Fulfilling information needs of online patients using domain knowledge in online health communities

Donghua Chen
Department of Information Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, P. R. China
Email: 15113181@bjtu.edu.cn

Runtong Zhang
Department of Information Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, P. R. China
Email: rtzhang@bjtu.edu.cn

Jiayi Feng
Department of Information Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, P. R. China
Email: 18113057@bjtu.edu.cn

Kecheng Liu
Informatics Research Centre, Henley Business School, University of Reading, Reading RG6 6UD, United Kingdom
Email: k.liu@henley.ac.uk

Abstract

Background: Online health communities (OHCs) experience difficulties in utilizing patient-reported posts to fulfill the information needs of online patients concerning health-related issues.

Objectives: We aim to propose a comprehensive method that leverages medical domain knowledge to extract useful information from posts to fulfill information needs of online patients.

Methods: A knowledge representation framework based on authoritative knowledge sources in the medical field for the OHC is proposed. On the basis of the framework, a health-related information extraction process for analyzing the posts in the OHC is proposed. Then, knowledge support rate (KSR) and effective information rate (EIR) are introduced as metrics to evaluate changes in knowledge extracted from the knowledge sources in terms of fulfilling the information needs of patients in the OHC.

Results: On the basis of a dataset with 372,343 posts in an OHC, experimental results indicate that our method effectively extracts relevant knowledge for online patients. Moreover, KSR and EIR are feasible metrics of changes in knowledge in terms of fulfilling the information needs.

Conclusions: The OHCs effectively fulfill the information needs of patients by utilizing authoritative domain knowledge in the medical field. Knowledge-based services for online patients facilitate an intelligent OHC in the future.

Keywords: online patients; information needs; online health communities; domain knowledge; information extraction

Key Messages

- Most online health communities cannot fulfill the information needs of online patients concerning health-related issues.
- Patient-reported posts contain useful semantic relationships in healthcare that help online patients make health-related decisions.
- Extracting medical knowledge within narrative posts by using domain knowledge precisely is challenging.
- Health-related information from narrative posts should include topics, domains, semantic relationships, and other implicit knowledge.
- Metrics on evaluating the changes in knowledge support in online health communities are important for implementation.

Background

Increasing use of online health communities (OHCs) offers an opportunity to enhance the abilities of online patients to obtain effective information concerning health-related issues and promote the involvement of communities in personal health management

(Grosberg et al., 2016). OHCs with effective health information in helping patients' decision-making concerning their illness are especially useful for patients with chronic diseases (Willis et al., 2017). The effective health information provides the basis of knowledge support for fulfilling information needs of online patients. The knowledge support in this context is regarded as the integration of knowledge sources in medical and healthcare domains in the communities. However, most OHCs at present merely help patients share their experiences of illness through initial posts written by online patients (Yan et al., 2016). In general, the posts contain a large amount of textual information, including descriptions of medicine, symptoms, and adverse drug events related to health issues. Proper utilization of such information helps predict remarkable health events in life-changing illnesses (Schmidt, 2012). Thus, an enhanced OHC should assist healthcare providers, online health information entrepreneurs, and developers in helping patients and caregivers make intelligent choices (Nath et al., 2016). Establishing and sustaining such a successful OHC require an enabler and strategic community management (Young, 2013). With the effective mechanism of health information management in OHCs, emotional, social, and technical contexts from patient-reported posts can contribute to fulfilling the information needs of online patients (Liu et al., 2014). However, the OHCs at present still experience difficulties in automatically extracting meaningful knowledge in a timely manner to assist patients in making decisions because of the complex semantics in posts (Foster, 2016). Therefore, the methodology to extract effective health information and accordingly assist patients in improving their health management should be determined (Rubrichi & Quaglini, 2012).

Knowledge in medical and healthcare domains provides authoritative references for enhancing computer-aided support in medical information systems and health-related websites. The Unified Medical Language System (UMLS) provided by the National Library of Medicine is the most important compendium of many controlled vocabularies in biomedical sciences. The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), maintained by the International Health Terminology Standards Development Organization, is a large knowledge source that defines substantial biomedical concepts and their relationships in UMLS. At present, SNOMED CT is widely used and recommended as a reference for the use of terminologies in medical informatics (Ronald & Nicolette, 2008). Moreover, SNOMED CT covers a multidisciplinary complex and biomedical chronic conditions, which are necessary to support online patients with chronic conditions (Sampalli et al., 2010). Although efforts have been devoted to developing and refining SNOMED CT in clinical use, further studies should be performed in the design, use, and maintenance of OHCs (Little et al., 2013). The integration of UMLS in the systems of OHCs also helps the communities to support complicated health-related services in the future (Albin et al., 2014). Although some published consumer vocabularies may be more suited to the proper analysis of colloquial posts in OHCs, UMLS sources, including SNOMED CT, are effective biomedical knowledge sources for the extraction of problems, procedures, medications, and clinical results; consequently, UMLS is widely used in healthcare domains (Tomasz et al., 2014; Nie & Shen, 2015). For example, medical knowledge has been introduced in web-based contexts to help knowledge discovery through social media practices (Kallinikos & Tempini, 2015). Therefore, utilization of authoritative knowledge sources in the medical field to enhance decision support is important for medical and healthcare systems.

