# Chapter 7
# Clustering Questions in Healthcare Social Question Answering Based on Design Science Theory


**Blooma John and Nilmini Wickramasinghe**

## 7.1 Introduction

In healthcare social media, users connect with patients and professionals without time and space boundaries to seek and share healthcare-related information (Denecke and Stewart 2011). A classic example of a Medicine 2.0 application is a healthcare Social Question Answering (SQA) service. Healthcare SQA services are redefining healthcare delivery and supporting patient empowerment. Healthcare SQA services allow users to seek information, communicate with others on similar problems, share health guidance, and compare treatment and medication strategies (Blooma and Wickramasinghe 2014). Examples of healthcare SQA services are *MedHelp*, *BabyHub*, and *Drugs.com*. The growing activities in online healthcare communities, asking questions and sharing answers, play an important role in users' health information inquiries (Zhang and Zhao 2013). Individual behaviors, in particular health-related behaviors such as physical activity, diet, sleep, smoking, and alcohol consumption, as well as adherence to medical treatments and help-seeking behavior (Hyyppä 2010), appear to be significant in SQA services.

On the other hand, there is a need to aid in assisting and mining the content shared to make the process of retrieving quality content, relevant to users, easier. For millions of users who ask questions in a healthcare SQA service like *Drugs.com*, the answers for the past questions submitted comprise a valuable knowledge repository. As the quality and the source of the questions and answers vary widely,

B. John (✉)
University of Canberra, Bruce, ACT, Australia
e-mail: blooma.john@canberra.edu.au

N. Wickramasinghe
Deakin University, Melbourne, VIC, Australia

Epworth HealthCare, Richmond, VIC, Australia



N. Wickramasinghe, J. L. Schaffer (eds.), *Theories to Inform Superior Health Informatics Research and Practice*, Healthcare Delivery in the Information Age,
https://doi.org/10.1007/978-3-319-72287-0_7

there is a need to further study the relationship between the users and the content they post, particularly with respect to health-related questions and answers. Earlier studies (Agichtein et al. 2008; Bian et al. 2009) classified high quality content from various SQA services. However, there were very few studies that focused on SQA services in the healthcare social media. Hence, the research question addressed in this paper is:

- How can we cluster similar questions shared on SQA services in healthcare social media based on the quality of the content shared by users?

We used quadri-link cluster analysis based on various features related to questions, answers, and users to cluster similar content and a similarity measure to classify content. Based on an earlier work (Blooma et al. 2016), the similarity measure was developed based on the user profile and the relationship network between the users and the types of information disclosed. We extended the features used by earlier studies (Bian et al. 2009; Agichtein et al. 2008) and used the similarity measure to cluster similar content based on a design science approach (Hevner et al. 2004).

This paper proceeds as follows. In the literature review, we review studies related to health informatics, design science, and cluster analysis. We then present the methodology and quadri-link cluster analysis to cluster similar questions. We describe our data collection and analysis procedures. We present the results of our pilot study and explain the precision based on a preliminary content analysis. We finally conclude with the contributions of this paper and propose future work.

## 7.2 Literature Review

Health informatics is a relentless pursuit of helping people to improve health by using information technology (Friedman 2012). Health informatics is an integration of elements from broadly defined information science and health science. On the other hand, studies in health informatics tend to miss either the health problem or the information technology problem. For example, De Vries et al. (2013) reviewed 55 heart failure risk computational models and showed clear evidence that only a few had been implemented in clinical practice. In general, previous studies highlighted that innovations fail to achieve sustainability because the health technology disregards the relationship between the technology and the people involved.

Today, social media has empowered users to post content that is publicly available, and the dangers and threats of inaccurate diagnoses are formidable (George et al. 2013). Hence, as we propose a novel method to cluster similar content shared by users of SQA services in healthcare social media, we based our similarity measures on the quality and the relationship of the content and users. For the design of this proposed cluster analysis, we used design science guidelines as reviewed in the following section.

