ***Update*** H_E
31: **end PARFOR**

$$Ī = \sum_{n=1}^{k} (Cr_n \times W_n) \quad (4)$$

where Cr_n refers to the particular criteria and W_n is the weight assigned to that criteria.

Algorithm 1 presents the steps to identify and rank the influential health expert users from the Twitter using the variant of HITS approach and the proposed influence metric. Line 2 of Algorithm 1 executes in $O(k)$, where k is the number of keywords. Line 3—Line 6 searches the repositories and have complexity $O(T \times k)$, where T represents the tweets. The operations in Line 7—Line 10 extract the users and have complexity $O(U \times T)$. Line 11—Line 16 execute in $O(\cup \times x \times k)$, where x be the number of tokens. Line 17 and Line 18 execute in $O(\cup \times k)$ and $O(m \times (\cup^2 + k^2))$, where m represents the number of iterations required by the variant of HITS to converge . Line 20 executes in $O(1)$ and each of Line 21—Line 25 take $O(T)$ to execute. Therefore, the total complexity from Line 19—Line 26 becomes $O(c \times 5T)$, where c being the number of candidate experts. Line 27—Line 29 execute in $O(24 \times T\ log(T))$, where $T\ log(T)$ is the time complexity to sort the list of top ranked experts. The total complexity of the algorithm to find the experts for a disease d becomes $O(d \times (k(1+T)) + (U(T+x)) + (K(1+U)) + (m \times (\cup^2 + k^2)) + (c + TlogT))$.

IV. Results and Discussion

We evaluated the effectiveness of our approach in terms of recommendation accuracy and scalability against varying workloads. Evaluation results for the expert user recommendation and scalability are presented in subsequent subsections.

A. Evaluation of Expert User Recommendation Module

The performance of the expert user recommendation module in terms of accuracy was evaluated and precision, recall, and F-measure [22] were used as the evaluation metrics.

Precision: The ratio of the accurately identified health experts (True Positives) to the total occurrences (True Positive (TP) + False Positive (FP)) is termed as precision and is given as:

$$Precison = \frac{TP}{TP + FP} \quad (5)$$

Recall: The identification probability of the randomly selected health expert from the total training set (True Positive (TP) + False Negative (FN)) is called recall and is given as:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

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

$$F - measure = \frac{2TP}{2TP + TP + FN} \quad (7)$$

The tweets were collected by using the twitteR package of *R* [23]. We evaluated the performance for correctly identifying the health experts by collecting over 20,000

profiles of Twitter users who used the health related terminologies in their tweets. Around 400,000 tweets related to the diabetes were collected from the Twitter by using the hypernyms, hyponyms, meronyms, holonym, and derivationally related terms through the WordNet. The aforementioned numbers also contain the tweets that were provided by the Symplur on request. The tweets repositories are maintained and updated by periodically executing the jobs in offline mode. The framework also performs the computations of the hub and authority scores to identify the candidate experts and the influential users in offline mode. The reason to perform the aforementioned tasks offline is that it requires huge amount of storage and processing that eventually results in high query response time. Therefore, our cloud based framework effectively stores the large amounts of Twitter data and performs intensive computation operations for the identification of health experts. Moreover, to minimize the query response time, the tweet repositories are preprocessed based on the geographical locations.

The performance of our approach was evaluated in terms of accuracy by comparing with the approaches presented in [15] and [24]. In addition, we also compared our approach with the popularity based ranking approach that only considers the frequency of keywords used by the health experts. The precision, recall, and F-measure for each of the approaches are presented in Fig. 2, Fig. 3, and Fig. 4, respectively where our proposed approach is termed as Influential User Recommendation (IUR). The performance of the IUR approach was observed to be sufficiently better as compared to the other approaches in terms of precision, recall, and F-measure for *Top-k* experts with *k*=(5, 10, 15, 20). However, the approach by Cheng *et al.* [23] also turned with high accuracy as compared to the approach presented in [15] and the popularity based approach. The popularity based approach attained low accuracy particularly for *Top-k* experts with *k*= (10, 15, 20). Interestingly the proposed IUR approach exhibited relatively high accuracy even at large *k*, *such as k*= (15, 20). The comparison of results shows that our proposed approach that first identifies the candidate experts and then calculates the influence of the candidates offers more accurate recommendations. In addition, offering users the facility to search and evaluate the experts by specifying four different criteria helps to obtain personalized recommendation about help experts.

B. Scalability Analysis

The systems based on the centralized computing models come across the issues of scalability because of their inability to cope with the ever changing processing requirements. Consequently, the deployment of decentralized cloud based methodologies that enable the concurrent processing of large data volumes is becoming inevitable. For a parallel algorithm to be scalable, with the increase in number of resources, for example the processors and the workload, the performance in terms of time efficiency and resources' utilization must be consistent or should not degrade substantially [25]. Therefore, we utilized the cloud services because they can be procured on-demand and according to requirements. Amazon Elastic Compute Cloud (EC2) [26] is an example of commercial cloud service provider that provides the processors, storage, and memory to host applications based on different pricing models. We evaluated the scalability of our approach by analyzing the effects of increase in workload and processors on the time consumption for: **(a)** the candidate expert identification module, **(b)** calculation of the influential users by considering all the possible permutations for a single query, and **(c)** weight assignment to four prioritized criteria.

