Artificial Intelligence in Medicine 81 (2017) 63–77

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

User recommendation in healthcare social media by assessing user similarity in heterogeneous network

Ling Jiang, Christopher C. Yang*

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA, 19104, United States

ARTICLE INFO

Article history:
Received 2 March 2017
Accepted 3 March 2017

Keywords:
Heterogeneous network mining
Similarity analysis
Healthcare informatics
Social media analytics
Recommendation systems

ABSTRACT

Objective: The rapid growth of online health social websites has captured a vast amount of healthcare information and made the information easy to access for health consumers. E-patients often use these social websites for informational and emotional support. However, health consumers could be easily overwhelmed by the overloaded information. Healthcare information searching can be very difficult for consumers, not to mention most of them are not skilled information searcher. In this work, we investigate the approaches for measuring user similarity in online health social websites. By recommending similar users to consumers, we can help them to seek informational and emotional support in a more efficient way.
Methods: We propose to represent the healthcare social media data as a heterogeneous healthcare information network and introduce the local and global structural approaches for measuring user similarity in a heterogeneous network. We compare the proposed structural approaches with the content-based approach.
Results: Experiments were conducted on a dataset collected from a popular online health social website, and the results showed that content-based approach performed better for inactive users, while structural approaches performed better for active users. Moreover, global structural approach outperformed local structural approach for all user groups. In addition, we conducted experiments on local and global structural approaches using different weight schemas for the edges in the network. Leverage performed the best for both local and global approaches. Finally, we integrated different approaches and demonstrated that hybrid method yielded better performance than the individual approach.
Conclusion: The results indicate that content-based methods can effectively capture the similarity of inactive users who usually have focused interests, while structural methods can achieve better performance when rich structural information is available. Local structural approach only considers direct connections between nodes in the network, while global structural approach takes the indirect connections into account. Therefore, the global similarity approach can deal with sparse networks and capture the implicit similarity between two users. Different approaches may capture different aspects of the similarity relationship between two users. When we combine different methods together, we could achieve a better performance than using each individual method.

1. Introduction

Online health social websites have grown substantially over the past decades. Health consumers participated in these websites to share healthcare information with peers. Through social networking with other consumers, they also provide social support for each other. Social support has been found to be important in helping patients to cope with stressful health conditions. It was demonstrated that social support could contribute to health outcomes by enhancing patient adherence to medical treatment [1]. Informational and emotional support are the most prevalent types of social support found in online health social websites [2].

Informational support help consumers to reduce uncertainty by offering facts or knowledge, including advice, information referral, insight from personal experiences, or opinions [2,3]. It was found that a majority of online posts in healthcare social websites were intended to seek informational support, such as asking for advice on drugs or treatments [2]. The information was considered as being particularly helpful to consumers newly joined

* Corresponding author.
E-mail addresses: ling.jiang@drexel.edu (L. Jiang), chris.yang@drexel.edu (C.C. Yang).

http://dx.doi.org/10.1016/j.artmed.2017.03.002
0933-3657/

the online communities [4]. They can learn highly technical and factual information about various medical conditions. The online social support was fulfilling a learning as well as support function, especially for consumers concerned with chronic or mental illness [4,5]. The rapid development of online healthcare social websites has supported consumers' healthcare information searching by significantly reducing barriers to healthcare information. Consumers can see what their peers are experiencing, and discuss healthcare topics with people who have similar concerns. However, this vast amount of healthcare information comes with challenges. Consumers could easily be overwhelmed by the ocean of information and realize that it is difficult to find the most relevant piece. In an online health social website with hundreds of thousands of threads on all kinds of health-related topics, searching can be very difficult for consumers, not to mention most of them are not skilled information searcher. Consumers use very different language from professional terms to express the healthcare issues [6], and this language problem directly leads to poor query formation [7]. In addition, consumers usually cannot fully understand their health conditions due to the lack of medical knowledge, which would further hinder effective information searching [7]. Under such circumstance, asking for informational support from similar consumers could be much more efficient in seeking health information than relying on browsing or using the website search function. Consequently, a good recommendation system that can match users to users for their information needs is desired.

Emotional support is another important social support consumers could benefit from an online health social website. Many patients describe their situations as "understandable only if you have gone through a similar situation." This understanding is part of empathy, which naturally stems from going through a similar situation [8]. Therefore, emotional support could be especially effective when it comes from similar consumers. And prior studies have shown that emotional support could help patients to maintain positive attitude towards health issues and to achieve better health outcomes [9,10]. So, social networking with similar consumers not only provides a shortcut to relevant healthcare information, but also helps consumers to find emotional resonance.

However, online health social websites are different from traditional social media sites. In traditional social media sites such as Facebook and Twitter, people keep in contact with a list of friends, and the networks are often built based on one's existing social ties [9]. People can follow their friends and keep track of news in their friends' daily lives. Social networking activities in such social media sites are based on the relationship between the participants but not necessarily based on the desire of in-depth discussion on healthcare specific issues. Nevertheless, in an online health social websites, social networking activities are mainly based on common concerns rather than long-term relationships. Consumers would participate in the same discussion thread or send messages to others because of similar interests, but they seldom maintain a long-term relationship with each other. The social ties are much weaker in online healthcare social websites [11]. In this case, consumers may not be able to quickly find the most similar users with themselves when they need informational or emotional support. Therefore, recommendation of similar users to consumers in online health social websites is desirable.

In this work, we propose to consider an online healthcare social website as a heterogeneous healthcare information network, and develop the content-based approach and structural approach for recommending similar users in such a network. In this paper, we used data collected from a popular online health social website MedHelp.[1] Millions of Consumers post discussion threads on a variety of healthcare topics in MedHelp every day [12], and we can collect the online threads for our study. In most existing studies on network science, information networks are usually assumed to be homogeneous, where nodes are objects of the same entity type and links are relationships from the same relation type. However, most real-world networks are heterogeneous, where nodes and relations are of different types [13]. Healthcare information network is one typical heterogeneous network. Different types of relations convey different semantic meanings, and treating all the nodes or links as of the same type may miss important semantic information [13].

The rest of the paper is organized as follows. In Section 2, we review related work on network construction, including named entity recognition and relation extraction, and then we discuss existing studies on similarity analysis. In Section 3, we introduce methods for constructing a heterogeneous healthcare information network from an online healthcare social website. In Section 4, we propose three different methods based on content-based approach and structural approach for recommending similar users in online health social websites. Section 5 presents the details on experiment, including data collection, ground truth, results and discussion. Finally, we conclude the work in Section 6.

2. Literature review

2.1. Constructing heterogeneous healthcare information network

Given a set of healthcare social media data, the question is how to construct an information network to best represent the data. In order to do that, we need to identify entities and relationship between entities from the data. In this section, we review related work on named entity recognition and relation extraction from healthcare social media data.

Many studies have been reported on named entity recognition from healthcare social media data in the past few years. Some researchers used lexicon to identify entities such as Adverse Drug Reactions (ADRs) and drugs. In [14], Leaman et al. created a lexicon by combining terms and concepts from four resources and used it to identify adverse reactions in user comments collected from DailyStrength. Patki et al. [15] expanded the lexicon used by Leaman et al. [14] by adding addition terms from two more sources: SIDER II [16] and Consumer Health Vocabulary (CHV) [17]. Benton et al. [18] scraped websites and databases that contain lists of dietary supplements, pharmaceuticals, and adverse events for terms to create a controlled vocabulary, and then augmented the vocabulary using the CHV.

Besides lexicon, language patterns were also utilized to automatically extract mentions of ADRs from user reviews about drugs in social network websites [19]. Pimpalkhute et al. proposed an approach to generate most probable misspelled drug name variants for querying social media postings. The underlying intuition is that, faced with the task of writing an unfamiliar, complex word (the drug name), users will tend to revert to phonetic spelling [20].

Some other researchers used machine learning techniques to extract ADRs from social media data. Using Twitter as data source, Ginn et al. [21] created a manually annotated corpus of 10,882 tweets that were used to train automated tools to extract ADRs from tweets. Nikfarjam et al. [22] extracted user posts from DailyStrength and Twitter, and introduced ADRMine, a machine learning sequence tagger for concept extraction from social media. ADRMine utilizes a variety of features such as context features, ADR lexicon, part-of-speech tags, negations, and a novel feature

[1] http://www.medhelp.org/.

for modeling words' semantic similarities. The experiment results showed that ADRMine outperforms several strong baseline systems in the ADR extraction task.