The utilization of patient-reported data in OHCs has been extensively explored to help patients seek effective health information for their decision making (Dobkin & Boothroyd, 2008). However, the dependence on medical experts in OHCs does not promote the sustained development of a community (Eijk et al., 2013). The manner by which patients seek health information may conversely affect the offline health-related behaviors of patients and should therefore be considered in the context of dynamics between patients and health professionals (Moreland et al., 2014). A drug indication database integrated from 12 knowledge sources or ontologies by terminological normalization to UMLS helps pharmaceutical researchers overcome information overload and accelerate drug discoveries (Sharp, 2017). However, difficulties are encountered in employing UMLS to existing methods. Considering that most questions related to patient care can be answered by existing online clinical knowledge, Morid et al. (2016) developed a Kernel-based Bayesian network classification model based on UMLS concepts and their semantic groups extracted from sentences, thereby outperforming general baseline methods. To understand the unstructured comments of patients about their care, Greaves et al. (2013) applied sentiment analysis techniques that can identify, extract, quantify, and study affective states and subjective information. Ren et al. (2014) used layered dynamic programming to analyze and predict user participations in OHCs. To identify groups at risk of certain diseases and pathologies, Alonso et al. (2016) proposed a straightforward information retrieval system for biomedical fields based on the UMLS Metathesaurus. Detecting different types of social support activities via text mining also helps understand and predict user participations in an OHC. Therefore, leveraging the patient-reported data in OHCs has great potential for improving the decision support of online patients.

Objectives

This study aims to utilize the domain knowledge in medical and healthcare domains in enhancing information extraction from posts and accordingly provide effective health information to online patients. First, knowledge sources from the domain knowledge are examined for knowledge representation in the OHCs. Second, an effective health-related information extraction process in the OHCs is proposed. Then, the metrics on evaluating the degree of knowledge support in the communities are also examined in terms of fulfilling information needs of the patients. Finally, the strength and limitations of the study in practices based on experimental results are discussed.

Methods

Knowledge representation framework

A knowledge representation framework (KRF) established in the OHC is the basis of utilizing existing authoritative knowledge sources, including UMLS and SNOMED CT, in the medical domain. Different knowledge sources have different mechanisms of knowledge representation in specific domains. Some mainly use the definitions of concepts and relationships among the concepts to represent medical knowledge. For example, International Classification of Diseases (ICD) standards that are used to classify diagnostic information have hierarchical knowledge structures. Thus, an independent KRF that can

integrate the most usable knowledge sources to facilitate information use in OHCs is necessary. Figure 1 illustrates the architecture of KRF for our analysis of narrative posts in OHCs.