Design science is an important and legitimate research paradigm in information systems (Gregor and Hevner 2013). Design science research involves constructing

a wide range of sociotechnical artifacts, such as new software, processes, algorithms, or systems intended to improve or solve an identified problem (Myers and Venable 2014). Hevner et al. (2004) presented seven guidelines for understanding, executing, and evaluating design science research. The seven guidelines are design as an artifact, problem relevance, design evaluation, research contributions, research rigor, design as a search process, and communication of research. Design science guidelines originated from information systems design theory originally proposed by Walls et al. (1992a; b) as "a prescriptive theory which integrates normative and descriptive theories into design paths intended to produce more effective information systems." Eventually, Peffers et al. (2007) expanded design theory into a design science research methodology by incorporating the principles, practices, and procedures required to carry out research by applying design science theory. They suggested that design science theory as a methodology needs to be consistent with prior literature, provide a nominal process model for doing design science research, and provide a mental model for presenting and evaluating design science research (Peffers et al. 2007). While we used cluster analysis, the nature of design science theory provides a foundation for more systematically specifying its design. Based on Arnott and Pervan (2012) and Xu et al. (2007), we used design science theory guidelines to propose a model to cluster similar questions in SQA services.

Cluster analysis groups together similar objects into meaningful clusters based on the similarities among the objects (Balijepally et al. 2011). Information systems research uses cluster analysis as an analytical tool for classifying configurations of various entities that comprise the information technology artifact. Because of the nature of the questions posed in SQA services, keywords alone do not provide a reliable basis for clustering user-generated questions effectively, particularly in SQA services for health care (Bian et al. 2009; Agichtein et al. 2008). To overcome the disadvantages of keyword-based clustering, extant research focuses on additional criteria. Blooma et al. (2016) used the content and user relationship to identify similar questions. Leung et al. (2008) introduced the notion of concept-based graphs. Bian et al. (2009) used a mutually coupled bipartite network to identify high quality content and users. However, there is a lack of studies that applied cluster analysis in health care to identify similar questions and reuse the existing answers for new questions.

Thus, in this paper, we propose a novel cluster analysis based on a design science approach by considering the relationship between the questions, answers, askers, and answers to cluster similar questions as detailed in the next section.

## 7.3 Methodology

In this section we identify a set of features related to questions, answers, and users to classify similar questions in healthcare social media. The features are based on Bian et al. (2009), Agichtein et al. (2008), Chan et al. (2010), and Angeletou et al. (2011).

**Table 7.1** Features for questions

| Feature | Description | References |
|---|---|---|
| Question | Words in the question subject and question detail | Bian et al. (2009), Agichtein et al. (2008) |
| Subject length | Number of words in the question subject | |
| Detail length | Number of words in the question detail | |
| Posting time | Date and time when the question was posted | |
| Question votes | Number of positive votes | |
| Number of answers | Number of answers received | |
| Punctuation density | Number of punctuation marks divided by the total number of characters | |
| Question's category | Tags/topics assigned to questions by the asker | |
| Number of words per sentence | Average number of words per sentence in the current question | |
| Capitalization errors | Number of sentences not starting with capitalized letters | |
| The Flesch–Kincaid (F–K) reading grade level | The FK reading score indicates the level of difficulty in reading | |

The four distinct sets of entities are questions, answers, users, and concepts. In this model we combine both answerer and asker into one entity—user. This is mainly because the actual role of the user is not determined a priori in this extended model when compared to Blooma et al. (2016); instead, the computed values of features regarding the social role will resolve their final role as an asker or an answerer. The second major inclusion in this model is that we considered concepts as a new entity that contains rich information. However, we simplified the concept extraction as the extraction of unique nonstop words as meaningful words. Although we used unique nonstop words in this study, we can further extend the concepts using a medical thesaurus to be very specific in healthcare medical terms.

In particular, the features used to represent questions are given in Table 7.1 and are mainly used to focus on intrinsic content quality metrics. The features are text related as the questions and answers we analyzed are primarily textual in nature. In addition to the content itself, there is a wide array of noncontent information available, from links between items to explicit and implicit features of the content, such as posting time, questions, votes, punctuation, typos, and semantic complexity measures.

