Identifying Similar Questions in Healthcare Social Question Answering Services: A Design Science Research

Completed Research

Blooma John
University of Canberra,
Australia
Blooma.John@canberra.edu.au

Nilmini Wickramasinghe
Deakin University & Epworth
HealthCare, Australia
nilmini.work@gmail.com

Jayan Kurian
University of Canberra,
Australia
Jayan.Kurian@canberra.edu.au

Abstract

Healthcare Social Question Answering (SQA) are services where users can ask, respond and receive answers for their posts from other social media users in health domain. The activities of social media users such as asking, responding, liking and posting comments results in building reusable content. This study identifies similar content (i.e. questions) from user posts which contributes towards providing better health care services. For identifying similar questions, this study uses a quadri-link cluster analysis to analyze the attributes of questions, answers and users. A design science methodology was used to develop the algorithm and calculate the similarity measures. The results of cluster analysis based on the proposed similarity measures on a pilot data set indicate that identifying similar questions will be a contribution in the transition of traditional healthcare services into social media enabled healthcare services. The results exemplify the future of digital transformation in health care SQA.

Keywords

Design science, health care social media, social question answering, cluster analysis

Introduction

Social Question Answering (SQA) is an emerging field where social media users can ask, respond or rate comments posted by others which results in the generation of new content created by social media users. These services implemented in the health domain are described as Healthcare Social Question Answering (SQA) services where social media users can seek and share information, network with others having similar medical conditions and discuss their treatment progress and outcomes (John et al., 2016a). Moreover, users also network with medical professionals to discus and seek advice related to their medical condition (Denecke & Stewart, 2011). Thus, healthcare SQA services are providing a platform for user empowerment where users can seek and share health related information and compare and get advice on medical conditions and recovery strategies. MedHelp, WebMD and Drugs.com are some of the popular healthcare SQA services. In addition to health-related information inquiries, specific topics that are widely discussed in healthcare Social Question Answering services are physical activities, diet, smoking and alcohol consumption. (Zhang & Zhao, 2013; Hyyppä, 2010).

The need to identify similar questions is important as it will facilitate efficient retrieval of quality answers to user questions. For example, in Drugs.com, a healthcare SQA service, responses are stored in a repository. Further, it is also important to study the relationship between users and the content they post since quality and source of questions and answers are important in retrieving the best answer for a new question. Though earlier studies (Agichtein, 2008; Bian et al., 2009) have used different data sources to retrieve quality answers, studies that examine quality of answers in health domain is sparse. Hence to address the research gap in literature, this study focusses on SQA services in healthcare social media. The research question addressed in this paper is:

- **How to identify similar questions shared on healthcare social question answering services?**

In this paper, we propose a quadri-link cluster analysis based on features related to questions, answers and users, to cluster similar content. To do this in a systematic fashion we employed a design science research methodology (DSRM) approach (Rai et al., 2017). Based on out earlier work (John et al., 2016b; John and Wickramasinghe, 2018), the similarity measure was developed with a relationship network between the users and the content shared. We propose the similarity measure to cluster similar content based on design science research (Hevner et al., 2004).

Literature Review

Today, users share content in social media, making it the most powerful tool and the threats of imprecise diagnoses are challenging (George et al., 2013). Social media sites are important tools today for personal health information exchange (Kotsenas et. al., 2018). Promotion of public health knowledge and patient advocacy were the two aspects found to be prevalent in health care social media (Xu et al., 2016). During conversations in social media, people may communicate their personal health and a potential driver for this form of revealing is the immediate positive outcomes that it can provide for contributors and the community (Kordzadeh & Warren, 2016). As the public discourse continues to mature within these virtual spaces, it will be critical to identify similar content to open new opportunities to engage patients and health professionals for better outcomes. Social media is a powerful tool, and offers collaboration between users and a mechanism for social interaction in health-related communication for a range of individuals (Moorhead et al., 2013). Although health care social media is rich, previous work tend to miss either side of the coin, the health side and the technology side. Earlier work fails to attain sustainability because the studies neglects the connection between the technology and the people involved (de Vries et al., 2013). Hence, in this paper, we base our similarity measures with the two aspects. The quality of the content and the relationship of the content with the users by using design science guidelines.