In terms of relation extraction, there is a large body of literature on drug-ADR relation extraction from social media data. Benton et al. [18] used co-occurrence statistics identified drug event pairs from a corpus of posts to breast cancer message boards. Bian et al. [23] used Natural Language Processing (NLP) techniques and a support vector machine (SVM) classifier trained on a manually labeled corpus of tweets to find drug users and potential adverse events. Yang et al. [24,25] employed association rule mining to identify relationship between drugs and ADRs in online posts, and proved that healthcare social media data is promising for ADR detection. Liu and Chen [26] developed a shortest dependency path kernel function and trained a SVM to learn patterns to detect drug and medical events in patient forum posts. They proposed generating syntactic and semantic features for relation instances based on the shortest dependency path from medical events to treatment entities.

2.2. Similarity analysis

Many studies have been reported for measuring user similarity using different methods for different purposes. Generally, methods for similarity analysis fall into two categories: content-based approaches and structural analysis approaches.

Content-based similarity has its roots in Information Retrieval (IR). Extensively studies of similarity of documents have been done in IR community [27]. Many recommendation systems utilize the similarity between users to predict users' ratings of items. A recommendation system that is based on user-based collaborative filtering algorithm [28] usually first finds similar uses for a target user, then uses the ratings of the similar users to estimate the target user's interests for certain items. Therefore, the measurement of similarity between users plays a fundamental role in these systems. A common method is calculating similarity based on users' profiles, which could be item ratings, browsing logs, clicks, queries and so on. Cosine similarity and Pearson's correlation are popular similarity measures in collaborative filtering [28–30]. Xiao et al. [31] used users' access logs to web pages to measure the similarity of users' interests. In order to measure the similarity more precisely, they also took into account the actual time when users were viewing the web pages. Pennacchiotti and Gurumurthy [32] used a topic model to represent users as mixtures of topics based on their Tweets. After that, they recommend to a target user those users with highly similar topic distributions.

Similarity based on structural analysis approaches are exploited frequently in information network studies. The problem can be viewed as calculating a measure of proximity or similarity between two nodes in a network. Some approaches adopted local similarity measures, based on the idea that two nodes are likely to be similar if they share a large number of neighbors in the network. The Friend-of-Friend algorithm, which is adopted by many popular online social websites such as Facebook.com, is based on the intuition that 'if many of my friends consider Alice a friend, perhaps Alice could be my friend too'[33]. The most basic implementation of this idea is to compute the number of common neighbors [34]. Jaccard's coefficient and Adamic/Adar index are alternative similar measures based on common neighbors [35]. SimRank algorithm proposed by Jeh and Widom computes a global similarity of the structural context in which objects occur [36]. The underlying intuition is "two objects are similar if they are related to similar objects". Due to the high computational complexity of SimRank, some studies were conducted on optimizing the algorithm [37,38]. In [39], Pan et al. used a Random Walk with Restart (RWR) algorithm to compute the affinity of one node to another. A random walker that starts from a node faces two choices at each step: either choosing randomly among the available edges to move, or jumping back to the starting node with a given probability. Symeonidis et al. introduced a transitive node similarity measure for friends recommendation in social networks [40]. Specifically, two persons that are connected with a path have a higher probability to know each other, proportionally to: (1) the length of the path they are connected with, and (2) the degree of similarity between the neighbor nodes that form that pathway. The structural features of a node can be also represented by its ego-centered network and the similarity problem can be also viewed as calculating the similarity of two ego-centered networks. Xiang et al. proposed a metric of similarity between two subgraphs in a network. The idea is that two subgraphs having more connections or being simultaneously closely connected to other subgraphs are likely to have higher proximity to each other [41]. Hsu et al. used the structural features of each user to recommend users with common interests in a social network [42]. Most existing structural similarity measures are explored in homogeneous networks. However, most information networks in real world are heterogeneous rather than homogeneous.

Recently, some researchers start to explore methods for measuring similarity of nodes in heterogeneous networks. Zhao et al. [43] enriched SimRank algorithm by encoding both in- and out-link relationships and proposed P-Rank to compute the structural similarity. Unlike SimRank that only considers in-linked neighbors, P-Rank also calculates the similarity of out-linked neighbors and used a parameter λ to balance the relative weight of in- and out-link directions. They applied P-Rank to a heterogeneous information network and the results demonstrated P-Rank as an effective similarity measurement. Sun et al. introduced the concept meta-path based similarity in [44]. A meta-path was defined as a path consisting of a sequence of relations defined between object types. Different meta-paths carry different semantic meanings in a heterogeneous information network. For example, "author-paper-author" denotes co-author relationship, while "author-paper-author-paper-author" indicates the two authors share a co-author. Taking into account the different semantic meanings associated with different types of links, it does not make sense anymore to mix all links together for similarity measurements. Under the meta-path framework, they proposed a novel similarity measure called PathSim, which turns out to be more meaningful in many scenarios compared with random-walk based similarity measures. Base on the meta-path concept, Yu et al. proposed a user guided similarity search in heterogeneous information network [45]. Given a user query with examples as guidance, one can understand the hidden semantic meanings of the query. Then training dataset with positive and negative entity pairs was used to learn a ranking model for finding the similar entities for the query.

In the current literature, only a few researchers have worked on developing techniques for measuring similarity of objects in heterogeneous network, and very limited number of them have studied heterogeneous healthcare information network. In this work, we proposed approaches for measuring similarity between health consumers in a heterogeneous healthcare information network.

3. Constructing heterogeneous healthcare information network from consumer-contributed content

In this section, we explore the methods for extracting entities and relations from given healthcare social media data. Here we first give a definition of information network.

An *Information Network* is a directed or undirected graph $G = (\nu, E; T, R)$ with an entity type mapping: $\varphi : \nu \rightarrow T$ and a link type mapping: $\phi : E \rightarrow \mathcal{R}$. Vertex $\upsilon \in \nu$ is an entity, while edge $e = \nu, u \in$

E represents a relationship between v and u, where $\upsilon, u \in v$. Type $t_i \in T = \{t_1, t_2, \ldots, t_n\}$ is an entity type, and $\varphi(\upsilon) \in T$. All vertices $v = \{V_1 \cup V_2 \ldots \cup V_n\}$ can be partitioned into n mutually exclusive subsets. Relation $r_j \in R = \{r_1, r_2, \ldots, r_m\}$ is a type of relationship, and $\phi(e) \in R$. All edges $E = \{E_1 \cup E_2 \ldots \cup E_m\}$ can be partitioned into m mutually exclusive subsets. In a weighted network, $w(e)$ denotes the weights of $e = v, u \in E$.

An information network can be directed or undirected, homogeneous or heterogeneous. Fig. 1 shows examples of three different types of information networks. Most current studies consider information networks as homogeneous, that is, $n = 1$ for $T = \{t_1, t_2, \ldots, t_n\}$ and $m = 1$ for $R = \{r_1, r_2, \ldots, r_m\}$. As illustrated in Fig. 1(a), a homogeneous information network contains only one type of node and relationship.

However, most information networks are heterogeneous in real world, where the number of entity types $n > 1$ or the number of relationship types $m > 1$. An online health social community can be abstracted as a heterogeneous network. The types of nodes may include users, and named entities appearing in threads such as drugs, diseases, and ADRs (Adverse Drug Reactions). The edges could consist of drug-treat-disease, drug-cause-ADR, user-take–drug, user-have-disease, and so on (Fig. 1(b)). In this case, the network is a directed heterogeneous information network. A directed network can be extracted when the relationship between a pair of nodes can be determined. For example, in a bibliographic network, the relationships among different types of entities, such as *paper*, *author*, and *topic*, can be explicitly determined from a relational bibliography database [46,47]. However, it is not a trivial task to effectively determine the relationships between different entities in healthcare social media data. We may use sophisticated natural language processing (NLP) techniques or thorough human annotation. But it is reported that using NLP tools to analyze social media data is quite challenging [48]. And human annotation would be very time consuming.

In this work, we represented the healthcare social media data as an undirected heterogeneous healthcare network. As shown in Fig. 1(c), the network contains four types of nodes, including users, drugs, diseases, and ADRs. And the entities are connected by their co-occurrence relationships in the data. That means we only consider one type of relationship here. We can expand the network by adding more types of nodes and identifying explicit relationships between entities in the future.

3.1. Entity recognition

As the constructed network will be used in following similarity analysis, the quality of the network is critical. Lexicon-based method uses exact match to find entities from the data, so it usually can achieve high precision. Therefore, we build dictionaries and used them for entity recognition.

Recognizing consumers, namely healthcare social website users, is straightforward. We extracted user names from corresponding field in the data.

For drugs, we used DrugBank[2] to build the dictionary. DrugBank is a comprehensive online database containing extensive biochemical and pharmacological information about drugs [49]. The database contains 7685 drug entries including 1549 FDA-approved small molecule drugs, 155 FDA-approved biotech (protein/peptide) drugs, 89 nutraceuticals and over 6000 experimental drugs [50]. This repository provides a comprehensive list of drugs for our study. Fig. 2 shows a record example of Abacavir, which is a generic name for the drug, followed by a list of brand names and brand mixture names. For each drug, we collect its generic name, brand names and bran mixture names.