With respect to an answer, in addition to the intrinsic content quality metrics as well as noncontent information, we also used the relationship features of questions and users. Word overlap and the ratio between the lengths of the question and the answer are features that are based on the relationship between the question and its answer. Features such as positive votes and negative votes are based on the

**Table 7.2** Features for answers

| Feature | Description | References |
|---|---|---|
| Overlap | Words shared between the question and answer | Bian et al. (2009), Agichtein et al. (2008) |
| Number of comments | Number of comments added by other participants | |
| Total positive votes | Total number of positive votes for the answer | |
| Total negative votes | Total number of negative votes for the answer | |
| Answer length | Number of words in the answer | |
| Unique words | Number of unique words in the answer | |
| QA ratio | Ratio between the question length and the answer length | |
| Number of words per sentence | Average number of words per sentence in the current question | |
| Capitalization errors | Number of sentences not starting with capitalized letters | |
| The Flesch–Kincaid (F–K) reading grade level | The FK reading score indicates the level of difficulty in reading | |

relationship between the answer and users. The list of features used for answers is listed in Table 7.2.

The user features were mainly adopted from Angeletou et al. (2011) and Chan et al. (2010). The features in-degree, out-degree, hub score, authority score, and initialization are used to reflect the structural network properties of a user within the community and the user's popularity in the community. These features, along with their relationship with the questioning and answering in which they participate, helped us enhance the quality of the cluster analysis, based on the relationship between the content and the users. It is highly important to put emphasis on the authority of the users' in a healthcare-related community to improve the quality of the outcome. The features used with respect to the users are detailed in Table 7.3.

Thus, to plot the relationship between the content and users, we separated the content and users into four distinct sets of entities and six distinct types of links as given in Fig. 7.1.

After we framed the quadri-link model, we calculated similar measures for questions, answers, and users. We introduced the similarity measure to compute the similarity score between two sets of questions, answers, or users by considering two components: link count to the same concepts and value of other features. We used the Jaccard similarity index (for a discrete set) and its extended general form for nonnegative real values as suggested by Charikar (2002). The equations for all three entities (answers, questions, and users) are similar, and they consist of four parts: the Jaccard index of the bag of words, the general Jaccard index of features, and the Jaccard indices of graph links to the two other clusters. The equations to calculate similarity between answers (sim($ai$, $aj$), questions sim($qi$, $qj$), and users sim($ui$, $uj$) are given below.

**Table 7.3** Features for users

| Feature | Description | References |
|---|---|---|
| Questions asked | Number of questions asked across all Q and A groups | Chan et al. (2010), Angeletou et al. (2011) |
| Total answers | Number of posted answers across all Q and A groups | |
| Votes | Total votes the user received | |
| In-degree | Number of other users answered by this user | |
| Out-degree | Number of other users that answered this user | |
| Hub score | Hub score for the user computed by the HITS algorithm | |
| Authority score | Authority score for the user computed by the HITS algorithm | |
| Average votes per answer | Average votes per answer divided by the total number of answers | |
| Initialization | Number of questions asked by this user divided by all questions asked | |
| % in-degree | The in-degree of this user divided by unique in-degrees | |

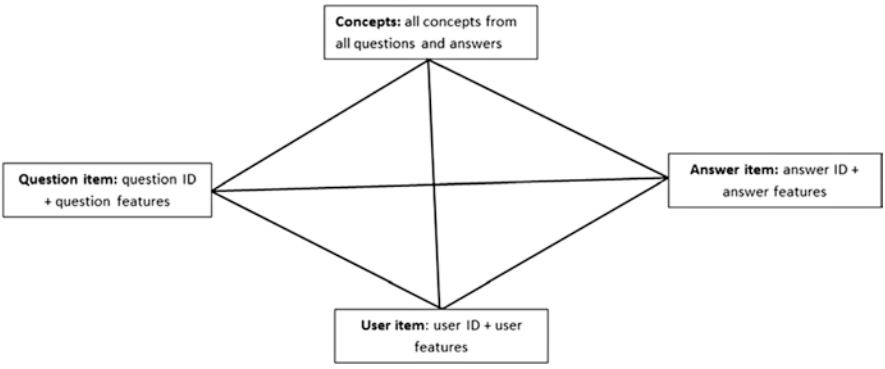



**Fig. 7.1** Quadri-link model

$$\begin{aligned}\mathrm{Sim}\left(a_i,a_j\right) &= \frac{\text{Common concepts}}{\text{Total distinct concepts of both } a_i \text{ and } a_j} \\ &+ \text{If their questions are the same or in the same cluster 0 else 1} \\ &\text{If their users are the same or in same cluster 0 else 1} \\ &+ \text{Jaccard similarity of features in the framework}\end{aligned} \tag{7.1}$$