Design science is a significant and genuine research paradigm in information systems (Gregor & Hevner, 2013). Design science is used to solve an identified problem by building socio-technical artefacts (Myers & Venable, 2014). Seven guidelines presented by Hevner et al. (2004) originated from information-systems design theory originally proposed by Walls et al. (1992). The steps proposed by Peffers et al. (2007) are problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. The design science theory as a methodology needs to be incorporating these steps. There is also a need to give a minimal procedure and a mental model for offering and assessing the design science research (Peffers et al., 2007). As we propose cluster analysis, the design science theory provides a groundwork for thoroughly specifying its design. Thus, this paper presents an instantiation of solving the problem of identifying similar questions to reuse the answers by developing computational model and algorithm termed as quadri-link cluster analysis, based on Arnott and Pervan (2012) and Xu et al. (2007). This is a classic example of computational genre of design science research (Rai, 2017). Computational design is the sciences of the artificial for changing existing situation into preferred ones by extending cluster analysis (John et al., 2016b) into health care SQA services.

Cluster analysis is a process of grouping together similar items into meaningful clusters based on similarities among the items (Balijepally et al., 2011). Cluster analysis is a widely used method in health-related projects. For example, Ahlqvist et al., (2018) used cluster analysis to identify novel subgroups of adult-onset diabetes and their association with outcomes. Cluster analysis is used in information systems research as an analytical tool for classifying configurations of numerous entities that include the information technology artefact. In addition, entities might be clustered based on unsupervised learning, using techniques such as hierarchical and k-means clustering. Regarding the detection of relevant content and communities, clustering methods such as hierarchical, k-means and fuzzy c-means clustering play a vital role (Prabhu et al. 2010). However, cluster analysis still faces challenges in discovering changing clusters in large-scale, dynamic and big data. Synchronized and unsynchronized trends that detect topics that are popular among highly clustered and distributed users evolved. As the questions asked in SQA services are very personal and detailed, keywords alone do not offer a dependable basis for clustering questions (Bian et al., 2009; Agichtein et al., 2008). This is even more significant for clustering questions and answers related to health. To overcome the disadvantages of keyword-based clustering, extant research focuses on additional criteria. John et al. (2016b) uses content and user relationship while Leung et al. (2008) presented the notion of concept-based graphs for clustering. Mutually coupled bipartite network was used by Bian et al. (2009). However, very few studies used cluster analysis in social media related to health questions for identifying similar questions in health care SQA services.

We propose a new cluster analysis by taking the relationship between the questions, answers, askers and answerers to identify similar questions based on design science research as detailed in the methodology.

Methodology

The four distinct sets of entities used for cluster analysis are questions, answers, users and concepts. In this model, we combine both answerer and asker into one entity – user. The reason for combining the answerer and asker is mainly because the actual role of the user is not determined a priori in this extended model when compared to John et al., (2016b). The features related to the users' social role will help us to identify the user as asker or answerer. Yet another addition in this model is that we used concepts as a new entity that contains rich information. We shortened the concept extraction as the extraction of unique non-stop words. The initial design of this algorithm is presented in our preliminary work (John and Wickramasinghe, 2018).