For disease, we utilized UMLS[3] to build the dictionary. The UMLS Metathesaurus contain health and biomedical concepts from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT [51]. UMLS provides a semantic network to categorize all concepts according to their semantic types. All concepts for diseases are organized under the semantic type "Disease or Syndrome". For each concept, UMLS includes a list of string names, which are lexical variants for the concept. We extracted all concepts with related string names from UMLS Metathesaurus to build a dictionary for diseases.

For ADRs, we extracted dictionary from SIDER (Side Effect Resource),[4] which contains information of recorded adverse drug reactions of marketed medicines [52]. The authors asserted that no such resource exists in machine-readable form so far to their knowledge [53]. However, the situation for ADRs is more complicated than drugs and diseases. Although some consumers could use different expressions for drugs or diseases, most consumers reach a consensus of names for drugs or diseases. For a drug, consumers often use the brand name of the drug. A few of them could also use the generic name. Likewise, people can also reach an agreement on disease names, except for some lexical variants, which have already been addressed in UMLS. Nevertheless, when it comes to ADRs, it becomes more complicated. Consumers often use a variety of expressions for ADRs. ADR is defined as an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product [54]. Strictly speaking, ADR is not a disease. Instead, it is a reaction which could be a disease, a symptom, or even just a sign. Therefore, it can be described in various different ways. For instance, if a consumer experienced hair loss after taking a drug, he/she might describe it as "hair loss", "hair falling", or "baldness". There are numerous ways of describing this ADR. Moreover, because of the language gap between consumers and health professionals, there is a mismatch between expressions used by consumers and professional vocabularies. Taking "hair loss" as example, the professional vocabulary for "hair loss" is "alopecia" in SIDER, which it is too professional for most consumers that they may not even know it. They are more likely to use "hair loss" instead of "alopecia". Therefore, it is necessary to study Consumer Health Vocabulary (CHV) to explore methods for addressing this problem.

The consumer vocabulary problem has long been recognized by researchers, and lots of efforts have been taken in developing CHV to bridge the language gap. One remarkable achievement is the development of the open access and collaborative (OAC) CHV, which is defined as "a collection of forms used in health-oriented communication for a particular task or need (e.g., information retrieval) by a substantial percentage of consumers from a specific discourse group and the relationship of the forms to professional concepts" [55,56]. It was developed by identifying Consumer-Friendly Display (CFD) names, that is, expressions describing medical concepts likely to be recognized by most consumers, and then map CFD names to professional vocabulary. Fig. 3 illustrates the record of "alopecia" in CHV Wiki.[5] As we can see, the third column shows the terminology, which is "alopecia" in this case, while the second column gives a CHV preferred name, which is the term that are recognized by most consumers. The first column contains all alternative terms conveying the same meaning with the professional term. Compared with professional terms, the alternative terms are more likely to be used by consumers. Therefore, these terms can be used for expanding the professional terms in

[2] http://www.drugbank.ca/.

[3] http://www.nlm.nih.gov/research/umls/.

[4] http://sideeffects.embl.de/.

[5] http://consumerhealthvocab.chpc.utah.edu/CHVwiki/.

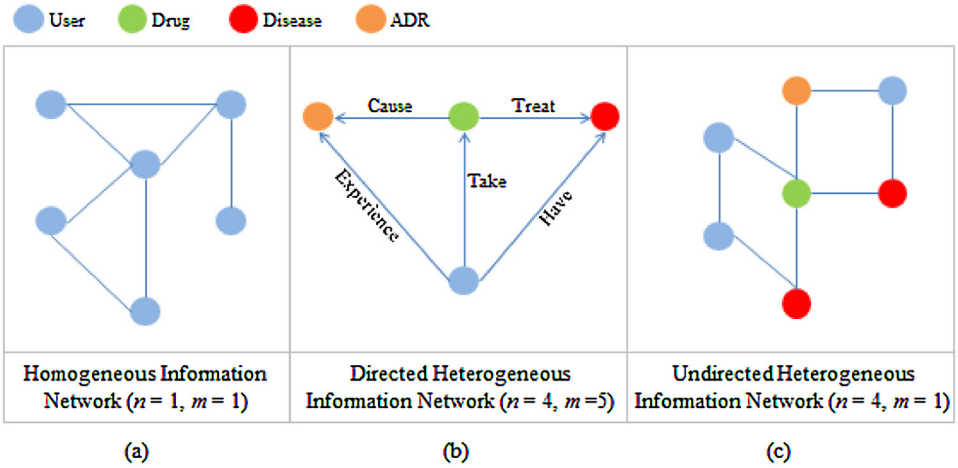


Fig. 1. Examples of Homogeneous Information Network (a), Directed (b) and Undirected (c) Heterogeneous Information Networks.

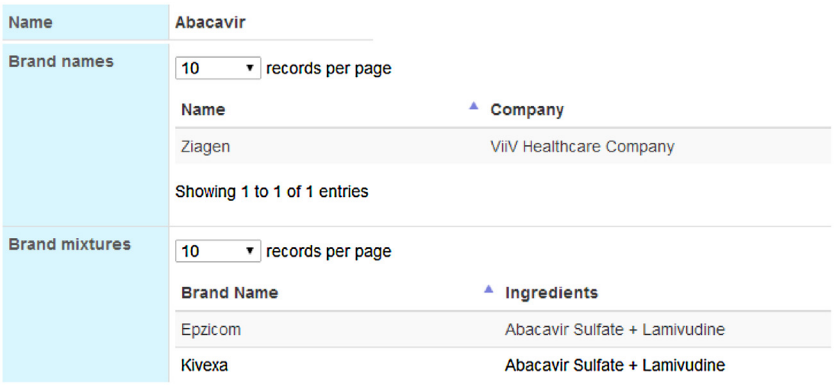


Fig. 2. Example Record of Drug "Abacavir" in DrugBank with Generic Name and Brand Names.

SIDER. In this paper we only used CHV to expand ADRs dictionaries. Even though consumers could also use different expressions for drugs or diseases, the problem is much less severer according to our observation. Most health consumers can easily refer to the indications on the prescribed to describe drugs and diseases; however, they may vary in the expression of their experience or reactions to drugs depending on their demographic backgrounds. The dictionaries based on DrugBank and UMLS can extract most valid drug and disease names. But we could consider using CHV on drugs and disease in our future work.

3.2. Relation extraction

Multiple entities are usually mentioned in a message posted by a health consumer and relations can be drawn between these entities. Fig. 4 illustrates an example of a message from MedHelp. A MedHelp user started a thread, asking about other consumers' experiences with some antidepressants. There are three entities that we can identify from this message: drug – Celexa, drug – Effexor, and disease – depression. The context of the message is that a user has depression and he/she took Celexa before, but now is on Effexor.

Co-occurrence analysis is the most direct way for relation extraction. The basic idea is that if two entities co-occur with each other, it implies an underlying relationship between them. As shown in Fig. 4, since the three entities were mentioned in the same message, there should be relationships between them according to co-occurrence analysis. Moreover, the user who posted this thread is also related to all entities in the message. And the two users in the same thread are connected to each other, because they co-occurred in a thread. Notice that the relations extracted is non-directed. Based on this method, we can construct a small heterogeneous health information network from the example thread as shown in Fig. 5.

4. Determine the similarity of users in healthcare heterogeneous information network

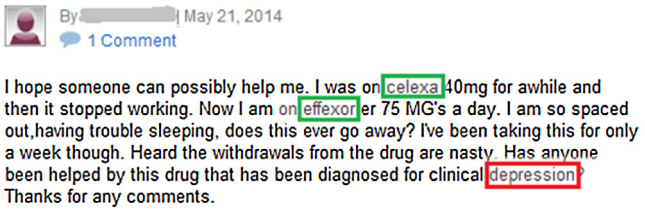
In this section, we present two types of approaches for measuring user similarity in the heterogeneous healthcare information network, namely content-based approach and structural approach. We talk about ContentSim, which is a content-based approach based on document similarity using TF-IDF, in Section 4.1. After that, we propose the local similarly method (ProfileNet) in Section 4.2 and the global similarity method (HealthRank) in Section 4.3.

4.1. Content-based similarity (ContentSim)

In our study, we are concerned with similar consumers who are interested in common health-related topics. In an online health social website, users participate in discussions on topics of interest. The messages authored by a user can best represent his/her interests. Thus the problem of user similarity can be transformed into the problem of content (text) similarity. Cosine similarity is a popular measurement for document similarity. A given document can be represented by a term vector that contains all the terms appear in the document. Cosine similarity measures the similarity of two documents by calculating the cosine of the angle between their

Term	CHV Preferred Name	UMLS Preferred Name
alopecia	hair loss	alopecia
alopecia disorders	hair loss	alopecia
alopecias	hair loss	alopecia
baldness	hair loss	alopecia
disorders hair loss	hair loss	alopecia
falling hair	hair loss	alopecia
falling hairs	hair loss	alopecia
falls hair	hair loss	alopecia
hair fall	hair loss	alopecia
hair falling	hair loss	alopecia
hair loss	hair loss	alopecia

Fig. 3. Example Record of "Alopecia" in CHV (first column – terms used by consumers; second column – CHV preferred term; third column – professional term).