$$\begin{aligned}\mathrm{Sim}\left(q_i,q_j\right) &= \frac{\text{Common concepts}}{\text{Total distinct concepts of both } q_i \text{ and } q_j} \\ &+ \text{If their answers are the same or in the same cluster 0 else 1} \\ &\text{If their users are the same or in same cluster 0 else 1} \\ &+ \text{Jaccard similarity of features in the framework}\end{aligned} \tag{7.2}$$

$$\begin{aligned}\mathrm{Sim}\left(u_i,u_j\right) &= \frac{\text{Common concepts}}{\text{Total distinct concepts of both } u_i \text{ and } u_j} \\ &+ \text{If their questions are the same or in the same cluster 0 else 1} \\ &+ \text{If their answers are the same or in same cluster 0 else 1} \\ &+ \text{Jaccard similarity of features in the framework}\end{aligned} \tag{7.3}$$

We used the above similarity measures to cluster similar answers, questions, and users iteratively. We used complete linkage similarity to find the most similar clusters (Defays 1977). We clustered all three entities (question, answer, and user) as we proceeded with the iteration. However, we needed to choose the order of clustering. The user needs information from the question and answer clusters rather than the individual questions or answers. Their user score will be computed last in the iteration. Similarly, questions need information from answers (two questions do not share the same individual answer); thus their score will be computed second. Consequently, the score of answers will be computed first since they only need the information of the individual entities from the other three sets. Therefore, the suggested order of clustering is answer, question, and user, and the proposed clustering algorithm is called quadri-link cluster analysis.

We also propose two ways to terminate the algorithm:

1. Run the algorithm until completion (everything is in one cluster), plot a dendrogram, and then select the number of clusters based on the plot.
2. Terminate upon reaching a certain condition (there is no similarity score above a certain threshold).

As we proceed, we first compute the feature value as shown in Tables 7.1, 7.2, and 7.3. Once we compute all the features, we feed the computed similarity scores to perform quadri- link cluster analysis. The steps used to calculate the quadri-link cluster analysis are listed below.

*The steps of quadri-link cluster analysis*:

1. Obtain the maximum similarity score of all answer clusters according to the similarity function above.
2. Merge the answer clusters with the highest similarity score.
3. Obtain the maximum similarity score of all question clusters according to the similarity function above.
4. Merge the question clusters with the highest similarity score.

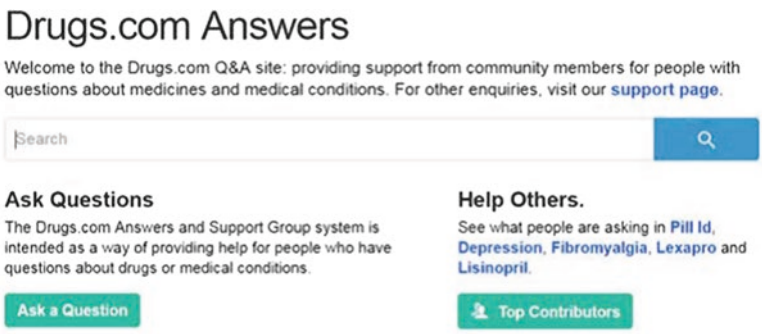



**Fig. 7.2** Drugs.com question and answer site

5. Obtain the maximum similarity score of all user clusters according to the similarity function above.
6. Merge the user clusters with the highest similarity score.
7. Repeat step 1–7, unless the similarity score falls below the threshold.