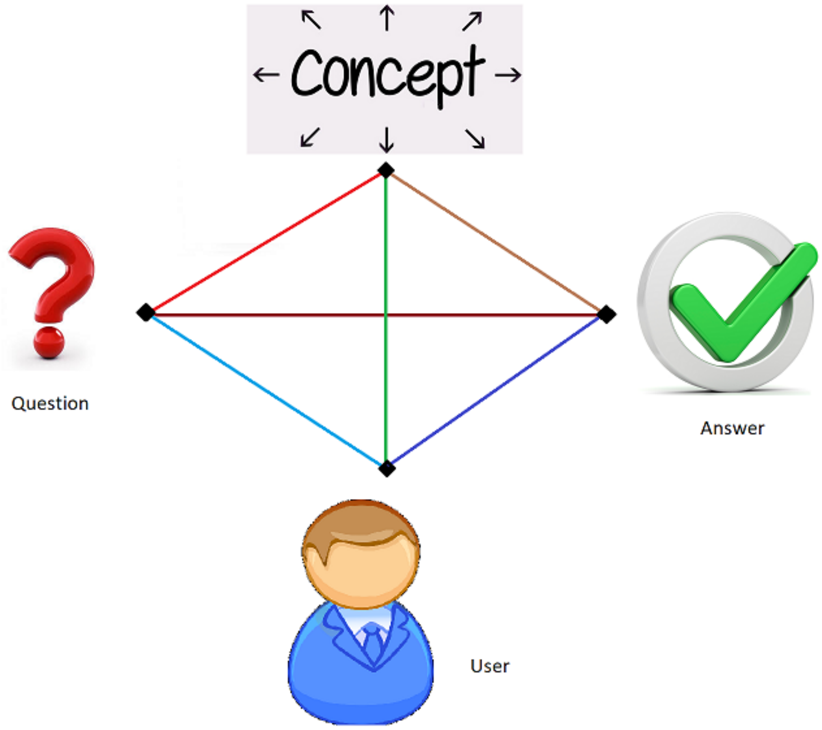

Figure 1. Quadri-Link Model

Features for questions are related to intrinsic content quality metrics (John et al., 2016b; Agichtein et al., 2008). First set of features are text related and involves the questions and answers. In addition to the text, there is a wide array of non-content information referred to as meta data available. Meta data ranges from links between items to explicit and implicit features of the content. To list a few features related to meta data are posting time, questions, votes, punctuation, typos and semantic complexity measures. For answers (Bian et al., 2009; Agichtein et al., 2008), we consider the intrinsic content quality metrics as well as non-content information. We also used the relationship features of questions and users. To consider the relationship between the question and its answer, we used word overlap and the ratio between the lengths of the question and the answer as features. The relationship between the answer and users are measured by using the number of positive votes and negative votes. The user features focused on the structural network properties of a user proposed are the features in-degree, out-degree, hub score, authority score and initialization (Angeletou et al. 2011; Chan et al., 2010). In-degree denotes the number of other users answered by this user. Out-degree denotes the number of other users who answered this user. The hub score and authority score is denoted by the score computed by the HITS algorithm for the user. Average votes per answer is calculated by the votes per answer divided by the total number of answers. Initialization is calculated by the number of questions asked by a user divided by all questions. In-degree score is calculated as the in-degree of this user divided by unique in-degrees. The quality of the cluster analysis based on the user features and the questions and answered related to the users helped improve the clustering outcome. The features used for questions, answers and users are given in Table 1 and extended from our preliminary work (John and Wickramasinghe, 2018).

Question Features	**Answer Features**	**User Features**
Number of words in the question	Overlap of words between question and answer	Number of Questions asked by the user
Length of the question subject	Number of comments for the answer	Total answers answered by the user
Length of the question detail	Total positive votes for the answer	Votes gained by the user
Question posting time	Total negative votes for the answer	In-degree score of the user
Question votes	Answer length	Out-degree score of the user
Number of answers for the question	Number of unique words in the answer	Hub score of the user
Punctuation density of the question	Question Answer (QA) ratio	Authority score of the user
Category of the question	Number of words per sentence in the answer	Average votes per answer
Number of words per sentence in the question	Capitalization errors in the answer	Initialization score of the user
Capitalization errors in the question	The Flesch–Kincaid (F–K) reading grade level of the answer	% In-degree score of the user
The Flesch–Kincaid (F–K) reading grade level of the question		

Table 1. Features related to Questions, Answers and Users

To calculate the relationship between the content and users, as illustrated in figure 1, we divided the content and users into four distinct sets of entities and six distinct types of links. The concepts in figure one represents unique non-stop words used. Questions, answers and users are mapped along with the concepts as given in the figure 1. The similar measures for the quadri-link model computes the similarity score between two sets of questions, answers or users by considering two components. We used the Jaccard similarity index (for a discrete set) and its extended general form for non-negative real values as suggested by Charikar (2002). The equations to calculate similarity between answers (sim(a_i,a_j)), questions (sim(q_i,q_j)) and users (sim(u_i,u_j)), are given in equation 1, equation 2 and equation 3 respectively. The flow of the steps used for quadri-link cluster analysis is illustrated in the flow chart given in Figure 2.