By [redacted] | May 21, 2014
1 Comment

I hope someone can possibly help me. I was on celexa 40mg for awhile and then it stopped working. Now I am on effexor er 75 MG's a day. I am so spaced out,having trouble sleeping, does this ever go away? I've been taking this for only a week though. Heard the withdrawals from the drug are nasty. Has anyone been helped by this drug that has been diagnosed for clinical depression? Thanks for any comments.

Fig. 4. Example of a Message from MedHelp.

term vectors. In this study, for each user, the messages authored by this user can be characterized by a term vector, and the value of each dimension in this vector corresponds to the frequency of each term in the messages. By calculating the cosine of two term vectors, we can measure how similar two set of messages are. Here the term vectors are simple word vectors that contain all words in the messages. The problem is formulated as follow.

Given a set of users $U = \{u_1, u_2, \ldots, u_n\}$ and a collection of messages$M = \{m_1, m_2, \ldots, m_n\}$, user u_i is the author of the messages m_i. Each user u_i can be represented by a term vector $\vec{t_i} = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$, which are all terms in m_i. t_{i1} is the TF-IDF value of the term in M. In order to measure the similarity of u_i and u_j, we adopt the cosine similarity to calculate the similarity between vector $\vec{t_i}$ and $\vec{t_j}$:

$$S(u_i, u_j) = \frac{\vec{t_i} \cdot \vec{t_j}}{\|\vec{t_i}\|\|\vec{t_j}\|} \quad (1)$$

Nevertheless, content similarity has some shortcomings. Since content-based approach measures the similarity using textual content, the quality of the content is critical. However, consumer-contributed content from online social websites are mainly composed of informal narrative content, and the poor quality of the content may impact the performance. Another challenge is that active users' interests cannot be easily represented by simple term vectors. Unlike traditional documents that are usually focused on a specific topic, an active user's messages could cover diverse topics. In that case, simple term vectors would not be able to best represent the user's complex interests. Therefore, we propose another approach based on structural information.

4.2. Local similarity (ProfileNet)

In this section, we propose a local similarity approach called ProfileNet to measure similarity between consumers in a heterogeneous healthcare information network. ProfileNet is proposed based on the concept ego-centered network. Ego-centered network is a specific type of information network. Given the definition of information network in the previous section, we define ego-centered network as follows:

An ego-centered network is an information network that consists of an ego node u and the nodes that node u is connected to in distance $d(d = 1, 2, \cdots l)$. An ego-centered network could also contain $n(n > 1)$ types of entities.

For example, for a user in a heterogeneous healthcare information network, his/her ego-centered network in distance $d = 1$ consists of the user node and all the nodes that the user node is directly connected to. And his/her ego-centered network in distance $d = 2$ contains all the nodes in distance 1 plus the nodes that are directly connected to the nodes in distance 1. In a heterogeneous healthcare information network, we can extract the ego-centered network for each user, and a user's profile can be represented by the entities that the user is associated with in his/her ego-centered network. If two consumers talked about many same entities in threads, it would be likely that they have very similar profiles.

Given the definition of the ego-centered network, we propose to measure the similarity of two users as follows:

$$t_i \in \{t_1, t_2, \ldots, t_n\} \quad (2)$$

where $t_i \in \{t_1, t_2, \ldots, t_n\}$ denotes different types of entities in the network, and α_{t_i} is the weight assigned to type t_i.

The Profile $P^d_{u_1}$ of a user node u_1 is defined as a vector of weights between u_1 and its neighbor nodes within distance d:

$$P^d_{u_1} = \left\{\vec{w_{t_1}}, \vec{w_{t_2}}, \ldots, \vec{w_{t_n}}\right\} \quad (3)$$

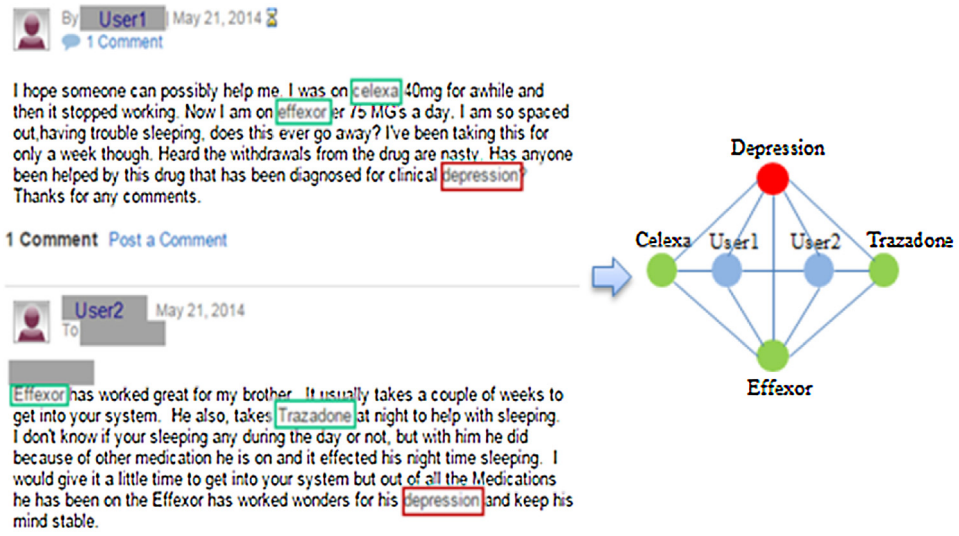


Fig. 5. Example of Constructing Heterogeneous Network from a Thread from MedHelp.

and

$$P_{u_1}^{d}\left(\vec{w}_{t_i}\right) = \left(w\left(e_{u_1v_1}\right), w\left(e_{u_1v_2}\right), \ldots, w\left(e_{u_1v_k}\right)\right) \tag{4}$$

where, $v_j \in \{v_1, v_2, \ldots, v_k\}$ are all the nodes of type t_i in the union of ego-centered networks of node u_1 and u_2, which means the type of v_j $\varphi(v) = t_i$. $w\left(e_{u_1v_j}\right)$ is the weight between node u_1 and node v_j, and it is calculate by the following equation:

$$w\left(e_{u_1v_j}\right) = w\left(e_{u_1x_1}\right) \times w\left(e_{x_1x_2}\right) \times \cdots \times w\left(e_{x_nv_j}\right) \tag{5}$$

where $\{x_1, x_2, \ldots, x_n\}$ are the nodes in the shortest path between node u_1 and node v_j. If multiple shortest paths exist between u_1 and v_j, we take the path with the largest weight.

Since consumers may have preference for different entity types, the profile is organized into several separate vectors based on the entity types. For example, two consumers may have similar profiles because they both talked about the same drug, or because they are interested in the same disease. In another word, consumers can be similar in different ways. By taking into account the heterogeneity of the network, we could provide personalized recommendation to consumers. For example, if consumers prefer to find similar users that are interested in same drugs, then we could set a higher weight for the drug entities vectors.

The weight schema of the edges of a heterogeneous network plays an important role in determining the similarity of users. In the Profile P_u^d, $w(e_{uv})$ denotes the weight between node u and node v. One method is considering all edges equally important and let $w((e_{uv})) = 1$. However, this is not the case for most real world information network. A consumer would be more interested in one drug than the other even if he/she talked about both in online forums. To deal with this problem, we may use the co-occurrence frequency between two nodes as the weight of the edge. Nevertheless, simply using co-occurrence frequency could favor the nodes with high frequency over those with much lower frequency. Another method is using association measurements to calculate the association strength between the two endpoints of an edge. In this study, we propose to use *leverage* and *lift*, which are defined as follows:

$$leverage(u, v) = support(e_{uv}) - support(u) \times support(v) \tag{6}$$

$$lift(u, v) = support(e_{uv}) / (support(u) \times support(v)) \tag{7}$$

where

$$support(e_{uv}) = freq(e_{uv}) / totalcount \tag{8}$$

$$support(u) = freq(u) / totalcount \tag{9}$$

where $freq(e_{uv})$ is the frequency of the edge, namely the co-occurrence frequency of node u and node v. $freq(u)$ denotes the frequency of node u and totalcount is the total number of messages in the dataset. $support(e_{uv})$ denotes the actual probability of co-occurrence of node u and node v the dataset. $support(u) \times support(v)$ is the probability of their co-occurrence if u and v are absolutely independent. Therefore, *leverage* measures the difference between the actual co-occurrence probability and the theoretical co-occurrence probability if the two nodes are independent, and *lift* reflects the division of the actual probability and theoretical probability. Both of them indicate the associations between a pair of nodes. For example, *lift* can also be written as following probability expression:

$$lift(u, v) = \frac{support(e_{uv})}{support(u) \times support(v)} = \frac{P(u, v)}{P(u) \times P(v)} = \frac{P(v|u)}{P(v)} \tag{10}$$

$lift(u, v) = 1$ denotes that node u and v are statistically independent with each other. $lift(u, v) > 1$ means that they are positively correlated, while $lift(u, v) < 1$ means they are negatively correlated. For both *leverage* and *lift*, the higher the values, the stronger the association strength.