While merging, if several clusters shared the same maximum score, they were merged in the same iteration. For example, if clusters A–B had 3.5, B–C had 3.5, and D–E had 3.5, then after merging we will have two new clusters A–B–C and D–E. For the general Jaccard index of features, we used a complete linkage criterion; thus during merging we used the minimum similarity score as the representative score of the new merged cluster. The similarity score between the new merged cluster A–B and cluster C was the minimum of {A–C, B–C}. For the graph link, the merge was an OR function of all sets of links. For example, when merging user clusters A–B, user A had links to question clusters C and D and answer cluster E, and user B had links to question cluster C and answer cluster F. The new user cluster A–B will have links to question clusters C and D and answer clusters E and F.

We conducted a pilot study by collecting publicly available data from the Drugs.com question and answer site[1] and tested the quadri-link cluster analysis to identify similar questions. A sample screenshot from Drugs.com is shown in Fig. 7.2. We collected 200 resolved questions related to obesity so that we could focus on the obesity domain in health care for testing and analysis. Along with the questions, we collected the answers for the resolved questions and the user details involved. We also collected the number of votes obtained for the answers.

Thus, to summarize, the main features that we collected for questions are question ID, user ID, title, description, topic, date, and total number of answers. For maintaining anonymity and privacy, identifications for questions, answers, and users were created as we collected the respective data. Similarly, the main features that we collected for answers are listed as answer ID, user ID, question ID, date, votes, text, and total comments. The main features that we collected for users are user ID, total questions asked by the user, total answers answered by the user, and

[1] https://www.drugs.com/answers/.

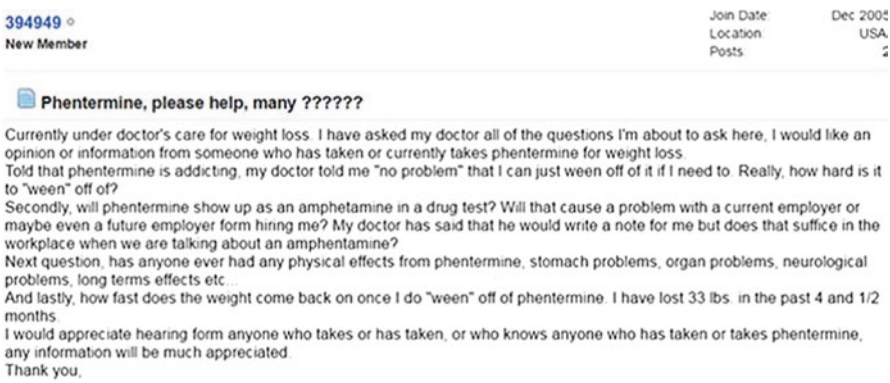
points earned by the user. Based on the collected data, the features listed in Tables 7.1, 7.2, and 7.3, respectively, for questions, answers, and users were calculated before we tested the proposed quadri-link cluster analysis. The results and findings from a pilot testing of the proposed quadri-link cluster analysis are discussed in the following section.

## 7.4 Results

We analyzed the content of the clustered questions by looking back at the answers and the users involved in the question to judge the similarity of the questions rather than the wording of the question itself. We found that the majority of the questions users asked contained a detailed description of a personal problem that required an answer. We also found that 79% of the questions had detailed descriptions with an average word count of 56.

The question with the longest description had 788 words, and it is a detailed explanation about diet, exercise, and metformin with a little personal history and story such as this "….So let me tell you my story and maybe you can share any ideas that you might have regarding what it is I suffer from,…." The question also had detailed answers, with the longest answer having 505 words. The answers, in turn, had comments that held the conversation lively. Yet another interesting fact was that the question was asked in 2012 and it continued getting replies, with the last answer posted in 2015. The first answer to the question was answered on the same day it was asked. Examples of the sample question and answer are given in Figs. 7.3 and 7.4. User details are the points they earned in participation and their questions and answers as illustrated in Fig. 7.5.

We then evaluated the results of the proposed quadri-link cluster analysis using three different combinations of datasets. We compared the results by using quadri-partite cluster analysis (Blooma et al. 2016).