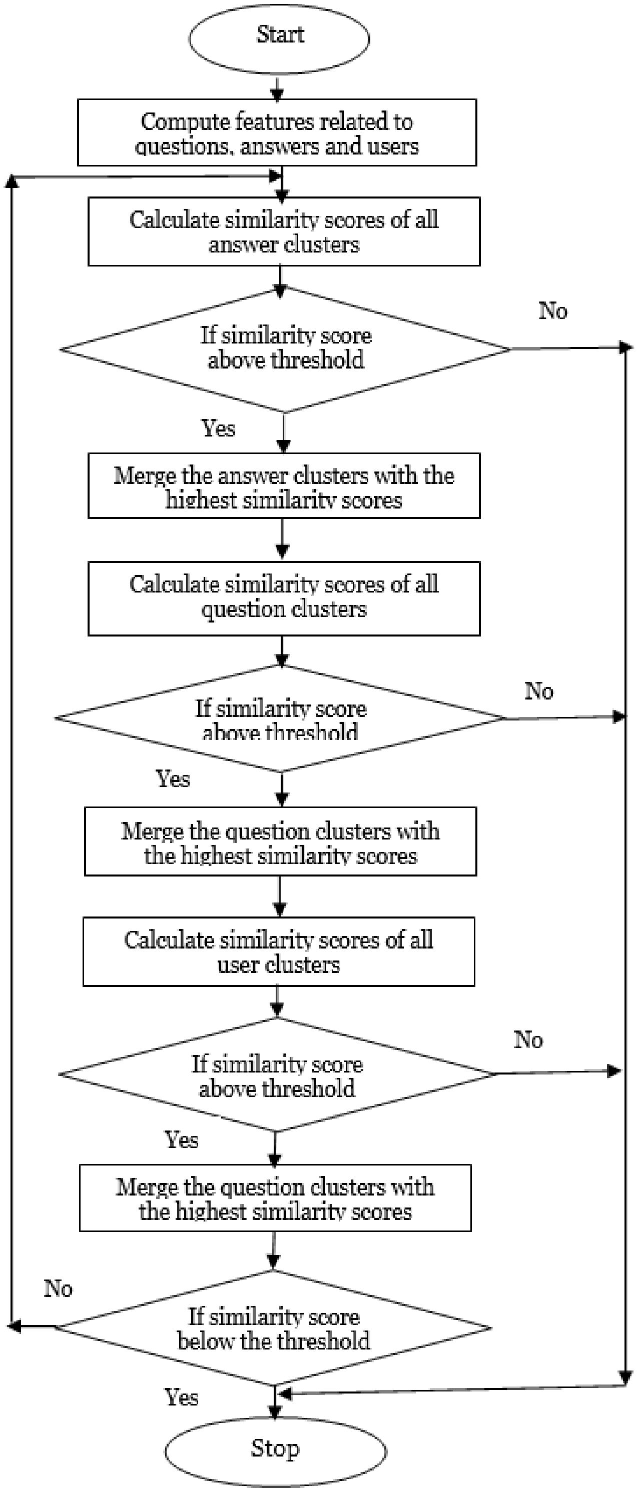
Start
Compute features related to questions, answers and users
Calculate similarity scores of all answer clusters
If similarity score above threshold
No
Yes
Merge the answer clusters with the highest similarity scores
Calculate similarity scores of all question clusters
If similarity score above threshold
No
Yes
Merge the question clusters with the highest similarity scores
Calculate similarity scores of all user clusters
If similarity score above threshold
No
Yes
Merge the question clusters with the highest similarity scores
No
If similarity score below the threshold
Yes
Stop

Figure 2. Flowchart illustrating the proposed quadri-link cluster analysis

$$Sim(a_i, a_j) = \frac{Common\ concepts}{Total\ distinct\ concepts\ of\ both\ ai\ and\ aj} + \textit{If their questions are the same or in the same cluster 0 else 1}$$

$$\textit{If their users are the same or in same cluster 0 else 1} + \textit{Jaccard similarity of features in the framework} \qquad \text{Equation 1}$$

$$Sim(q_i, q_j) = \frac{Common\ concepts}{Total\ distinct\ concepts\ of\ both\ qi\ and\ qj} + \textit{If their answers are the same or in the same cluster 0 else 1}$$

$$\textit{If their users are the same or in same cluster 0 else 1} + \textit{Jaccard similarity of features in the framework} \qquad \text{Equation 2}$$

$$Sim(u_i, u_j) = \frac{Common\ concepts}{Total\ distinct\ concepts\ of\ both\ ui\ and\ uj} + \textit{If their questions are the same or in the same cluster 0 else 1}$$

$$\textit{If their answers are the same or in same cluster 0 else 1} + \textit{Jaccard similarity of features in the framework} \qquad \text{Equation 3}$$