Fig. 6 shows an example of two users' profiles in distance 1. There are four types of nodes in the network: user, drug, disease, and ADR. For a user, there could be three types of profile vectors: drug, disease, and ADR. In Fig. 6, we consider an unweighted network. According to the formulation of the Profile P_u^d, if we assign the same weight to the three types of entities, then the similarity between User1 and User2 should be 0.5. In Fig. 7, we used co-occurrence frequency as the weights of the edges. In this case, we can get a similarity score of 0.47.

Fig. 8 demonstrates the two users' ProfileNet similarity in distance 2. As the distance increases, more related nodes would be included in a user's profile. Meanwhile, more irrelevant nodes could also be includes. Therefore, the similarity score would not necessarily increase or decrease as the distance increases. It depends on the users' ego-centered network structure. If more related nodes were included, the similarity score will increase. Otherwise, it will decrease. In this example, the similarity score decreased a little bit.

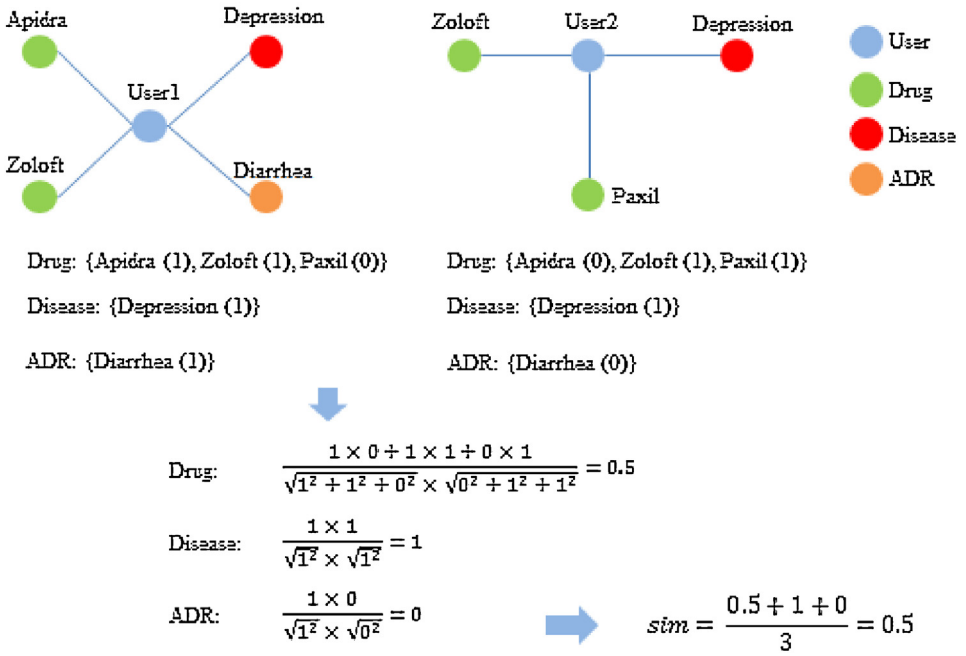


Fig. 6. Example of Calculating Similarity between User1 and User2 using ProfileNet in Distance 1 (Unweighted).

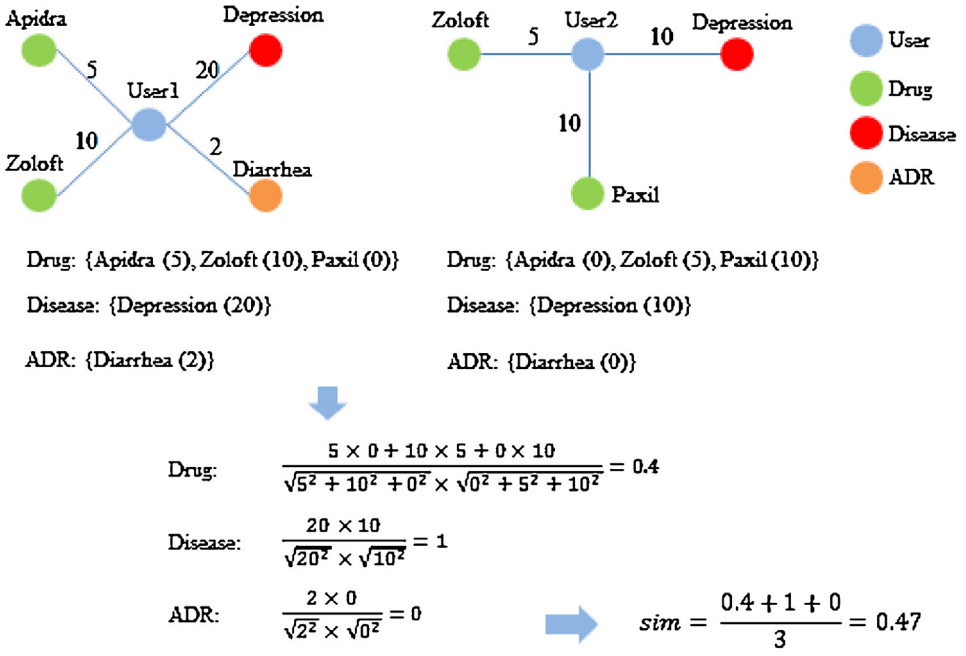


Fig. 7. Example of Calculating Similarity between User1 and User2 using ProfileNet in Distance 1 (Weighted – co-occurrence).

4.3. Global similarity (HealthRank)

The proposed ProfileNet measures similarity between consumers from a local perspective. It considers the overlap of two consumers' neighbors within certain distance in the network. However, the local similarity has one shortcoming. The similarity of two consumers cannot be measured if they have no common neighbors in certain distance. For example, in Fig. 9, *User1* and *User2* are directly connected to two different drugs *Drug1* and *Drug2*. In this case, the local similarity would consider them dissimilar in terms of Drug in distance 1. Nevertheless, even if they are not connected to the identical Drug Entity in the network, there could be some degree of similarity between them if *Drug1* and *Drug2* are both used to treat the same disease *Disease1*. And the two drugs could be connected to *Disease1* somewhere in the network. We may address the problem by increasing the distance d in ProfileNet. However, simply increasing d would include more and more irrelevant nodes and impact the performance. To deal with this problem, we should estimate the similarity between consumers from a global point of view.

HealthRank is inspired by SimRank [36], which is a well-known algorithm for global similarity, based on the idea that two objects are similar if they are referenced by similar objects. SimRank is first proposed for measuring nodes similarity in homogeneous directed

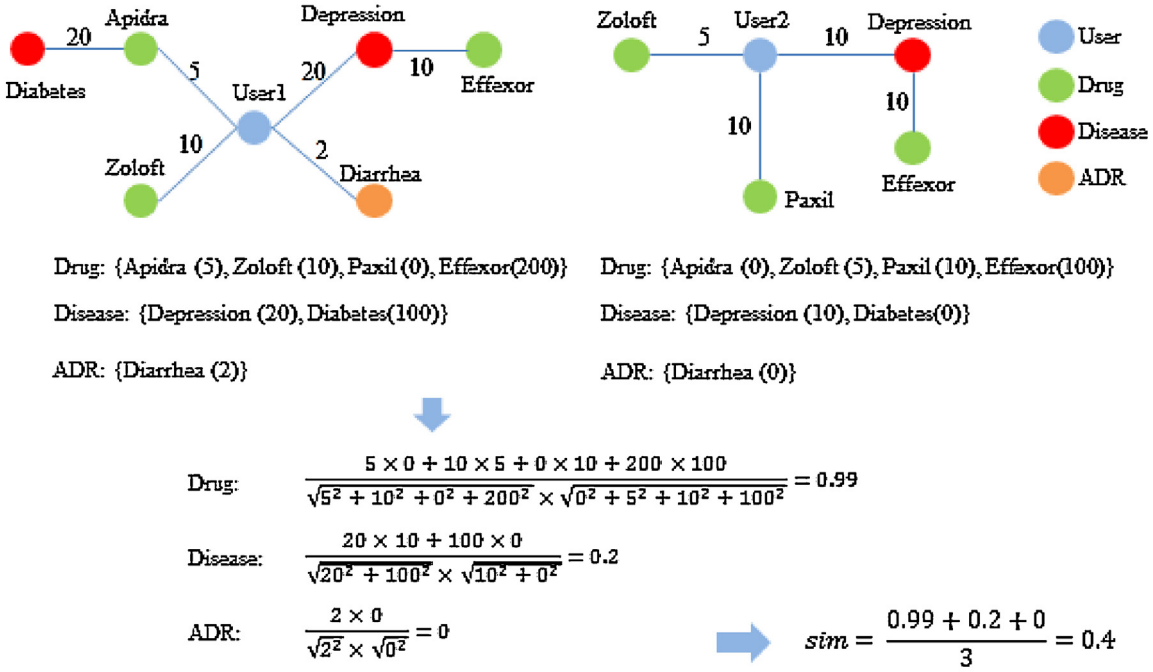


Fig. 8. Example of Calculating Similarity between User1 and User2 using ProfileNet in Distance 2 (Weighted – co-occurrence).