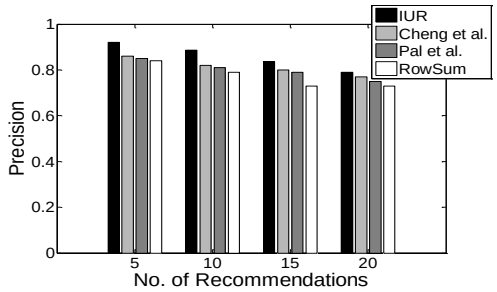

Fig. 2: Precision comparison of IUR with other approaches

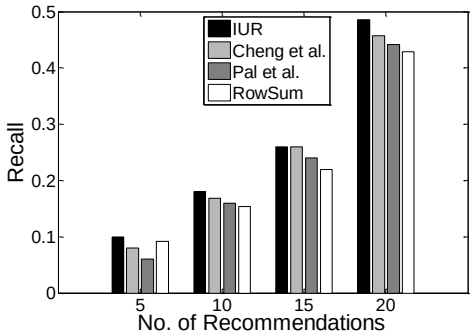

Fig. 3: Recall comparison of IUR with other approaches

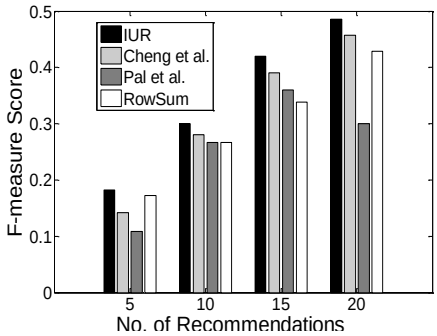

Fig. 4: F-measure comparison of IUR with other approaches

Each of the aforementioned tasks is performed offline and the repositories are updated periodically to avoid the overheads arising due to online processing. The influence is calculated based on the importance of the criteria indicated by the users. Because there are four criteria over which users can view the ranking decisions, it makes a total of 24 different possible combinations to evaluate the final ranking or influence of an expert for a single query. Obviously, it is

impractical to calculate ranking score for each of the combinations at run time to manage the queries of users from different geographical regions. Therefore, executing the parallel and periodic jobs not only avoids high processing delays but also ensures the availability of updated information at all the times. We evaluated the performance of all of the modules in terms of time consumption by increasing the number of users and the number of processor. The time consumption to identify the expert users from 20,542, 41,084, and 82,168 user profiles by varying the number of processors was observed.

Fig. 5 shows the scalability results with different workloads and number of processors to identify the candidate health experts using the variant of HITS approach. The results show that increasing number of users two times resulted in sudden increase in the processing time. However, substantial decreases in time consumption were observed by increasing the number of processor. On average, by increasing the number of user profiles twice increases the time consumption by approximately 38.72% whereas increasing one processor resulted in an average decrease of 16.27% for the candidate experts identification task. It is also important to note that by increasing the number of processors more than a certain limit, relatively small decreases in processing time were observed. The reason is that this time also includes the overheads, such as the processor start up time and the communication time between the two processors. For large number of processors, the aforementioned overheads also increase and consequently affect the total execution time. Fig. 6 shows the execution time corresponding to the three workloads for influential user identification module. The influential users' identification module calculates the number of followers of each of the experts, performs sentiment analysis, and calculates the health related tweets and the retweets. Consequently, for each candidate expert, four different tasks are to be performed, which requires parallel task processing to speed up the query response time. By increasing the number of profiles twice, the average combined increase in time consumption is 72.03% whereas an average decrease of approximately 66.37% is observed by increasing one processor at a time. Fig. 7 shows the processing time for weight assignment to various decision criteria. For each user query, the framework performs weight assignment according to 24 different combinations. This requires sufficient computations that result in increased processing time, if performed online. It also appears from Fig. 7 that the time consumption for weight assignment task is sufficiently less than the two other modules. The reason is that weight assignment is only subtask of the process of influential user identification that has to be performed repeatedly.

It is evident from the above discussion and results that all the tasks starting from the tweets extraction to the influential user identification require enormous processing time and resources. Therefore, query response time can only be reduced if all the tasks demanding heaving computations are preprocessed and periodically updated to ensure the delivery of the most recent information about health experts. The experimental results also reveal that with the increase in workload and processors, our algorithm substantially maintains the efficiency in terms of time consumption. Therefore, the proposed cloud based approach is highly effective and can scale up and scale down depending upon the workloads.

