Similarity of medical concepts in question and answering of health communities

Hamid Naderi
Mashhad University of Medical Sciences, Iran

Sina Madani
Vanderbilt University Medical Center, USA

Behzad Kiani
Kobra Etminani
Mashhad University of Medical Sciences, Iran

Abstract
The ability to automatically categorize submitted questions based on topics and suggest similar question and answer to the users reduces the number of redundant questions. Our objective was to compare intra-topic and inter-topic similarity between question and answers by using concept-based similarity computing analysis. We gathered existing question and answers from several popular online health communities. Then, Unified Medical Language System concepts related to selected questions and experts in different topics were extracted and weighted by term frequency -inverse document frequency values. Finally, the similarity between weighted vectors of Unified Medical Language System concepts was computed. Our result showed a considerable gap between intra-topic and inter-topic similarities in such a way that the average of intra-topic similarity (0.095, 0.192, and 0.110, respectively) was higher than the average of inter-topic similarity (0.012, 0.025, and 0.018, respectively) for questions of the top 3 popular online communities including NetWellness, WebMD, and Yahoo Answers. Similarity scores between the content of questions answered by experts in the same and different topics were calculated as 0.51 and 0.11, respectively. Concept-based similarity computing methods can be used in developing intelligent question and answering retrieval systems that contain auto recommendation functionality for similar questions and experts.

Corresponding author:
Kobra Etminani, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, 564-7948-917, Iran.
Email: EtminaniK@mums.ac.ir

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Introduction

A great number of people seek their health-related questions from online resources when they face a specific health condition and do not have enough knowledge to answer it.1 These types of healthcare consumers play a more effective role in managing their disease(s) and treatment plans. Therefore, they need to gather and process health-related information from various resources and make a final decision.2,3 Also, it has been shown that 71 percent of health consumers between the ages of 18 and 20 years look for health information in online resources.4 However, despite the availability of healthcare information on various websites, consumers are often confused and cannot find what exactly they are looking for.5,6 Several studies showed7–9 that most consumers tried to find health information by searching two or three keywords that hardly describe their actual medical condition. The keywords used in these simple searches typically do not match to the structured content of online databases; hence, the query returns unrelated results. Therefore, the performance of searching health information by consumers in online search engines is typically low.

Online question and answering (Q&A) communities are rapidly growing. Users can ask their healthcare questions in various topics in Q&A websites, sometimes anonymously. They usually can find answer for their questions quickly without any cost. History of questions which has been sent to online communities is a valuable resource to capture consumers’ vocabularies in a given domain. However, the lack of standard concepts in such questions makes it difficult for consumers to find question relevant to a particular health problem. Therefore, one of the most challenging hurdles in online Q&A communities is duplicated questions. Such duplicates cause a large number of similar questions piling up on websites which makes the search and browse for other consumers quite frustrating. Similarly, medical experts need to find comparable questions and respond with relevant information. In these instances, online community’s support team need to find a way to classify similar Q&As through a well-defined content management framework. Therefore, identification of similar questions in Q&A retrieval systems is essentially important from both consumers’ perspective and medical experts.

A retrieval system can automatically suggest similar Q&As to users by considering inter-topic similarity. Also, instead of having a backup team manually sorting questions into different topics, the system can automatically categorize questions in various topics based on inter-topic similarity. The goal of this study was to compare inter- and intra-topic similarity between Q&As submitted to the online health communities. We used a concept-based similarity computing method to facilitate finding similar healthcare-related questions and medical experts.

Method

Figure 1 shows the step by step process of Q&A gathering, concept extraction, and similarity calculation.
Step 1: source Q&A communities

Yahoo Answers was launched in 2005 as a major social Q&A website that covers a wide range of domains and has the most number of visits among Q&A websites in United States. Health category in Yahoo Answers is a major topic with more than 20 subcategories such as diabetes, women’s health, and cancer. WebMD is another major health information website that was created more than 20 years ago. This online health community has more than 100 million visitors per month that seek diverse health-related services such as Q&A on various health topics, medication information, and clinical symptom. NetWellness is a non-profit community health service with around 500 healthcare subject matter experts. The website was funded by three universities, including Case Western Reserve University, Ohio State University, and the University of Cincinnati. More than 70,000 questions in distinctive health domains have been answered by medical experts since 1995 on the website. Content evaluation of NetWellness, based on feedbacks received from end users, shows Q&As posted on this website were greatly helpful for healthcare consumers.