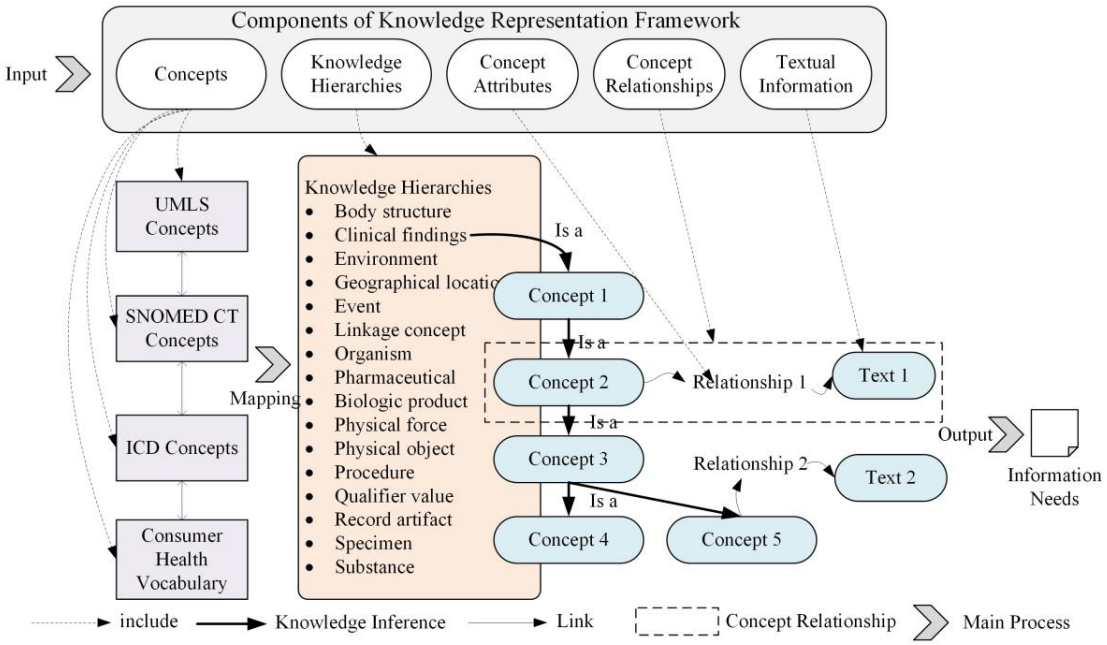

Figure 1. Architecture of the knowledge representation framework

The proposed KRF's components illustrated in Figure 1 used to represent medical domain knowledge include concepts, knowledge hierarchies, concept attributes, concept relationships, and textual information. The components contribute to building an effective knowledge support and organizing different knowledge sources in OHCs. The process of building the KRF in the figure is as follows. At first, the concepts in the figure are organized as a combination of different knowledge sources in the medical domain, including UMLS concepts, SNOMED CT concepts, ICD concepts, and consumer health vocabulary. Here the knowledge sources construct a huge medical and health-related dictionary (Concept 1, Concept 2, …). Then, knowledge hierarchies from the knowledge sources are used to provide a hierarchy of massive concepts in the dictionary. Then, the concept attributes are used to represent the properties of the concepts. On the basis of existing concept relationships in the knowledge sources (Relationship 1, Relationship 2, …), relationships such as the "is-a" relationship among the concepts in the framework are reorganized to provide an integrated knowledge base for further use. Finally, textual information (Text 1, Text 2, …) associated with the concepts is inferred so the framework provides additional information related to the initial input that contains limited information provided by online patients. Using semantic and knowledge relationships in the knowledge sources can facilitate the development of OHCs (Ravoire et al., 2017). In sum, the architecture in the figure provides the basis of knowledge support for analyzing patient-reported data in OHCs by integrating multiple knowledge sources in medical domains.

For bridging the online posts in OHCs and the KRF, a knowledge expansion process is illustrated in Figure 2 to facilitate the use of the framework for analyzing the posts. When the patients in different disease-based support groups browse health-related posts,

generally, many factors have influence in the choice of these online patients in seeking additional knowledge support. The process extracts biomedical texts from the posts of patients in their groups and implements machine learning techniques, including text categorization and text clustering. The system then evaluates the extracted knowledge existing in a patient's post and accordingly predicts the specific knowledge-based services that the patients are interested in. Finally, the online patients can acquire knowledge-based services. Using the KRF, the knowledge within the posts can be expanded with more implicit knowledge from external knowledge sources. The process can be used in further applications of machine learning techniques, such as linguistics and topic modeling (Yang et al., 2016).

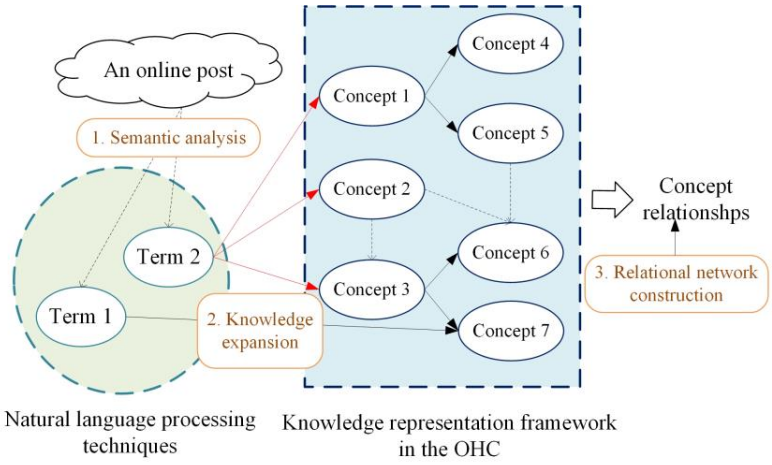

Figure 2. Knowledge expansion process based on the extracted terms from an online post

Information extraction process

On the basis of the KRF in Figure 1 and the knowledge expansion process in Figure 2, an information extraction process to extract effective health-related information from inputs of online patients is proposed. Figure 3 illustrates main steps of the process: determining the topics of the posts, identifying domains of the posts, obtaining relationships among the terms, and implementing knowledge expansion for the text. The details of each step in Figure 3 are as follow:

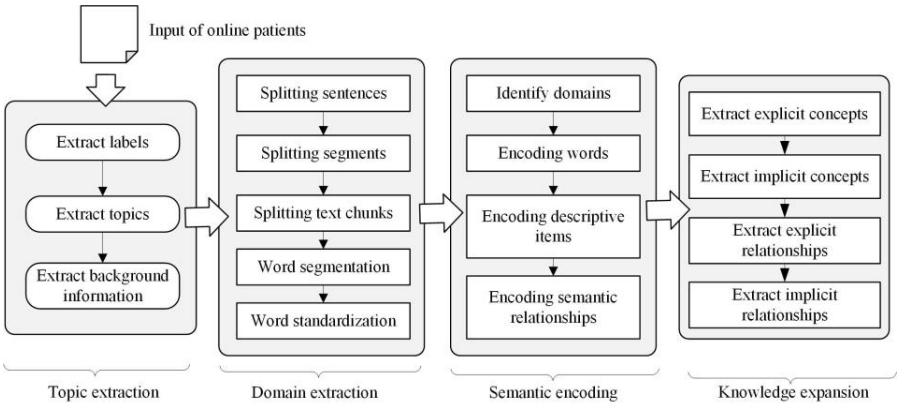

Figure 3. Overview of a health-related information extraction process for posts

The first step is to determine the topics of the inputs of online patients in the OHCs. The complexity of the input of an online patient can be reduced by associating the posts with specific health-related topics. However, the accuracy of determining topics without prior knowledge involved in the topic extraction is often not ideal. Thus, fetching the topics of the post initially is necessary. For example, a post relevant to diabetes can be associated with a diabetes knowledge model summarized by experts in the knowledge sources. In this case, original posts are classified into the predefined topics that can be used in the knowledge model (Souvignet et al., 2016). Here, this step divides the free texts in a post into two types of narrative data: metadata items and free texts. The metadata items in a post, such as explicit labels and background information, are analyzed to obtain basic topics of the post. The free text within the posts is then analyzed to obtain keywords related to the medical domain within the post.