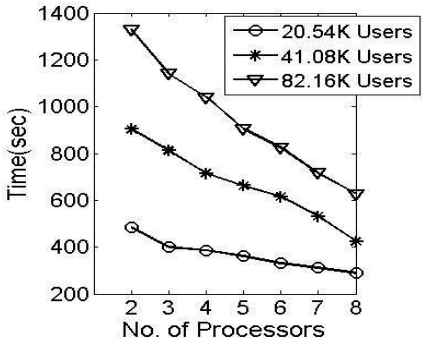

Fig. 5: Execution time analysis for different no. of users and processors to identify candidate experts

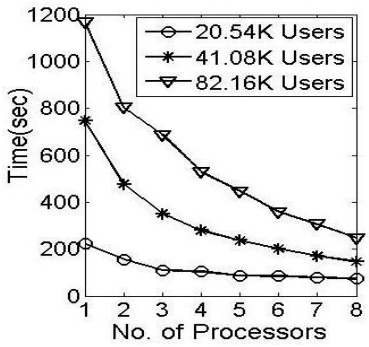

Fig. 6: Execution time analysis for different no. of users and processors to identify influential users

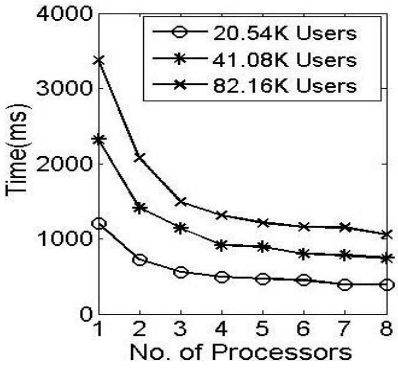

Fig. 7: Execution time analysis for different no. of users and processors for weight assignment

V. CONCLUSIONS

In this paper, we proposed a framework that enables the users to interact with the health experts from Twitter to seek the advice at no cost. The framework utilizes the cloud infrastructure to manage huge tweet repositories. The variant of the HITS algorithm is employed to identify the candidate experts. The approach effectively identified the candidate experts by considering the use of distinctive keywords, importance of the keywords, and the importance of the experts using the keywords. To make the ranking process more effective, we further introduced an influence metric that identifies the influential users from the list of candidate experts. Experimental results demonstrate that the proposed framework is highly effective in terms of accuracy as compared to other approaches. Moreover, the performance of the system in terms of execution time is preserved at high workload which indicates the scalability of the system. We are optimistic that the research will be helpful to fully utilize the potential of the online health communities by offering free of cost interaction services among the patients and doctors.

REFERENCES

[1] 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*, vol. 43-44, pp. 99-109, 2015.

[2] K. Mille, "Big Data Analytics in Biomedical Research," *Biomedical Computation Review*, pp. 14-2, 20121.

[3] H. Chen, R.H.L. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS quarterly* 36, no. 4, pp. 1165-1188, 2012.

[4] "Patientslikeme",http://www.patientslikeme.com/, accessed on March 7, 2015.

[5] S. Fox, M. Duggan, "Health online 2013," http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf, accessed on March 25, 2015.

[6] "Health Fact Sheet", http://www.pewinternet.org/fact-sheets/health-fact-sheet/, accessed on March 7, 2015.

[7] A. Abbas, L. Zhang, and S. U. Khan, "A Survey on Context-aware Recommender Systems Based on Computational Intelligence Techniques," *Computing*, DOI 10.1007/s00607-015-0448-7.

[8] 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, pp. 1431-1441, 2014.

[9] M. Ali, S. U. Khan, and A. V. Vasilakos, "Security in Cloud Computing: Opportunities and Challenges," *Information Sciences*, vol. 305, pp. 357-388, 2015.

[10] B. Kayyali, D. Knott, and S. V. Kuiken, "The big-data revolution in US health care: Accelerating value and innovation," Mc Kinsey & Company, pp. 1-13, 2013.

[11] "Healthcare Social Media Analytics," http://www.symplur.com/healthcare-social-media-analytics/, accessed on March 10, 2015.

[12] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University, Press, 2010.

[13] J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261-270, 2010.

[14] 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*, pp. 1-7, 2014.

[15] 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.

[16] 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*, pp. 575-590, 2012.

[17] X. Tang, and C. C. Yang, "Identifing influential users in an online healthcare social network," In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 43-48, 2010.

[18] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM* 38, no. 11, 1995, pp. 39-41.

[19] H. Park, J. Yoon, and K. Kim, "Identifying patent infringement using SAO based semantic technological similarities," *Scientometrics* 90, no. 2, pp. 515-529, 2012.

[20] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55-60.

[21] T. Solymosi, and J. Dombi, "A method for determining the weights of criteria: the centralized weights," *European Journal of Operational Research* 26, no. 1, pp. 35-41, 1986.

[22] P. Bedi and R. Sharma, "Trust based recommender system using ant colony for trust computation," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1183-1190, 2012.

[23] "twitteR: R based Twitter client," http://cran.r-project.org/web/packages/twitteR/index.html, accessed on March 21, 2015.

[24] 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 March 21, 2015.

[25] 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.

[26] "Amazon EC2 Pricing," http://aws.amazon.com/ec2/pricing/, accessed on March 19, 2015.