Step 2: Q&A random selection

A PHP script was written to collect Q&As from source communities and store them in a MySQL database (version 5.6). Subsequently, 3000 Q&As were randomly selected for our study. The health topics of the three sources (Yahoo Answers, WebMD, and NetWellness) were not exactly the same. Therefore, we randomly selected 75 Q&As (five Q&As for each topic) to calculate the inter-topic and intra-topic similarities. Also, 400 Q&As were randomly selected to calculate expert–expert similarities. For example, questions 1 and 2 (provided in the Online Supplementary Material) are two sample questions extracted from NetWellness community. Our objective was to extract Unified Medical Language System (UMLS) concepts from these questions and calculate similarity scores among them based on the steps depicted in Figure 1.

Step 3: concept extraction and mapping

In this step, selected Q&As were inserted into a concept extraction and mapping pipeline where MetaMap was used to extract clinical entities and map them to UMLS Concept Unique Identifiers.
(CUIs). MetaMap data source was restricted to Consumer Health Vocabulary (CHV) terminology as the most relevant resource for mapping consumer healthcare terms to the equivalent controlled vocabularies. MetaMap processed the two sample questions (provided in the Online Supplementary Material). Each health consumer term was mapped to a UMLS CUI and equivalent preferred term from controlled vocabularies. MetaMap assigns a threshold of up to 1000 which indicates centrality and coverage of concept in the processed text. Supplementary Table 1 contains the output of concept extraction and mapping phase for the content of sample questions.

**Step 4: term frequency–inverse document frequency calculation**

Term frequency–inverse document frequency (TF-IDF) score indicates the importance of a term based on the number of word occurrences in a document. Considering this frequency indicator, a concept that appears several times in a question gets a higher weight. On the other hand, a general concept that appears frequently in most of the questions is considered as a common concept with low weight. In this phase, TF-IDF weighted vectors were calculated based on extracted concepts for questions (question TF-IDF calculation) and experts (expert TF-IDF calculation). Equation (1) shows how to compute TF-IDF weight of a concept

\[
W_{TF-IDF}(t) = TF(t) \times IDF(t)
\]

where the value of \( TF \) and \( IDF \) for a specific term is calculated based on equations (2) and (3), respectively

\[
TF(t) = \left( \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Maximum number of terms in the document}} \right)
\]

\[
IDF(t) = \log_2 \left( \frac{\text{Total number of documents}}{\text{Number of documents with term } t} \right) + 1
\]

For example, TF-IDF for concept “C0023884” in question 1 which maps to “Liver” preferred term is calculated as follows

\[
TF(C0023884 \text{ in question } 1) = \frac{1}{8}
\]

\[
IDF(C0023884 \text{ in question } 1) = \log_2 \frac{3000}{35} + 1
\]

\[
W_{TF-IDF}(C0023884 \text{ in question } 1) = \left( \frac{1}{8} \right) \times \left( \log_2 \frac{3000}{35} + 1 \right) = 0.92
\]
All these steps were performed on concepts extracted from selected Q&As. Supplementary Table 2 shows TF-IDF numbers for extracted CUIs of sample questions.

MetaMap weights (threshold numbers) were normalized in the range of [0–1]. Then, the total weight of a concept is obtained by multiplying the MetaMap normalized weight and the TF-IDF weight.

**Step 5: similarity calculation**

The similarity measure is a function that computes the degree of similarity between two vectors of UMLS concepts. UMLS concepts relevant to a question and expert are considered as weighted vectors. So, similarity measure was used to represent the similarity between a pair of questions and experts.

- **Question–question similarity calculation**

In this phase, similarity between two questions is calculated. This similarity is divided into question–question intra-topic similarity and question–question inter-topic similarity. In question–question intra-topic similarity, Q&As of the same topics were randomly selected from our corpus and the similarity between them were computed. Also, to compute question–question inter-topic similarity, Q&As from different topics of source community were selected and cosine similarity between them was considered. Supplementary Figure 1 shows the calculation method of question–question similarity between vectors of questions concepts.