We used the complete linkage similarity measures (Defays, 1977). We selected the order of clustering as given in the flow chart (Figure 2). As the user score is computed after answers and questions because the user score requires information from the question and answer clusters. Hence, the user score will be computed last in the iteration. Correspondingly, questions require information from answers and thus question's similarity score will be calculated second. Subsequently, the score of answers will be calculated first. Therefore, the order of clustering is first, answer, then, question and finally, user. Thus, proposed clustering algorithm is called quadri-link cluster analysis. Threshold used to terminate the flow of clustering is based on the similarity measure of the questions, answers and users (Gower at al.,1969). Threshold determines the stopping condition of the algorithm. In quadri-link cluster analysis, the similarity measure threshold can be variated based on the need.

Results

For this pilot study, we collected openly accessible data from Drugs.com. Drugs.com is one of the popular source of drug information available online. It provides free, peer-reviewed and accurate data on more than 24,000 prescription drugs. The social question answering service in Drugs.com is also popular in asking and answering questions by various users. We collected questions answers and user details from Drungs.com and tested the quadri-link cluster analysis to identify similar questions. The pilot study used 200 resolved questions related to obesity for testing the proposed quadri-link cluster analysis. Based on the data collected, the features listed in Table 1 for questions, answers and users were calculated before we tested the proposed quadri-link cluster analysis.

We then evaluated the results of the quadri-link cluster analysis using the sample dataset. The top ten clustered questions are listed in Table 2. We compared the results by using quadripartite cluster analysis (John et al., 2016). We recruited two expert medical practitioners to rate the similarity of the questions. The medical experts were given a sample set of 10 questions and were asked to evaluate. Their evaluation received an inter-rater agreement 80%. After they evaluated, their evaluation was compared and discussed with respect to the deviations. They were also asked to write justification about their decision as they evaluated the rest of the questions independently. Once the two evaluators were clear about the process, they evaluated all 200 questions independently. After their evaluation was over, they had to do a few rounds to meetings to discuss regarding the discrepancies and finally come with the concluding evaluation that both agree.

The results and findings from a pilot testing of the proposed quadri-link cluster analysis are discussed as follows. We evaluated the accuracy of the clustered questions by reading through the questions and the various answers obtained for the questions as well as the profile of the users involved in asking the question as well as answering and commenting on the question. We found that most of the questions users asked contained a detailed description of a personal problems. 79% of the questions had detailed descriptions. The average word count of questions was 56. The question with the longest description had 788 words. The longest question was about a detailed explanation about diet, exercise and Metformin. An example for long history is "….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,…". The long questions also had comprehensive answers, with the lengthiest answer having 505 words. A question that was asked in 2012 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. So, users are referring to questions even after years to gain insights from similar cases.

In quadripartite cluster analysis, the most similar questions 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, there is an immense need to put more emphasis on the medical terms. Second 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.

Question 1	**Question 2**
"Does Trazodone cause constipation or weight gain?"	*"Does Buspar cause weight gain or constipation?"*
"Kindly inform me about the recommended medicine that causes weight to lose?"	*"Just to get rid of obesity, which drug is beneficial?"*
"How much weight gain to expect taking Loestrin Fe?"	*"How much weight gain should I expect from Valtrex?"*
"Does cyclobenzaprine cause weight gain?"	*"Does polyethylene glycol 3350 cause weight gain?"*
"Feedback on Prozac?"	*"Prozac--- like it?"*
"Does Cymbalta (60mg) cause weight gain or cause water?"	*"Does this Effexor cause weight gain?"*
"Does Lupron injections cause weight gain in women?"	*"Does Trileptal or the Generic cause weight gain?"*
"Does it make you put on weight omeprazole?"	*"Does Fosamax cause you to put on weight?"*
"Does Lupron injections cause weight gain in women?"	*"Does this drug cause increased appetite or weight gain?"*
"Does cyclobenzaprine cause weight gain?"	*"Does daily dosage of 25mg of spironolactone cause weight gain?"*

Table 2. Top ten similar questions identified by quadri-link cluster analysis

On the other hand, 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?"*. These two questions had word match. On analysis, it was interesting to note that both questions were asked by the same user on the same day. The questions were answered by different users. However, the answers were not the same. To add on the answers agreed that the drugs caused weight gain or constipation. This example is an evidence of the accuracy of quadri-link cluster analysis. On expert evaluation, the questions mention about the side effects of two different anti-depressants. But Trazodone is drug to treat depression and Buspar is a drug to treat Anxiety disorders. Buspar may be used to enhance the effectiveness of antidepressants in certain patients. Another example 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. The feedback from expert evaluation was that both questions are not formulated as a question, informal language is used with limited information about the answer they are seeking. Yet another example is a case where the questions were asked and answered by different users. The questions are *"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"*. Experts evaluated as similar questions because both questions were asking for the side effects and information about Prozac.