394949
New Member
Join Date: Dec 2005
Location: USA
Posts:

**Phentermine, please help, many ??????**

Currently under doctor's care for weight loss. I have asked my doctor all of the questions I'm about to ask here, I would like an opinion or information from someone who has taken or currently takes phentermine for weight loss.
Told that phentermine is addicting, my doctor told me "no problem" that I can just ween off of it if I need to. Really, how hard is it to "ween" off of?
Secondly, will phentermine show up as an amphetamine in a drug test? Will that cause a problem with a current employer or maybe even a future employer form hiring me? My doctor has said that he would write a note for me but does that suffice in the workplace when we are talking about an amphentamine?
Next question, has anyone ever had any physical effects from phentermine, stomach problems, organ problems, neurological problems, long terms effects etc...
And lastly, how fast does the weight come back on once I do "ween" off of phentermine. I have lost 33 lbs. in the past 4 and 1/2 months.
I would appreciate hearing form anyone who takes or has taken, or who knows anyone who has taken or takes phentermine, any information will be much appreciated.
Thank you,

**Fig. 7.3** A sample question

Wow, You have Alot going on that is for sure. First off are you diabetic & if so I assume that is why they are wanting to put you on Metformin.
I did loose weight by taking Metformin but the main reason for that was because it makes you extremely Nauseated... If you are diabetic the best medication I found to take weight off is Victoza it is an injection you do.
But it is a new medication for Diabetes & they are trying to get the FDA to let them put this in a pill form for the purpose for people to use this to lose weight. I am not sure if they have approved that form yet.

**Fig. 7.4** A sample answer

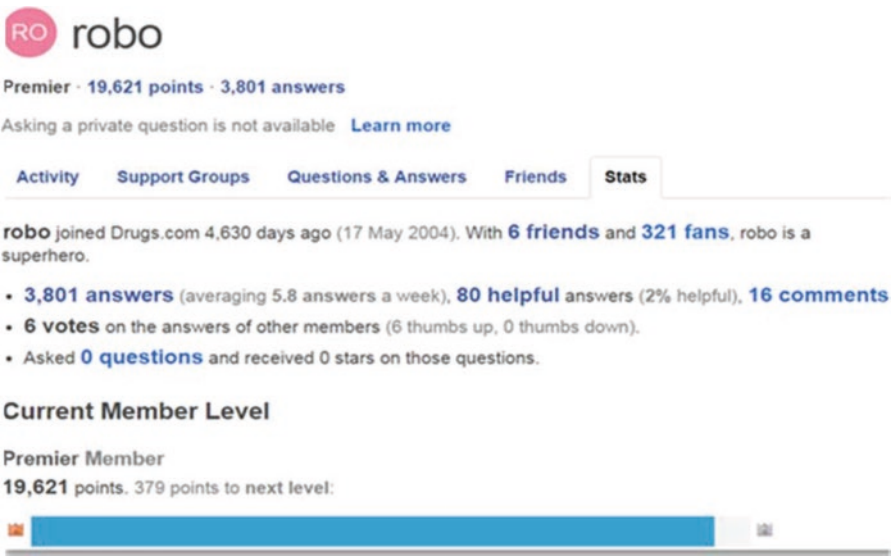



**Fig. 7.5** A sample user profile

In quadripartite cluster analysis, the most similar questions that evolved were "Do Chlordiazepoxide/clidinium pills cause weight gain?" and "Can periactin increase your appetite?". In this case, the user and the response were the same and that lead to the clustering, which is a result of the algorithm used. However, as suggested by the study, there is an immense need to put more emphasis on the medical terms. Yet another set of similar questions identified were "Drug induced weight gain; any solutions?" and "Can you take this with antidepressants?". In this case, the questions are very vague. The question leads to a sequence of discussions to describe the medicine that led to the question. The same user answered the question and the answer was to ask more details regarding the medicine. This process led to the questions being clustered as similar. Hence, based on the analysis, we extended the quadripartite cluster analysis to gauge the features as in quadri-link cluster analysis.