- **Expert–expert similarity calculation**

We generated profile for some experts based on answered questions. Then we computed the similarity between profile of experts in same (expert–expert intra-topic similarity) and different topics (expert–expert inter-topic similarity). Expert–expert similarity displays the similarity of the concepts in the questions answered by the experts. This similarity can help find experts with similar profiles and activities. Intra-topic similarity shows the similarity between content of questions answered by experts in the same topic. On the other hand, inter-topic similarity calculates the similarity between concepts of questions that are answered by experts in different health topics. Supplementary Figure 2 shows the calculation method of expert–expert similarity between vectors of extracted concepts related to experts.

For example, cosine similarity between the TF-IDF vectors of the two sample questions was calculated based on equation (4)

$$\text{Similarity}(q_1, q_2) = \frac{q_1 \cdot q_2}{\|q_1\| \cdot \|q_2\|}$$

where $\|q_1\|$ was computed by following formula and $w_i^2$ is square of total weight for each concept

$$\|q_1\| = \sqrt{\sum_{i=1}^{n} w_i^2}$$
Table 1 shows the final weighted vectors for questions 1 and 2. MetaMap assigns a threshold number to each extracted concept from 600 to 1000. This value indicates coverage and centrality of the concept in the processed text of the question. In this table, MetaMap average is the average of threshold numbers assigned by MetaMap for extracted concepts. Also, total weight is an aggregated weight that for each concept was calculated by multiplying MetaMap weight (semantic weight) by TF-IDF weight. Finally, TF-IDF weight is a numerical value that indicates the importance of a concept based on TF-IDF method.

Finally, the similarity between these two sample questions based on extracted concepts is 0.38

\[
\text{Similarity}(q_1, q_2) = \frac{11.56}{5.81 \times 5.12} = 0.38
\]

The last step is calculating cosine similarity between two questions or experts. Supplementary Figure 1 illustrates all of these steps.

**Result**

In our study, Yahoo Answers, WebMD, and NetWellness were considered as the source of Q&A communities. One thousand questions, together with their answers, were randomly selected from each source. First, question–question similarity was calculated in two steps: intra-topic similarity
and inter-topic similarity. In inter-topic similarity, 75 Q&As from 15 different topics (five Q&As in each topic) were randomly selected from our corpus and the similarity between questions of the same topic for each community was computed. For example, the similarity between selected Q&As of “Breast Cancer” topic in NetWellness community was 0.157. Also, the average of intra-topic similarity was calculated separately for each source community. Supplementary Table 3 shows computed values of intra-topic similarity for selected topics.

Subsequently, 75 questions were selected from 15 topics mentioned above to compute the similarity between questions of different topics. Supplementary Table 4 contains the result of inter-topic similarity for a few selected topics. For example, the average similarity between Q&As of “Diabetes” and “Breast Cancer” in NetWellness was 0.015. Supplementary Figure 3 shows the similarity distribution in details.

Our result showed a considerable gap between intra-topic and inter-topic similarity. This difference in the similarity of the two groups was apparent in nearly all three studied sources.

In the second phase of our study, we computed expert–expert similarities. To calculate the similarity between questions that had been answered by an expert, 400 answered questions by 20 experts (20 Q&As for each expert) were randomly selected from NetWellness. Supplementary Table 5 indicates distribution of selected experts in different health topics.

After the concept extraction phase for selected Q&As, TF-IDF weighting vectors were computed based on equation (2) for each expert. Similarity between experts was also calculated according to the cosine similarity between their TF-IDF vectors. Table 2 shows similarity numbers as a 20 × 20 matrix considering expert-related topics. Since the similarity matrix is symmetric, the similarity results were represented by an upper triangular matrix which all the entries below the main diagonal are zero. The value of each cell in this matrix indicates the similarity between experts in related row and column. The similarity between a vector and itself is 1; therefore, the values on the main diagonal of similarity matrix are all 1. As an example, the similarity between expert 2 and expert 3, who are both represented “Breast Cancer” topic, is 0.64. On the other hand, the similarity between expert 3 and expert 9, who are assigned to other topics “Breast Cancer” and “Diet and Nutrition,” are calculated as 0.04.