Fig. 9. Problem with Local Similarity (users may not have common immediate neighbors but they can be connected through other entities with a certain distance).

graph, and the authors only considered in-links for similarity calculation.

Given two nodes u and v in a directed network, $I(u)$ and $I(v)$ denote the set of in-neighbors of u and v, respectively. Individual in-neighbors are denoted as $I_i(u)$, where $1 \le i \le |I(u)|$. SimRank calculate similarity of two nodes based on the following equation:

$$s_0(u,v) = \begin{cases} 0 & (if u \neq v) \\ 1 & (if u = v) \end{cases} \tag{11}$$

$$s_{k+1}(u,v) = \frac{C}{|I(u)||I(v)|}\sum_{i=1}^{|I(u)|}\sum_{j=1}^{|I(v)|} s_k\left(I_i(u), I_j(v)\right) \tag{12}$$

where C is a constant between 0 and 1. Since it is possible that u or v have no in-neighbors, $s(u,v) = 0$ when $I(u) = \phi$ or $I(v) = \phi$.

SimRank calculate similarities between nodes by iteratively updating the similarities values of each pair of nodes in the network. However, the conditions are different when it comes to heterogeneous information network. Since nodes in heterogeneous information network belong to different types of entities, it would be misleading to calculate the similarity of two nodes of different type. For example, it is impossible to measure the similarity between a consumer and a disease. The global similarity should be aggregated from the similarities of nodes of the same type. In addition, the heterogeneous health information network constructed in previous section is an un-directed network. We need to consider all neighbors of a node rather than just in-neighbors as in a directed network. In light of this, we propose a global similarity approach called HealthRank for heterogeneous healthcare information network.

Given two nodes u and v in an un-directed heterogeneous information network, $N(u)$ and $N(v)$ denote the set of neighbors of u and v, respectively. Individual neighbors are denoted as $N_i(u)$, where $1 \le i \le |N(u)|$. $\varphi(u)$ denotes the entity type of u. The similarity of two nodes is calculated using the following equation:

$$s_0(u,v) = \begin{cases} 0 & (if u \neq v) \\ 1 & (if u = v) \end{cases} \tag{13}$$

$$s_{k+1}(u,v) = \begin{cases} 0 & (if \varphi(u) \neq \varphi(v)) \\ \frac{C}{|N(u)||N(v)|}\sum_{i=1}^{|N(u)|}\sum_{j=1}^{|N(v)|} s_k\left(N_i(u), N_j(v)\right) & (if \varphi(u) = \varphi(v)) \end{cases} \tag{14}$$

This approach considers all links equally important, but we could also use different weight schema to reformulate Eq. (13) as follows:

$$s_{k+1}(u,v) = \frac{C}{|N(u)||N(v)|}\sum_{i=1}^{|N(u)|}\sum_{j=1}^{|N(v)|} s_k\left(N_i(u), N_j(v)\right) \times w\left(e_{uN_i(u)}\right) \times w\left(e_{vN_j(v)}\right) \tag{15}$$

Unlike local similarity method, global similarity method calculated the similarity of each pair of nodes in the network at the same time. Even if two nodes do not share any common neighbors, there might still be certain degree of similarity between them if they are connected to similar nodes.

5. Experiment

In order to test the proposed methods, we collected data from a popular online health social website. In this section, we introduce the experiments and discuss the results.

5.1. Dataset

We collected the dataset from a popular online health social website MedHelp. Every month over 12 million people [12] seek healthcare information, receive support, share experiences on this

Table 1
General Information of the Constructed Network.

Network Information				
# Nodes	# Links	Avg. Degree	Density	Diameter
5130	69487	13.564	0.005	10
Number of Nodes				
#User Node	#Disease Node	#Drug Node	#ADR Node	
3059	1223	381	467	

website. New threads are initiated every day by consumers, talking about various kinds of medical topics. This public available consumer-contributed content is ideal for our study. In MedHelp, there are hundreds of Medical Support Communities, each of which focuses on a specific type of medical condition, such as heart disease, lung cancer, etc. Consumers can join in any communities of interest by starting posts and participate in discussions. We collected all threads from 8 communities, including heart disease, depression, lung cancer, breast cancer, skin cancer, gastric cancer, stomach cancer, and dental health. Then we randomly extract 200 threads with more than one participant from each community and combine them into a dataset. The dataset contains a total of 1327 threads. In each thread, there are several messages authored by individual users. There are a total of 6654 messages in the dataset.

5.2. Heterogeneous healthcare information network

In this study, we used message as analysis unit to extract entities as nodes and relations as links to construct the network. The reason we chose message over thread for nodes and links extraction is that topic digression usually occurs in thread. Especially in a long thread with tens of messages, messages posted by different users might be about totally different diseases or drugs. Under this circumstance, using co-occurrence analysis for relation extraction from the whole thread would generate a large number of false links. A message is usually focused on a single topic, so the entities in one message are more likely to be related to each other. However, user connections were extracted based on threads, because participating in the same thread indicates the co-occurrence between users. For node frequency and link frequency, multiple appearances in one message only accounted for one occurrence.

Given the above dataset, we used the proposed entity recognition and relation extraction methods to construct a heterogeneous healthcare information network. The network contains four types of nodes, including User, Drug, Disease, and ADR. Table 1 gives some general information of the constructed heterogeneous healthcare information network. As we can see, the network is quite sparse. We then applied the proposed similarity analysis approaches to this network.

5.3. Ground truth

Our study is motivated to find similar users from online health social websites and recommend to consumers. Some consumers with similar interests might have already interacted with each other in online threads, however, there are still some others that have never talked to each other and they might not even know the existence of each other. We should be able to measure not only the explicit similarity between consumers, but more importantly, the implicit similarity. In order to set up the ground truth for the experiment, two human annotators were recruited to go through all the messages authors by the consumers and determine whether a given pair of consumers is similar to each other.

In order to set up the ground truth, we select a sample of users from the dataset to conduct the experiment. Intuitively, the more messages a user posted, the more content would be available for the content-based similarity. In addition, the user would also have more connections to other nodes in the network. Based on this idea, we could speculate that similarity analysis methods might have different performance for active user and inactive users. Therefore, we put users into three groups with node frequency $5 < freq < 10$, $10 \leq freq \geq 50$, and $freq > 50$ respectively. We exclude users who have frequency less than 5 because these users are likely participating in the discussion randomly without particular interests. Then we randomly selected 5 users from each group. For each of the 15 users, we randomly select 25 users, and we have 375 pairs of users. The two human annotators then evaluate the 375 pairs of users' messages independently, and determine if each pair of users has similar healthcare interests. Here we only consider high level healthcare interest, such as breast cancer, dental health or diabetes, rather than detailed healthcare concerns such as lymph node, tumor size, or metastasis for breast cancer. Each pair of users was labeled as similar or dissimilar to each other. The two annotators are graduate students with healthcare informatics training and have extensive experience working with medical professionals and researchers.

We used the weighted Kappa [57] measure to compute the inter-rater agreement. Weighted Kappa measure is a statistical measure extended from Kappa measure for computing the agreement between two ordered lists. It has a maximum value of 1 which indicates a perfect agreement between two raters and a value of 0 when the agreement is not better than by chance. In general, a weighted kappa measure value larger than 0.8 is considered to be a very good agreement between the two raters. In our experiment, the weighted Kappa measure was 0.95, which means that two annotators reached a very good agreement and the ground truth was reliable.