The second step is to identify specific medical domains for each term extracted from an online post. The narrative text of the posts in this step is split into a list of meaningful terms. In general, most health-related posts contain terms with significant medical meanings. The meanings are relevant to different domains, such as the domains related to knowledge of symptoms, medicine, and treatment. Here, the part of speech of the terms is labelled, and then the domains of each term are identified. For example, the term "aspirin" belongs to a medicine domain. Many domains are considered in the method because the post may also contain other information that is useful for the analysis. The different knowledge sources in the KRF contribute to identifying multiple domains of each term for analysis. Then, the terms are encoded by using concept IDs from the KRF. An encoding template is considered a quadruple that includes the code, semantic type, knowledge source, and preferred name according to the UMLS Metathesaurus. This template accurately represents the extracted terms to some extent. Considering the domains, a 2-tuple (domain, code) is used to describe encoded codes for a specific term. For example, given the term "aspirin," we encode the term as 2-tuple, that is, (SNOMED CT, 387458008|aspirin|), which indicates that the term "aspirin" belongs to a domain in SNOMED CT encoded as ID: 387458008. When the encoding process is completed, all of the terms within a post are encoded into different domains.

Then, the relationships among the terms in the post are obtained. Two types of possible items should be illustrated in the text: descriptive and semantic relationship items. The descriptive items are used to represent the basic structure of adjacent terms and their location relationships and are usually associated with syntactic structure, which can be found on the basis of the results of MetaMap (Chiaramello et al., 2016). The semantic relationship items are used to infer the specialization or generalization relationships from the KRF, such as the UMLS (Jin et al., 2015). Therefore, the process first explores descriptive items that describe the patterns of expressions based on the term-encoded texts in the encoding process. After this operation is accomplished, many descriptive items can be discovered and re-encoded according to their formation in the text. Some semantic tags are used to encode the text corresponding to semantic relationships found in the framework. The encoded texts are then mapped to the terminology sets in the framework. In sum, tagging such existing relationships into the item-encoded texts helps represent semantic meanings of the free text. Thereafter, implicit semantic relationships among the terms within the text can be obtained.

Finally, the knowledge expansion process described in Figure 2 is implemented. On the basis of the result acquired from the previous step, additional semantic relationships

from the framework are easily extracted. On the one hand, the system queries MetaMap to fetch all of the concepts that possess UMLS's Concept ID and convert the UMLS's Concept ID into SNOMED CT's Concept ID. On the other hand, the concepts are associated with the specific concepts in posts according to the knowledge hierarchies in Figure 1 and the attributes and relationships of the concepts extracted directly from the posts. By reorganizing the knowledge relationships of terms, additional implicit domain knowledge from the knowledge sources is discovered, and the system in the OHC provides the most useful and relevant information to a specific patient according to his/her sharing experiences. The final result of this process provides explicit and implicit knowledge in a post to fulfill the information needs of the patients in the OHCs.

Evaluation metrics

On the basis of the expanded knowledge from the input of online patients in the OHC, two proposed metrics aim to measure knowledge changes during the process of fulfilling information needs. These metrics are Knowledge Support Rate (KSR) and Effective Information Rate (EIR), which are important in terms of providing knowledge support for online patients in OHCs. The effectiveness of the two metrics here depends on the number of relevant concepts extracted from the posts. Generally, traditional methods based on massive posts depends on statistics for text mining. However, with the integration of authoritative knowledge sources in medical domains, each meaningful term extracted in the post by following the information extraction process in this study can be regarded as useful information for online patients. In addition, the knowledge in the KSR and EIR is extracted based on the semantic relationships among the terms in medical and healthcare domains; therefore, the relevance among the terms is stronger than that using existing statistical methods in the literature. Thus, the two metrics can be used to estimate the change in knowledge during the process.

Knowledge support rate

The expanded knowledge provides a comprehensive understanding of the input of an online patient. Determining the degree of knowledge support for fulfilling the information needs of online patients concerning health-related issues is important. KSR is proposed as a metric to represent the degree of knowledge support in the OHC facilitated with the proposed method. The metric measures the rate of implicit knowledge in a post associated with other knowledge sources in comparison with the explicit knowledge in the post in each analysis for the post. The definition of KSR is defined as in (1).

$$\mathrm{KSR} = \frac{Implicit\ knowledge - Explict\ knowledge}{Explit\ knowledge} \tag{1}$$

The explicit knowledge in (1) represents the healthcare and medical knowledge within a post. However, the implicit knowledge in (1) represents the amount of knowledge extracted from the KRF. The explicit knowledge in (1) can be calculated according to the number of concepts identified from the text. The implicit knowledge is calculated according to the number of concepts associated with the explicit concepts on the basis of the semantic relationships found in the framework.