Figure 2 shows the integrated intra-topic and inter-topic similarity of the mentioned experts. The size of the circles represents the degree of similarity. Larger circles indicate greater similarity, where they can be seen in the main diagonal.

**Discussion**

In this study, the similarity between submitted questions by healthcare consumers in online communities was investigated. There was a significant difference between intra-topic similarity and inter-topic similarity which could be helpful for finding similar questions and experts. Therefore, it seems that concept-based similarity computing method has potentials in developing intelligent Q&A retrieval systems that contain auto recommendation of similar questions and experts.

There are three main steps in health information search process: (1) selecting information sources, (2) evaluating collected data, and (3) processing and preparing information for users. People usually ask about their health related problems from several sources. The result of a study which examined consumer health information searching practice showed that online users usually consult more than three sources during their request for information. The majority of consumers (90.5%) will usually look to a second source, where more than half of them (76.2%) consult a third source to find their intended information. There are several reasons for such behavior. First, they do not get a compelling answer from a given resource. Second, most of health consumers seek for a more comprehensive answer or read additional opinions. Third, some people hop from one
Table 2. Expert–expert similarity matrix.

| Topic                  | Expert | E1  | E2  | E3  | E4  | E5  | E6  | E7  | E8  | E9  | E10 | E11 | E12 | E13 | E14 | E15 | E16 | E17 | E18 | E19 | E20 |
|------------------------|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Breast Cancer          |        | 1.00| 0.61| 0.60| 0.15| 0.11| 0.06| 0.12| 0.12| 0.02| 0.10| 0.12| 0.07| 0.10| 0.08| 0.10| 0.05| 0.07| 0.04| 0.10| 0.11|
| Dental and Oral Health (Adults) |        | 1.00| 0.64| 0.12| 0.11| 0.05| 0.09| 0.06| 0.03| 0.09| 0.10| 0.07| 0.10| 0.07| 0.12| 0.08| 0.07| 0.05| 0.08| 0.09|
| Diet and Nutrition     |        |     |     | E3  | 1.00| 0.16| 0.15| 0.07| 0.08| 0.05| 0.04| 0.19| 0.19| 0.15| 0.15| 0.12| 0.18| 0.13| 0.09| 0.07| 0.17| 0.14|
| Digestive Disorders    |        |     |     |     |     |     |     |     |     |     | E4  | 1.00| 0.55| 0.45| 0.09| 0.05| 0.05| 0.11| 0.14| 0.13| 0.17| 0.20| 0.15|
| Ear, Nose, and Throat Disorders |        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.11| 0.10|
| Eye and Vision Care    |        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.11| 0.10|
| Kidney Diseases        |        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.11| 0.10|
Consumers usually evaluate the quality of online information (like Q&As seen on the websites) based on the credibility of author(s) and references. Also, appearance of similar results for a health condition search in several online sources of information may indicate the validity of the information. Thus, allocating an online environment where users can consolidate information from multiple resources with reliable references increases their confidence about such information. Concept-based similarity computing method can help developers to create a “Meta” Q&A health community from available online Q&A health societies. A lot of questions from several sources
can be automatically integrated in various topics by similarity between their concepts. This method will simplify information search process from multiple sources and facilitates information retrieval among a huge number of historical Q&A that are stored as valuable knowledge bases.

Comparing the results of intra- and inter-topic similarity could indicate that the similarity among questions of the same topics is significantly higher than the similarity between questions from different topics. Such results can be achieved by gathering questions from several Q&As communities as the source of information from various health topics. Also, expert–expert similarity has been calculated for different health topics. Our results indicate the similarity between the content of questions that were answered by experts for the same topic (intra-topic similarity) is higher than the similarity between concepts of questions that were answered by experts in a different health topic (inter-topic similarity) (Figure 2). Therefore, identification of similar experts can be done in an automated fashion. Online community’s web services, by using our method, can recommend similar experts to consumers who could not get compelling answer for their health problem from other experts. Finally, from the viewpoint of the online community organizer, calculating cosine similarity based on extracted concepts can help community support team programmatically define a process for categorizing questions and experts by topics.

**Limitation**

The sample questions in this study were only gathered from three popular online health communities. Considering submitted Q&As from more online communities can lead to a more accurate and comprehensive results.