5.4. Evaluation measurement

In this experiment, we use F1 score to measure the performance. For each target user of the 15 users, similarity analysis algorithms were used to calculate the similarity scores between the target user and the randomly selected 25 users. Then we got a list of users ranking in a descending order of the similarity scores. We consider the Top k ($1 \leq k \leq 25$) users are similar to the target user and use them for recommendation. If one user in Top k is considered as similar to the target user by the ground truth, then we have a True Positive (TP). However, if one user is determined as dissimilar with the target user but is included in Top k, then we have a False Positive (FP). If one user is considered as dissimilar with the target user and is not in the Top k, then it is a True Negative (TN). Finally, if one user is considered as similar user to the target user but it is not ranked in Top k, then we have a False Negative (FN). Based on the definition, we can calculate the following:

$$precision = TP/(TP + FP) \tag{16}$$

$$recall = TP/(TP + FN) \tag{17}$$

$$F1 = 2 \times precision \times recall/(precision + recall) \tag{18}$$

5.5. Results and discussion

Given the above dataset, ground truth, and evaluation measurement, we then conducted a series of experiments. For each target user of the 15 users, 25 users were randomly selected to set up the ground truth. According to the ground truth, an average of 5.27 out of 25 users is similar to a target user. Therefore, we set k = 1, 2, 3, 4, 5 for Top k and calculate the average F1 score from Top 1 to Top 5 for

Table 2
F1 Score for Different Approaches (Top1–Top 5) (Unweighted).

User Frequency	Top K	ContentSim	HealthRank	ProfileNet ($d=1$)	ProfileNet ($d=2$)	ProfileNet ($d=3$)
<10	1	0.34	0.26	0.26	0.34	0.00
	2	0.57	0.39	0.43	0.40	0.00
	3	0.74	0.48	0.45	0.47	0.07
	4	0.78	0.59	0.44	0.55	0.18
	5	0.73	0.60	0.45	0.62	0.19
	Avg.	**0.63**	**0.46**	**0.41**	**0.48**	**0.09**
10–50	1	0.32	0.27	0.19	0.19	0.19
	2	0.44	0.46	0.39	0.39	0.27
	3	0.49	0.60	0.48	0.48	0.28
	4	0.55	0.70	0.65	0.47	0.29
	5	0.53	0.74	0.66	0.50	0.29
	Avg.	**0.47**	**0.55**	**0.47**	**0.40**	**0.26**
>50	1	0.07	0.28	0.28	0.04	0.04
	2	0.14	0.47	0.39	0.16	0.14
	3	0.20	0.56	0.48	0.29	0.17
	4	0.31	0.73	0.56	0.35	0.16
	5	0.44	0.76	0.62	0.31	0.19
	Avg.	**0.23**	**0.56**	**0.46**	**0.23**	**0.14**

each proposed method. There were three groups of users, and each group contained 5 users. We used the average F1 score of users in each group for evaluation.

5.5.1. Comparison among methods

We applied the proposed methods to the dataset. Next, we will compare the performance of different methods and discuss the results in details. In this part of experiment, we treated all links in the network equally and let for all edges. And we tested the performance of ProfileNet within distance $d = 1, 2, 3$. Table 2 gives the detailed average F1 score for each group of users from Top 1 to Top 5 using different similarity analysis methods. Fig. 10 shows averaged F1 score, precision and recall of Top 1 to Top 5 for ContentSim, HealthRank, and ProfileNet ($d = 1$) respectively.

5.5.1.1. ProfileNet within different distance. A noticeable pattern in the results is that when the distance d increases, the performance of ProfileNet reduces except for users with less than 10 messages. ProfileNet measures similarity based on users' neighbors. Theoretically, when the distance increases, more nodes would be considered as the central node's neighbors. Therefore, more noise could be added into the calculation. However, for users who posted less than 10 messages, there might be only a very small set of neighbors, which makes it hard to find similar users. In this case, increasing d could be likely to bring in relevant nodes rather than noisy ones. But as d increases and exceeds certain threshold, more noise could be brought in. In this experiment, performed better for inactive users, and performed better for other user groups. Overall, $d = 1$ performed better on average.

5.5.1.2. Content-based similarity vs structural-based similarity. Another interesting trend could be observed between content-based method and structural methods in different user groups. As we can see, ContentSim achieve an average F1 score higher than 0.6 when users participate in a very small set of threads. As users get more active, ContentSim's ability in identifying similar users reduces. The F1 score of ContentSim for users who have posted more than 50 messages is as low as 0.23. Contrarily, the performance of both ProfileNet ($d = 1$) and HealthRank increases as users get more active.

The trend can be explained by the different underlying basic ideas of different methods. Content-based method calculates similarity between two users based on the threads authored by each user. Active users usually participate in a great number of threads, and theses threads could be about a diversity of topics. On the contrary, inactive users who only posted a few threads are likely to be interested in only one or two topics. Therefore, inactive users have more focused interests, while active users' interests are diluted by being spread over a large number of threads. In this case, content-based similarity could yield better results for inactive users. On the other hand, structural methods measure similarity based on users' structural context such as users' neighbors. Active users are connected to much more entities than inactive users. Thus, the local or global network of an active user bears richer structural information than that of an inactive user. Therefore, structural approaches performed better for active users.

5.5.1.3. Local similarity vs global similarity. Next, we compared the performance of local similarity (ProfileNet) and global similarity (HealthRank). Both methods outperformed ContentSim. However, global similarity (HealthRank) outperformed local similarity (ProfileNet) for all user groups.

ProfileNet measures the similarity between two users locally, based on their direct neighbors. However, this method cannot calculate the similarity score if the two users have no common neighbors. Especially in a sparse network, it is highly possible that two users are not directly connected to the same set of nodes but they are similar to some extent. One potential solution is increasing the distance d. For example, if one user is linked to "Prozac" while another user is connected to "Zoloft", ProfileNet with $d = 1$ would consider them as dissimilar. But if we let $d = 2$, we might find that both "Prozac" and "Zoloft" are connected to the disease "Depression", and now the two users are both related to "Depression". In this case, we could say that they are similar. Nevertheless, increasing d could also bring in noisy nodes and impact the performance, which is discussed above. Global similarity method addressed this problem well. By measuring similarity between two users based on the similarity of their neighbors, global similarity could overcome the scarcity problem and calculate similarity scores even if two users do not have immediate common neighbors.

5.5.2. Weight schema for structural methods

We also conducted experiments on local and global similarity approaches using different weight schemas for the edges in the network. Since ProfileNet with $d = 1$ performed better on average for all user groups, we only conducted experiment on ProfileNet with. Figs. 11 and 12 demonstrate the results for local and global

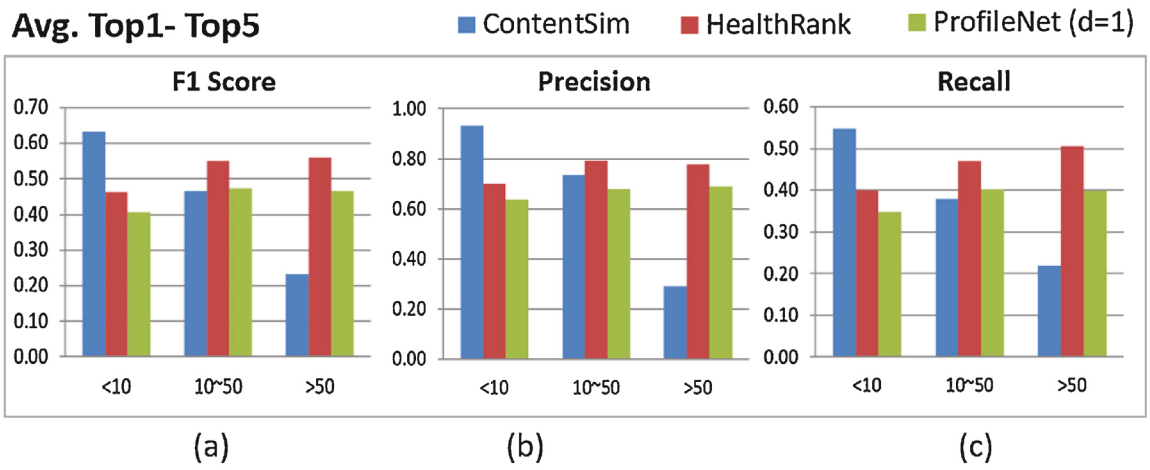


Fig. 10. F1 Score (a), Precision (b) and Recall (c) for Different Approaches (Averaged Top1–Top 5) (Unweighted).

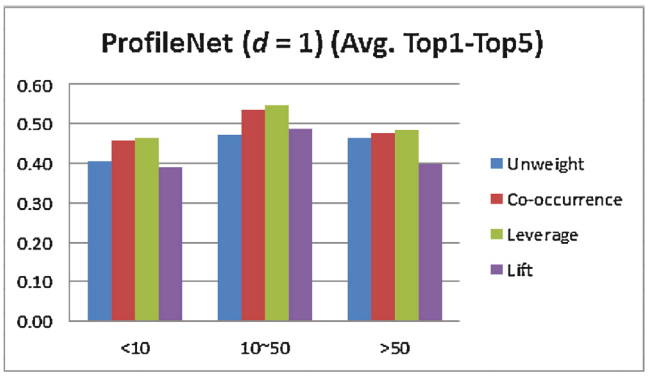


Fig. 11. F1 Score for ProfileNet with Different Weight Schema.