In the quadri-link cluster analysis, the first pair of questions that were clustered are "Does trazodone cause constipation or weight gain?" and "Does buspar cause weight

gain or constipation?". In this case, they were found to be similar. On analysis, it was interesting to note that both of the questions were asked by the same user on the same day. The questions had a different cohort of users answering the question; however, the answers were not the same but did agree that the drugs caused weight gain or constipation. This is clear evidence of the precision of quadri-link cluster analysis. Another example of similar questions is "Kindly inform me about the recommended medicine that causes weight to lose?" and "Just to get rid of obesity, which drug is beneficial?". On analysis, it is not only the questions that were similar but also the same user asked both questions. The answers for the questions were differently worded but had the same meaning. Hence, this is yet more clear evidence of the precision of quadri-link cluster analysis. Another example of similar questions asked and answered by different users that were clustered is "Feedback on prozac? Can anybody give me feedback on Prozac, especially, energy level, weight loss …" and "Prozac—like it? Can someone give me some information about Prozac?"

On the other hand, quadri-link cluster analysis did not distinguish between different medical terms when the questions were phrased the same except for the medical terms. For example, "Does cyclobenzaprine cause weight gain?" and "Does polyethylene glycol 3350 cause weight gain?" were clustered, and "Does cymbalta (60 mg) cause weight gain?" and "Does this effexor cause weight gain?" were found similar. "Does Lupron injections cause weight gain in women?" and "Does Trileptal or the Generic cause weight gain?" are yet another combination of questions that were found similar. Although quadri-link cluster analysis was found to be more precise than quadripartite cluster analysis, there is a need to emphasize the use of a medical thesaurus to identify similar medical terms to avoid clustering questions that have all of their features similar except for the medical terms (Zhang and Zhao 2013; Blooma and Wickramasinghe 2016).

## 7.5 Discussion

The application of design science theory as a cluster analysis technique is a novel step toward identifying similar questions. We present the findings of this study based on the five activities as detailed by Peffers et al. (2007).

Activity 1—Problem identification and motivation: The specific research problem is identifying similar questions in healthcare SQA services. Health information needs and the omnipresence of social media have made users seek and answer queries in various arenas like SQA services. Hence, the motivation for this study was significant as the quality and the source of the questions and answers varied widely. Moreover, little is known about how to identify similar questions to reuse the content collected in healthcare SQA services.

Activity 2—Define the objective of a solution: We focused on developing quadri-link cluster analysis based on the relationship between content and the users involved to solve the complexity of searching through user-generated content in

healthcare social media. Most importantly, we produced a viable artifact in the form of quadri-link cluster analysis. Thus, applying design science theory, we aided in assisting and mining the content shared to make the process of retrieving quality content relevant to users easier.

Activity 3—Design and Development: The designed artifact in this study is an instantiation of quadri-link cluster analysis. As we designed the artifact, we defined the list of features used with respect to questions, answers, and users. We then framed the quadri-link model comprised of questions, answers, users, and concepts to calculate the similarity measure. We finally developed the similarity measure and the algorithm to cluster similar questions.

Activity 4—Evaluation: We evaluated the artifact based on a pilot study as detailed in the results section. The evaluation resulted in clear evidence of improved precision of quadri-link cluster analysis. Although the findings from this study are a pilot attempt, there is a need to extend the analysis to various variations of similarity measures and datasets collected from other SQA services, like MedHelp, to evaluate the precision of quadri-link cluster analysis. This study can be extended to apply and evaluate other types of content in social media to cluster similar content, like tweets in Twitter and postings and comments in Facebook.

Activity 5—Communication: We published the algorithm and its applicability on healthcare social media. We also emphasize the fact that the nature and use of design science theory and quadri-link cluster analysis will be compared and contrasted in future studies.

## 7.6 Conclusion

Healthcare SQA services enable patient empowerment and the better transfer of pertinent information and germane knowledge with the potential end result being superior healthcare delivery. Furthermore, regarding the nature of chronic diseases where prevention is a key factor, a healthcare SQA service plays a major role in supporting healthier lifestyle practices. The pervasive nature of healthcare SQA services means that this is a benefit that most, if not all, people can enjoy. In many ways a healthcare SQA service has the potential to revolutionize current healthcare delivery practices and/or roles. In addition, it has a key role to play regarding public health and enabling education and change of lifestyle for all. Hence, this research sheds light on how various factors that influence specific types of health outcomes contribute to both theory and practice.