**Conclusion**

Concept-based similarity computing methods can be used in developing intelligent Q&A retrieval systems that contain auto recommendation functionality for similar questions and experts. Also, applying concept-based similarity methods have potentials to integrate multiple communities where users can find additional information from similar questions that were asked in other communities.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethical approval**

This article does not contain any studies with human participants or animals performed by any of the authors.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by Medical Research Council of Mashhad University of Medical Sciences (grant number 951317).

**ORCID iDs**

Behzad Kiani https://orcid.org/0000-0002-8816-328X
Kobra Etminani https://orcid.org/0000-0003-2006-6229
Supplemental material

Supplemental material for this article is available online.

References

1. Wildemuth BM, de Biek R, Friedman CP, et al. Information-seeking behaviors of medical students: a classification of questions asked of librarians and physicians. *Bull Med Libr Assoc* 1994; 82(3): 295–304.
2. Anderson RM and Funnell MM. Patient empowerment: reflections on the challenge of fostering the adoption of a new paradigm. *Patient Educ Couns* 2005; 57(2): 153–157.
3. Lewis D, Chang BL and Friedman CP. Consumer health informatics. In: Lewis D, Eysenbach G and Kukafka R (eds) *Consumer health informatics: informing consumers and improving health care*. New York: Springer, 2005, pp. 1–7.
4. Fox S. The Social Life of Health Information, 2011. Available at: http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/ (accessed 12 June 2018).
5. Arora NK, Hesse BW, Rimer BK, et al. Frustrated and confused: the American public rates its cancer-related information-seeking experiences. *J Gen Intern Med* 2008; 23(3): 223–228.
6. Zeng QT, Kogan S, Plovnick RM, et al. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *Int J Med Inform* 2004; 73(1): 45–55.
7. Sillicence E, Briggs P, Fishwick L, et al. Trust and mistrust of online health sites. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Vienna, 24–29 April 2004, pp. 663–670. New York: ACM.
8. Toms EG and Latter C. How consumers search for health information. *Health Informatics J* 2007; 13: 223–235.
9. Eysenbach G and Kohler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 2002; 324(7337): 573–577.
10. Aizawa A. An information-theoretic perspective of tf–idf measures. *Inform Process Manag* 2003; 39: 45–65.
11. Zhang Y. Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. In: *Proceedings of the 1st ACM international health informatics symposium*, Arlington, VI, 11–12 November 2010, pp. 210–219. New York: ACM.
12. Corp WH. WebMD announces fourth quarter and year end financial results, 2015. Available at: https://www.prnewswire.com/news-releases/webmd-announces-fourth-quarter-and-year-end-financial-results-300040633.html (accessed 12 June 2018).
13. Lauckner C and Hsieh G. The presentation of health-related search results and its impact on negative emotional outcomes. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Paris, 27 April–2 May 2013, pp. 333–342. New York: ACM.
14. Lefebvre RC and Bornkessel AS. Digital social networks and health. *Circulation* 2013; 127: 1829–1836.
15. Lau AY and Coiera EW. Impact of web searching and social feedback on consumer decision making: a prospective online experiment. *J Med Internet Res* 2008; 10(1): e2.
16. Marine S, Embi PJ, McCuistion M, et al. Netwellness 1995–2005: ten years of experience and growth as a non-profit consumer health information and ask-an-expert service. *AMIA Annu Symp Proc* 2005; 2005: 1043.
17. Aronson AR and Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17(3): 229–236.
18. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001: 17-21.
19. Wu HC, Luk RWP, Wong KF, et al. Interpreting TF-IDF term weights as making relevance decisions. *J ACM Trans. Inf. Syst* 2008; 26: 1–37.
20. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *J Doc* 1972; 28: 11–21.
21. Salton G and Buckley C. Term-weighting approaches in automatic text retrieval. *Inform Process Manag* 1988; 24: 513–523.

22. Salton G and McGill MJ. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1986, p. 400.

23. Savolainen R. Time as a context of information seeking. *Libr Inform Sci Res* 2006; 28: 110–127.

24. Zhang Y. Consumer health information searching process in real life settings. *Proc Am Soc Inform Sci Technol* 2012; 49: 1–10.