It was also found that the quadri-link cluster analysis was not able to differentiate between medical terms. For example, *"Does cyclobenzaprine cause weight gain?"* and "*Does polyethylene glycol 3350 cause weight gain?"* were clustered. Based on expert evaluation, the questions are similar as both the questions ask about weight gain as a side effect for two drugs. However, there are no similarities in the questions with respect to the medical terms. Cyclobenzaprine is a muscle relaxant and PG 3350 is a laxative for occasional constipation. Another example, "*Does cymbalta (60mg) cause weight gain?"* and "*Does this effexor cause weight gain?"* were found similar. In this case, both the questions mention about the weight gain effects of anti-depressants. However, experts also added that Cymbalta and Effexor are two different antidepressants in different classification. Their actions are not similar and thus their side effects. Another example is *"Does Lupron injections cause weight gain in women?"* and *"Does Trileptal or the Generic cause weight gain?"*. First question is about a drug for the treatment of cancer and its side effect as weight gain in females and second question mention about an anti-convulsant and its side effect as weight in general. Lupron is a drug used to treat patients with cancer and Trileptal is an anticonvulsant generally used for anxiety disorders. Hence, to conclude it is clearly evident that quadri-link cluster analysis is more accurate than quadripartite cluster analysis. However, there is a need to accentuate the use of a medical thesaurus so that the medical terms are used to avoid issues highlighted in clustering questions that use similar words except for the medical terms (Zhang and Zhao, 2013; John and Wickramasinghe, 2018).

Conclusion

The findings of the cluster analysis are best explained based on the six activities of design science research as detailed by Peffers et al. (2007).

- *Problem identification and motivation:* The problem identified in this paper will help in understanding the linkage between the digital artifcat designed and its human experiences by embracing design thinking practices.
- *Define the objective of a solution:* Developing quadri-link cluster analysis based on the relationship between content and the users defines the objective of the solution that result in a viable artefact in the form of "quadri-link cluster analysis".
- *Design and Development:* The designed artefact in this paper is an instantiation of quadri-link cluster analysis. We identified the list of features used with respect to questions, answers and users. We framed the quadri-link model to calculate the similarity measure. We finally developed the similarity measure and the algorithm to cluster similar questions as illustrated in the flowchart.
- Demonstration: We demonstrated the efficacy of the artifact to solve the problem of clustering user generated questions in health care SQA services by experimentation of this pilot study.
- *Evaluation:* We evaluated the artefact based on 200 questions and related features collected from Drugs.com. The evaluation demonstrated clear evidence of improved precision of quadri-link cluster analysis.

- *Communication:* We published the algorithm and its applicability on healthcare social media. This study can be extended to apply and evaluate other types of content in social media to cluster similar content, like tweets and comments in social media (Schneeweiss, 2014).

This research has implications for theory and practice. Potential benefits of identifying similar content in health care SQA services improves the advantages of digitization in healthcare social media (Mihailescu et al., 2017). The implications are manifold and include increases in productivity of health care services, innovations in value creation, as well as novel forms of interactions with patients and customers, among others. Identifying similar content using the proposed quadri-link cluster analysis, will be a part of transforming digital content as patients have become empowered consumers (Matt et al., 2015). Today, patients and professionals are armed with access to instant information and big data thus leading to growing their expectations for personalized engagement in all aspects of their lives (Murdoch, et al., 2013). Identifying and delivering the needs of customers tend to shift to value-based care by meeting patient and population wellness objectives and superior healthcare delivery. A key success in learning from health care SQA services will be to remain focused on our goal of gaining actionable insights into the best ways to treat the health issue. This can eventually be extended to the big health care data (Schneeweiss, 2014). Furthermore, 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 social media means that this is a benefit that most, if not all, people can enjoy. In many ways, it has the potential to revolutionize current healthcare delivery practices. In addition, it has a key role to play regarding public health and enabling digital transformation and change of lifestyle for all.