In particular, the proposed quadri-link cluster analysis contributes to the healthcare informatics domain by introducing content-, user-, and concept-based clustering applicable to sort the issues faced in content-based similarity. Based on a design science approach, the quadri-link model, similarity measures, and the cluster analysis algorithm showed that similarity was improved with respect to not only the words in the questions but also the context, which in turn improved the precision. Although the results in this paper were tested using 200 questions and the top results

analyzed, there is a need to test quadri-link cluster analysis for various combinations of data as part of future work. As highlighted in the earlier section, there is also a need to integrate a medical thesaurus so that we can further extend the concepts to be very specific for medical terms.

## References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high- quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 183–194).

Angeletou, S., Rowe, M., & Alani, H. (2011). Modelling and analysis of user behaviour in online communities. In *The Semantic Web-ISWC 2011* (pp. 35–50). Berlin: Springer.

Arnott, D., & Pervan, G. (2012). Design science in decision support systems research: An assessment using the Hevner, March, Park, and Ram Guidelines. *Journal of the Association for Information Systems, 13*(11), 923.

Balijepally, V., Mangalaraj, G., & Iyengar, K. (2011). Are we wielding this hammer correctly? A reflective review of the application of cluster analysis in information systems research. *Journal of the Association for Information Systems, 12*(5), 375.

Bian, J., Liu, Y., Zhou, D., Agichtein, E., & Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 51–60).

Blooma, M. J., & Wickramasinghe, N. (2014, December). Research issues in healthcare social question answering services. In *Australian Conference on Information Systems 2014*, Auckland, New Zealand.

Blooma, M. J., & Wickramasinghe, N. (2016). Prevalence of social question answering in healthcare social media. In *Contemporary consumer health informatics*. Cham: Springer.

Blooma, M. J., Chua, A. Y. K., Goh, D. H., & Wickramasinghe, N. (2016, October). Graph-based cluster analysis to identify similar questions: A design science approach. *Journal of the Association for Information Systems, 17*, 590.

Chan, J., Hayes, C., & Daly, E. M. (2010). Decomposing discussion forums and boards using user roles. *ICWSM, 10*, 215–218.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *STOC 02, Montreal, Quebec, Canada*.

De Vries, J. J. G., Geleijnse, G., Tesanovic, A., & Van de Ven, A. R. T. (2013, September). Heart failure risk models and their readiness for clinical practice. In *2013 IEEE international conference on healthcare informatics (ICHI)* (pp. 239–247). IEEE.

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal, 20*(4), 364–366.

Denecke, K., & Stewart, A. (2011). Learning from medical social media data: Current state and future challenges. In *Social media tools and platforms in learning environments* (pp. 353–372). Berlin: Springer.

George, D. R., Rovniak, L. S., & Kraschnewski, J. L. (2013). Dangers and opportunities for social media in medicine. *Clinical Obstetrics and Gynecology, 56*(3), 453. https://doi.org/10.1097/GRF.0b013e318297dc38.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337–355.

Friedman, C. P. (2012). What informatics is and isn't. *Journal of the American Medical Informatics Association, 20*(2), 224–226.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 3.

Hyyppä, M. T. (2010). *Healthy ties: Social capital, population health and survival*. New York: Springer.

Leung, K. W. T., Ng, W., & Lee, D. L. (2008). Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering, 20*(11), 1505–1518.

Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management, 51*(6), 801–809.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992a). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36–59.

Walls, J., Widmeyer, G., & El Sawy, O. (1992b). Building an information system design theory for vigilant EIS. *Information Systems Research*, *3*(1), 36–59. Retrieved from http://www.jstor.org/stable/23010780

Xu, J., Wang, G. A., Li, J., & Chau, M. (2007). Complex problem solving: identity matching based on social contextual information. *Journal of the Association for Information Systems, 8*(10), 524.

Zhang, J., & Zhao, Y. (2013). A user term visualization analysis based on a social question and answer log. *Information Processing and Management, 49*(5), 1019–1048.