Effective information rate

Domain knowledge extracted to fulfill the information needs of online patients may conversely affect the behaviors of online patients who are looking for health-related information from peers with trusted experiences (Pavlinek et al., 2017). For example, after a patient finishes reading a post online, and assuming that this patient chooses a keyword extracted from the post by using the proposed method, the OHC navigates the patient to a specific page that contains the chosen keywords. To evaluate the influence of this generated knowledge on the choices of patients and to determine whether the chosen keywords satisfy the actual health-related conditions of patients in their profiles, we propose the EIR as a metric to measure the effectiveness of information provided to the patient concerning health-related issues. EIR is defined as follows:

$$\mathrm{EIR} = \frac{KSR(c) - KSR(p)}{KSR(p)}, \tag{2}$$

where KSR(c) represents the KSR in the current context while KSR(p) represents the KSR in the previous context. Thus, the EIR reflects effective information of the transition from the previous page to the current page in the context of the OHC. The metric is used for the OHC to evaluate the knowledge changes according to the selection of the patient with information needs.

In sum, KSR and EIR are used as metrics to evaluate the information and knowledge changes during the process of information seeking of online patients in OHCs.

Results

Datasets

The patient-reported data in OHCs and authoritative knowledge sources are needed in this study to facilitate the proposed method for fulfilling information needs of online patients. A dataset of online posts from an OHC is used to examine the feasibility of the method in fulfilling information needs of online patients. The dataset consisting of 372,343 patient-reported posts is obtained from an OHC (https://www.healthboards.com/). Each post in the dataset has a topic related to drugs that online patients used. Samples of the posts are listed in Table 1. The website is a typical OHC because online patients in the websites exchange their experience of illness and the treatment of medicine through social networks. Therefore, the OHC is chosen as a case study of examining our proposed method. Although other OHCs may have different styles of narrative posts for different purposes, the motivation of our study focused on fulfilling information needs of online patients is still suitable for most OHCs. The dataset contains English free texts of drug-specific posts and discussions. Its terms of use prevent privacy disclosure; thus, public information, such as captions and narrative texts, is collected from the posts for academic research. Aside from the dataset, the UMLS Metathesaurus and SNOMED CT terminology sets are used in establishing the KRF for knowledge support. On the basis of the dataset and the framework, we use KSR and EIR to measure the results in different steps of the method.

Table 1. Sample of patient-generated online posts in the dataset

Topic of drug	Caption of online post	Narrative text of online post
Casodex	Post Op Watchful Waiting	I agree with that everything should be done to persuade your husband to seek early treatment with a post-op rising PSA. Gleason 9 is a very aggressive cancer and will progress rapidly. So far it seems it has only spread regionally with positive margins and extra-capsular extension.
Aspirin	CAVERNOUS HEMANGIOMA - anyone?	I have a cavernous hemangioma in the lower left occipital region of the brain. I'm 44 and the first symptoms of any problems happened when I was about 16. I started to get migraine-like headaches. My dad suffers from migraine, so this was the thought of our family doctor. To be on the safe side he recommended that I see a neurologist.
Bactrim	UTI problems not being solved	They say that 80% of UTI's are caused by the E.coli bacteria but there are two other bacterias that the D-Mannose is not effective in fighting. So I started to take a natural antibiotic called Oil of Oregano (nasty stuff). It is supposed to be very strong. Then after two days I saw blood in my urine. Has anyone else had success with that?
Reglan	fentanyl patch and nausea	I am so sorry to hear that you are going through this. I am very prone to nausea with many medications, fentanyl included. When I first was on the patch, I used Reglan and Phenergan to help and within 2-3 weeks it was much better. I would call your doctor and see if maybe you can go up more gradually (ie the 12.5 mcg patch + a 50 mcg patch for a while) to see if that helps.

Evaluating knowledge support

The EIR based on the aforementioned dataset is calculated to examine the feasibility of the metric to evaluate the degree of knowledge support to analyze the inputs of online patients. The dataset includes 316 types of drugs mentioned in the posts. To examine the amount of implicit knowledge obtained from the KRF, we experiment the knowledge expansion process for each drug-related topic. A summary of the number of drug-related topics in different ranges of KSR is presented in Table 2. As shown in the table, 280 of the KSRs fall in the [0, 100] range, which account for 88.60% of all KSRs. Other KSRs occupy 12.40%. Thus, we conclude that most of the KSRs lie in the [0, 100] range. Figure 4 illustrates the distribution of the number of drug topics in different sub ranges of [0, 100]. In the figure, about 42.43% of the KSRs are in the [0, 5] range, whereas the others have a relatively low number below 40. Table 2 and Figure 4 present the distribution of KSRs when the KSR is used to measure the knowledge obtained from the framework.