To conclude, as Sir Isaac Newton famously said that he has only seen further by “standing on the shoulders of giants.”, this is only a stepping stone towards facilitating the use of filtered archived answers in health care SQA services regarded as explicit knowledge for the given target questions based on the proposed quadri-link cluster analysis.

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, 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), pp. 923.

Ahlqvist, E., et. al., 2018. “Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables,” *The Lancet Diabetes & Endocrinology*, 6(5), pp. 361-369.

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), pp. 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.

De Vries, J. J. G., Geleijnse, G., Tesanovic, A., & Van de Ven, A. R. T. 2013. Heart failure risk models and their readiness for clinical practice. In Healthcare Informatics (ICHI), 2013 IEEE International Conference, pp. 239-247. IEEE.

Chan, J., Hayes, C. & Daly, E. M. 2010. “Decomposing Discussion Forums and Boards Using User Roles,” *In Proceedings of the ICWSM*, 10, pp. 215–218.

Charikar, M.S. 2002. “Similarity estimation techniques from rounding algorithms.” *In Proceedings of the STOC 02,* Montreal, Quebec, Canada.

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. Springer Berlin Heidelberg.

Gower, J., & Ross, G. 1969. “Minimum Spanning Trees and Single Linkage Cluster Analysis”. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 18(1), pp. 54-64.

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

George, D. R., Rovniak, L. S., & Kraschnewski, J. L. 2013. “Dangers and opportunities for social media in medicine,” *Clinical Obstetrics and Gynecology*, 56(3).

Hevner, A. R., March, S. T., Park, J, & Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly*, (28: 1).

Hyyppä, M.T., 2010. “Healthy ties: Social capital, population health and survival”. *Springer Science & Business Media.*

John B., Gururajan R., Wickramasinghe N. 2016a. “The Prevalence of Social Question Answering in Health-Care Social Media,” In: Wickramasinghe N., Troshani I., Tan J. (eds) *Contemporary Consumer Health Informatics*. Healthcare Delivery in the Information Age. Springer, Cham.

John B., Chua, A. Y. K., Goh, D. H. & Wickramasinghe, N. October 2016b.“Graph-based Cluster Analysis to Identify Similar Questions: A Design Science Approach”, *Journal of the Association for Information Systems:* 17 (9), Article 2.

John, B., & Wickramasinghe, N. 2018. “Clustering Questions in Healthcare Social Question Answering Based on Design Science Theory*,” In Theories to Inform Superior Health Informatics Research and Practice*, pp. 95-108. Springer, Cham.

Kotsenas, A. L., Arce, M., Aase, L., Timimi, F. K., Young, C., & Wald, J. T. 2018. “The strategic imperative for the use of social media in health care,” *Journal of the American College of Radiology,* 15(1), pp. 155-161.

Kordzadeh, N., Warren, J. and Seifi, A. 2016. “Antecedents of privacy calculus components in virtual health communities,” *International Journal of Information Management*, 36(5), pp.724-734.

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), pp. 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), pp. 801-809.

Matt, C., Hess, T., Benlian, A. 2015. “Digital Transformation Strategies,” *Business and Information Systems Engineering*, 57(5), pp. 339–343.

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. 2013. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. *Journal of Medical Internet Research,* 15(4), e85.

Mihailescu, M., Mihailescu, D., & Carlsson, S. 2017. “Understanding Healthcare Digitalization: A Critical Realist Approach”. *In Proceedings of the 2017 International Conference on Information Systems.*

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *The Journal of the American Medical Association*, 309(13), pp. 1351-1352.

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) pp. 77.

Prabhu, J., Sudharshan, M., Saravanan, M., & Prasad, G. 2010. “Augmenting rapid clustering method for social network analysis,” *In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference* pp. 407-408. IEEE.

Rai. A. (2017) Diversity of Design Science Research. *MIS Quarterly*, 41(1).

Schneeweiss, S. 2014. “Learning from Big Health Care Data,” *The New England Journal of Medicine* 2014, 370, pp. 2161-2163.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. 1992. “Building an information system design theory for vigilant EIS,” *Information systems research,* 3(1), pp. 36-59.

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), pp. 524.

Xu, S., Markson, C., Costello, K. L., Xing, C. Y., Demissie, K., & Llanos, A. A. 2016. Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter. *JMIR Public Health and Surveillance,* 2(1), e17.

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