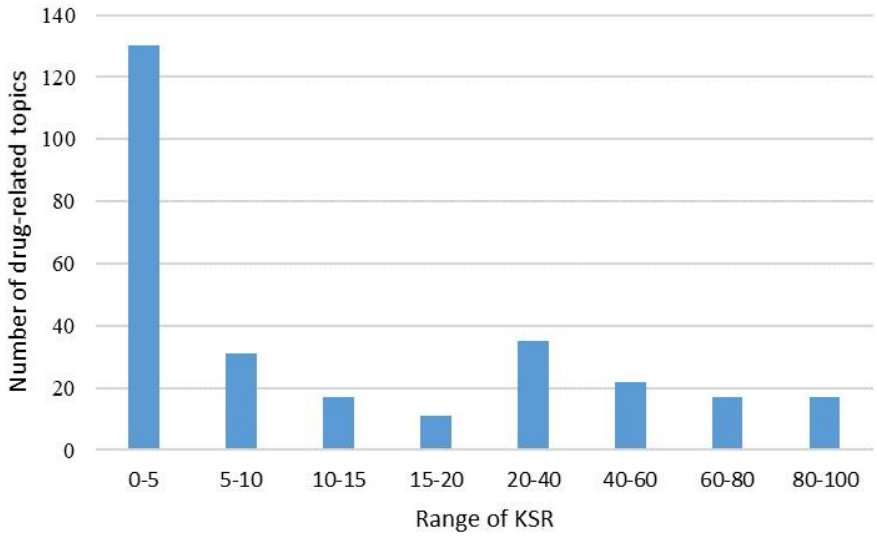

Figure 4. Distribution of the KSRs among the drug-related topics

Table 2. Summary of KSRs for drug-related topics

Range of KSR	Number of drug-related topic
0–100	280
100–200	15
200–300	5
300–400	7
400–500	2
500–1000	3
1000–1500	3
1500–2000	1

Fulfilling information needs

On the basis of the KSRs in the experiment, we calculate the EIR during the changes of drug-related topics. As stated previously, EIR indicates the degree of fulfilling the information needs of online patients in the OHC. For example, regarding Azilect → Ecotrin, a patient who is currently at a page with a topic of Azilect selects another page with a topic of Ecotrin. EIR is used to estimate the information changes in the pages where the patient stays. In the table, if the value of EIR is positive, then the patient is led to a web page that contains more health-related knowledge. If the value is negative, then the knowledge in the current page is less than that in the previous ones. Otherwise, the information between two pages is unchanged.

Table 3 described three examples of the EIR calculation in the process of information seeking. They are "Azilect → Ecotrin," "Apidra → Xanax," and "Valium → Valium." The first column in the table presents the process of changing topics by online patients. The examples in the table represent different cases of information seeking for online patients.

The KSRs of the cases are calculated in advance. The different values of EIR indicate that the changes of online patients' selection in the OHC contribute to adapting the behavior of online patients to help the patient obtain effective health-related information in the OHC. In sum, EIR is an important metric to estimate knowledge-based changes to fulfill the information needs of online patients in the OHC.

Table 3. Calculation of the EIR during information seeking

Process of information seeking	KSR of Current Page	KSR of Previous Page	Value of EIR	Meaning
Azilect → Ecotrin	1741.157	1351.439	0.2884	Effective
Apidra → Xanax	78.2424	88.24113	-0.1133	Not Effective
Valium → Valium	0.25197	0.25197	0	No Changes

Discussion

In this study, we propose a comprehensive method to utilize patient-reported data and authoritative knowledge sources in medical domains in fulfilling the information needs of online patients in the OHCs. The KRF illustrated in Figure 1 for organizing and representing multiple knowledge sources in medical domains in the OHCs is proposed. On the basis of the framework, a health-related information extraction process illustrated in Figure 3 is illustrated. Then, KSR and EIR are proposed as metrics. Experimental results indicate that the method is suitable for the OHCs to analyze patient-generated posts and fulfill the information needs of online patients concerning health issues.

Authoritative knowledge sources in the medical field help facilitate OHCs to support their online patients. The use of authoritative knowledge sources also improves the patients' ability to seek effective health information related to their concerns. The proposed KRF utilizes the UMLS Metathesaurus and SNOMED CT in constructing a large knowledge source as the basis of domain knowledge support in the OHCs. The extracted medical knowledge obtained from the posts where online patients share their experiences with other patients with similar health-related issues is valuable to promote knowledge-based services to the patients promptly. The results show the basic use of the method based on the knowledge sources for fulfilling the information needs of online patients in OHCs. Most machine learning techniques can be easily integrated in the communities as the first step to analyze patient-reported data because such techniques largely depend on the effective extraction of healthcare terms within the posts (Reifegerste et al., 2017).

Knowledge expansion based on explicit knowledge of the input of online patients is a key to promote intelligent knowledge-based services of OHCs. In our experiment, two methods are applied to process the online posts. One method is referred to as the keyword-based matching method that depends on keywords extracted from a dictionary with unspecific domains to provide recommendation information, which is widely used in most OHCs because it is easy to use (Albert et al., 2015; Akay et al., 2014; Osborne et al., 2016). Another method is the proposed method with the framework in Figure 2. The outputs from these two methods are compared. The results indicate that our method is more feasible to

obtain implicit knowledge from the posts than the keyword-based matching method. Thus, it is concluded that the key factor in maximizing the use of the posts in OHCs is fully integrating and utilizing the existing knowledge sources in medical domains rather than focusing on extracting explicit knowledge within the posts.

The information needs of online patients should be fulfilled by existing patient-reported data from the posts and their behaviors. Online patients in an OHC always have the desire to obtain useful health-related information to solve their health-related issues. An intelligent OHC is one that provides an automated decision-making process to promote knowledge-based services to patients. KSR and EIR are the key to achieve the goal. As shown in Table 2 and Figure 4, most KSRs are larger than 1, indicating that the obtained relevant knowledge is much greater than that of the explicit knowledge within the post. Thus, the extended knowledge is able to fulfill more information needs of online patients. Conversely, the enhanced knowledge support affects the behaviors of the patients during the process of seeking health-related information. At present, most OHCs fail to guide the patient with proper and precise health-related information. In sum, the proposed method and the corresponding metrics are important to facilitate intelligent OHCs in the future.

This study has limitations. First, the KRF we used is composed of two types of knowledge sources in the medical field, namely, UMLS and SNOMED CT. In fact, many other knowledge sources in the medical and healthcare domains can be integrated in the framework. Different capabilities of knowledge sources in the framework exert different influences on the accuracy of the extracted knowledge from the posts. However, UMLS and SNOMED CT are the most widely used and are comprehensive enough to represent the knowledge in health-related posts. We use tuples to represent the relationships among the concepts from the framework, which may not be very representative in some cases. Other knowledge sources may also have other media to represent their expert knowledge. Second, the patient-generated contents in the OHCs vary depending on the types of OHCs. In this study, we focus on the analysis of narrative posts written by online patients when they are sharing experiences with others in the OHCs. Other types of patient-generated data, such as posted images and the data from online health tools, also influence information needs of patients. Open access consumer health vocabularies are useful in analyzing the posts that contain massive non-standard medical terms. Therefore, utilization of the vocabularies in the KRF illustrated in Figure 1 may improve the accuracy of our method. However, in this study, we focused on applying domain knowledge in analyzing explicit and implicit knowledge within the posts and did not consider the narrative styles of online posts that contain non-standard terms and slangs in the medical field. Thus, the context in the OHCs to fulfill information needs is complex. The information extraction process in this study is suitable to process the narrative posts and is possibly unable to handle other types of patient-generated data in the OHCs. Third, this comprehensive method does not consider the involvement of machine-leaning techniques and other artificial intelligent techniques to improve the accuracy of knowledge inference and fulfill information needs. However, the proposed method based on domain knowledge is simple enough for the designers of the OHCs to use in practices. In fact, the output of the information extraction method based on the framework provides the basis of the textual feature selection on the posts when we apply machine learning techniques, such as clustering analysis, in the posts. Thus, the proposed method is also the initial and key stage of the future implementation of such machine learning techniques. Fourth, UMLS and SNOMED CT are products that require licenses to

employ in the OHCs. Developing open-source licenses for utilizing the knowledge is important to promote knowledge-based online communities. KSR is a simple metric to measure how much knowledge is obtained. It depends on the number of concepts and relationships among the concepts; sometimes, it cannot handle complicated cases in online environments. We introduce EIR for the OHCs to estimate changes in knowledge in different contexts when the patient is seeking relevant health-related information. Finally, the collected dataset is relevant to drug use and written in natural language; hence, our experimental results may not have a reference value for specific OHCs, such as PatientsLikeMe (https://www.patientslikeme.com/), a website that uses journals and charts to quantify patient data. In sum, the proposed method is feasible to fulfill the information needs of online patients in the OHCs, although many further works should be considered in real life.

Conclusion

This study proposes the KRF to integrate multiple knowledge sources in medical domains for improving the knowledge-based services in OHCs. The framework is proven effective and feasible to utilize the domain knowledge in analyzing online posts. On the basis of the dataset of online posts collected from an OHC, the information extraction process with domain knowledge is also feasible to gather knowledge within the post and effectively obtain implicit knowledge from the framework. KSR and EIR are proposed as metrics to provide an easy and simple way to estimate changes in knowledge during the process of information seeking of online patients. The proposed method applied in the OHC is useful in promoting an intelligent OHC in terms of fulfilling the information needs of online patients in the OHC.

Conflict of interest statement

The authors declare no conflict of interest.

References

Akay, A., Dragomir, A., & Erlandsson, B. E. Network-based modeling and intelligent data mining of social media for improving care. *IEEE Journal of Biomedical and Health Informatics* 2015, **19**, 210-218.

Albert, P., Hartzler, A. L., Jina, H., Mcdonald, D. W., Wanda, P. Automatically detecting failures in natural language processing tools for online community text. Journal of Medical Internet Research. 2015, 17(8), e212.

Albin, A., Ji, X., Borlawsky, T. B., Ye, Z., Lin, S., Payne, P. R., ... & Xiang, Y. Enabling online studies of conceptual relationships between medical terms: developing an efficient web platform. *JMIR Medical Informatics* 2014, **2**, e23.

Alonso, I., & Contreras, D. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach. *Expert Systems with Applications* 2016, **44**, 386-399.

Chiaramello, E., Pinciroli, F., Bonalumi, A., Caroli, A., & Tognola, G. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of Biomedical Informatics* 2016, **63**, 22-32.

Dobkin, P. L., & Boothroyd, L. J. Organizing health services for patients with chronic pain: when there is a will there is a way. *Pain Medicine* 2008, ***9***, 881-889.

Foster, D. 'Keep complaining til someone listens': Exchanges of tacit healthcare knowledge in online illness communities. *Social Science & Medicine* 2016, **166**, 25-32.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research* 2013, **15**, e239.

Grosberg, D., Grinvald, H., Reuveni, H., & Magnezi, R. (2016). Frequent surfing on social health networks is associated with increased knowledge and patient health activation. *Journal of Medical Internet Research* 2013, **18**, e212.

Jin, J., Yan, X., Li, Y., & Li, Y. How users adopt healthcare information: an empirical study of an online Q&A community. *International Journal of Medical Informatics* 2016, **86**, 91-103.

Kallinikos, J., & Tempini, N. Patient data as medical facts: Social media practices as a foundation for medical knowledge creation. *Information Systems Research* 2014, **25**, 817-833.

Lee, D., Cornet, R., Lau, F., & De Keizer, N. A survey of SNOMED CT implementations. *Journal of biomedical informatics* 2013, **46**, 87-96.

Little, M., Wicks, P., Vaughan, T., & Pentland, A. Quantifying short-term dynamics of parkinson's disease using self-reported symptom data from an internet social network. *Journal of Medical Internet Research* 2013, **15**, e20.

Liu, Y., Xu, S., Yoon, H. J., & Tourassi, G. Extracting patient demographics and personal medical information from online health forums. In *AMIA Annual Symposium Proceedings* 2014*,* 1825.

Moreland, J., French, T. L., & Cumming, G. P. The prevalence of online health information seeking among patients in Scotland: a cross-sectional exploratory study. *JMIR Research Protocols* 2015, **4**, e85.

Morid, M. A., Fiszman, M., Raja, K., Jonnalagadda, S. R., & Del Fiol, G. Classification of clinically useful sentences in clinical evidence resources. *Journal of Biomedical Informatics* 2016, **60**, 14-22.

Nath, C., Huh, J., Adupa, A. K., & Jonnalagadda, S. R. Website sharing in online health communities: a descriptive analysis. *Journal of Medical Internet Research* 2016, **18**, e11.

Nie, J. Y., & Shen, W. (2015, October). Flexible Concept Matching for Medical Information Retrieval. In IEEE Conference on *Systems, Man, and Cybernetics (SMC)* 2015, October 9-12, Kowlon, China 1901-1906.

Osborne, J. D., Wyatt, M., Westfall, A. O., Willig, J., Bethard, S., & Gordon, G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association* 2016, **23**, 1077-1084.

Pavlinek, M., & Podgorelec, V. Text classification method based on self-training and LDA topic models. *Expert Systems with Applications* 2017, **80**, 83-93.

Ravoire, S., Lang, M., Perrin, E., Audry, A., Bilbault, P., Chekroun, M., ... & Malbezin, M. Advantages and limitations of online communities of patients for research on health products. *Therapie* 2017, **72**, 135-143.

Reifegerste, D., Wasgien, K., & Hagen, L. M. Online social support for obese adults: Exploring the role of forum activity. *International Journal of Medical Informatics* 2017, **101**, 1-8.

Ren, K., Lai, A. M. , Mukhopadhyay, A., Machiraju, R., Huang, K. Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. Medical Genomics, 2014, 7(Suppl 1): S11.

Ronald, C., Nicolette, D. K. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making, **2008**, 8(Suppl 1), S2.

Rubrichi, S., & Quaglini, S. Summary of Product Characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics* 2012, **45**, 231-239.

Sampalli, T., Shepherd, M., Duffy, J., & Fox, R. An evaluation of SNOMED CT® in the domain of complex chronic conditions. *International Journal of Integrated Care* 2010, **10**, e038.

Schmidt, C. W. Trending now: using social media to predict and track disease outbreaks. *Environmental Health Perspectives* 2012, **120**, a30.

Sharp, M. E. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics* 2017, **8**, 2.

Souvignet, J., Declerck, G., Asfari, H., Jaulent, M. C., & Bousquet, C. OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. *Journal of Biomedical Informatics* 2016, **63**, 100-107.

Tomasz, A., Naoki, S., Mary, S. Next generation phenotyping using the unified medical language system. JMIR Medical Informatics,2014, 2, e5.

van der Eijk, M., Faber, M. J., Aarts, J. W., Kremer, J. A., Munneke, M., & Bloem, B. R. Using online health communities to deliver patient-centered care to people with chronic conditions. *Journal of Medical Internet Research* 2013, **15**, e115.

Willis, E., & Royne, M. B. Online health communities and chronic disease self-management. *Health Communication* 2017, **32**, 269-278.

Yan, Z., Wang, T., Chen, Y., & Zhang, H. Knowledge sharing in online health communities: A social exchange theory perspective. *Information & Management* 2016, **53**, 643-653.

Yang, F. C., Lee, A. J., & Kuo, S. C. Mining health social media with sentiment analysis. *Journal of Medical Systems* 2016, **40**, 236.

Young, C. Community management that works: how to build and sustain a thriving online health community. *Journal of Medical Internet Research* 2013, **15**, e119.