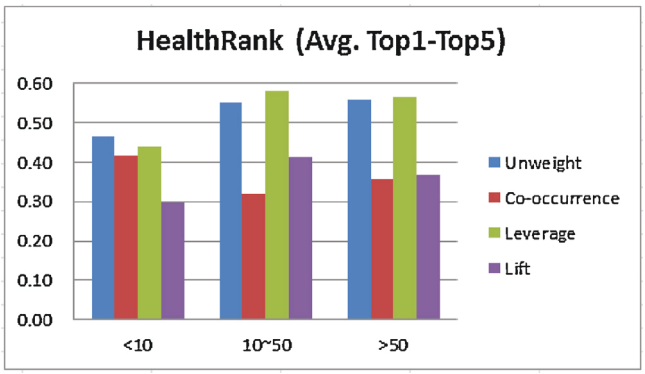


Fig. 12. F1 Score for HealthRank with Different Weight Schema.

Table 3
F1 Score for ProfileNet and HealthRank with Different Weight Schema.

User Frequency	Unweight	Co-occurrence	Leverage	Lift
HealthRank				
<10	0.46	0.42	0.44	0.30
10–50	0.55	0.32	0.58	0.41
>50	0.56	0.36	0.57	0.37
ProfileNet ($d = 1$)				
<10	0.41	0.46	0.47	0.39
10–50	0.47	0.54	0.55	0.49
>50	0.46	0.48	0.48	0.40

approaches respectively, and Table 3 shows the detailed F1 score of different weight schemas for both approaches.

As we can see, for ProfileNet, Leverage and Co-occurrence were comparable and performed better than the other two. As for the HealthRank, Leverage and Unweight performed better. Co-occurrence frequency is the easiest way to measure the strength of a connection between two nodes. However, it performs the worst for HealthRank. A possible reason could be that the similarity can be biased due to some nodes with high frequencies. The weighted global similarity approach calculates the similarity by multiplying the weights, so the nodes with very high frequency are more likely to have higher similarity scores with other nodes. The local similarity calculates the cosine similarity between two profiles, so it mainly depends on the angle between the two vectors and is less sensitive to the absolute value of the co-occurrence frequency. Leverage performs the best for both approaches.

5.5.3. Hybrid methods

From the above results, we can see that content-based method performed differently from structural-based methods for different user groups. Local and global similarity methods also have different results among different user groups. We may speculate that different methods may capture different aspects of the similarity relationship between two users. If we integrate the methods, we may get a more comprehensive measurement of user similarity. Therefore, we used logical operator AND/OR to integrate different similarity measuring methods. Specifically, given two different methods M1 and M2, if the similarity relationship of two users were captured by both methods, M1 AND M2 considers them as similar to each other. If the similarity relationship were captured by either method, M1 OR M2 considers the users as similar to each other. As shown in the previous section, using leverage as the link weight produced the best results in structural similarity. As a result, we used leverage as the link weight in this part of experiment.

We first used logical operators to combine content-based method and structural methods. In this work we introduced two different structural methods ProfileNet and HealthRank. Therefore, we combined the content-based method with each of the two structural methods respectively. As we can see in Figs. 13 and 14, the two sets of combination have very similar results. AND operator generated the worst performance, while OR operator produced the best performance except for active user group. Both of "ContentSim OR ProfileNet" and "ContentSim OR HealthRank" generated the best performance in inactive user group, and the performance reduced as users get more active. This result may be influenced by the trend of content-based method. In spite of that, we can tell that the OR operator captured the strength of both methods and improved the performance for some user groups.

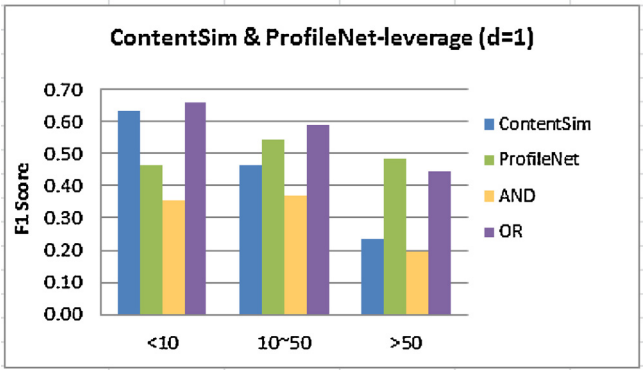


Fig. 13. F1 Score for the Combination of ContentSim and ProfileNet.

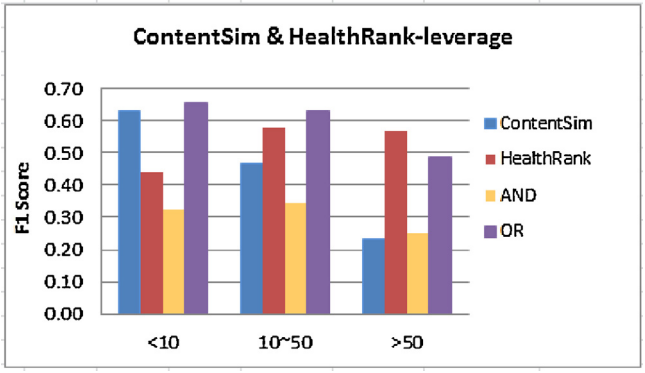


Fig. 14. F1 Score for the Combination of ContentSim and HealthRank.

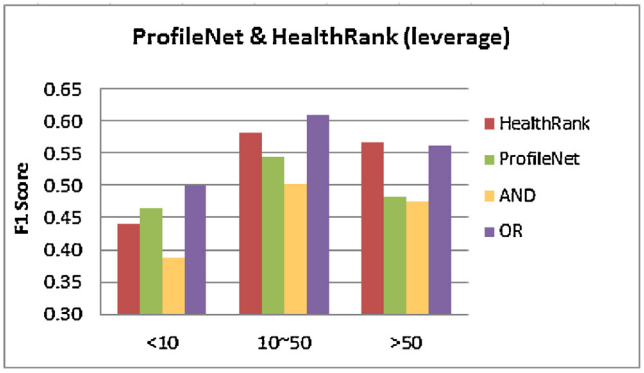


Fig. 15. F1 Score for the Combination of ProfileNet and HealthRank.

We then combined the local and global structural similarity methods, and Fig. 15 illustrates the results. The AND operator still performed the worst. Local OR global structural similarity produced the best results, and performed comparably with HealthRank for active users. The results implied that local and global structural could capture different aspects of the relationship of two users. When we combine the two methods together, we could get a better performance than either individual method.

Finally, we combined the content-based method with "ProfileNet OR HealthRank" to test if we can further improve the performance. As shown in Fig. 16, the OR operator performed better than the other methods except HealthRank in active user group. The OR combination of the methods is significantly impacted by the content-based method, which performed poorly for active users. Therefore, HealthRank still outperformed the OR operator here.

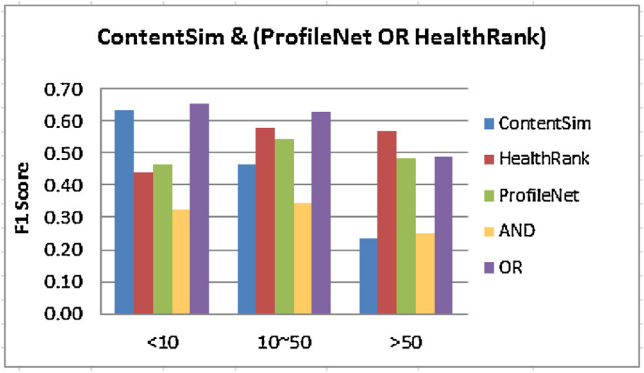


Fig. 16. F1 Score for the Combination of ContentSim and (ProfileNet OR HealthRank).

5.6. Limitations

In this study, we used external dictionaries to extract different types of named entities from healthcare social media data. Since the consumer language problem is quite severe for ADRs, we used CHV to address the problem. We observed that most consumers used formal expressions for drugs and diseases, so we only used professional vocabularies for identifying drug and disease names. However, we may still miss some consumer expressions for drugs and diseases. Consumers would refer to drugs or diseases using their own slang or daily language instead of universally recognized names. This could be improved in the future by applying CHV to identifying drug and disease names.

In order to extract relationships between different entities, we applied co-occurrence analysis to find connections among entities. Therefore, we only captured one type of relationship between different types of entities. To determine the explicit relationship, we may need to resort to sophisticated natural language processing and semantics analysis techniques. Also, the relation extraction can be supported by using external ontologies or domain expert knowledge. We may use existing knowledge-based relationships to build connections between entities, such as drug-treat-disease, drug-cause-ADRs, and so on. In this work, we captured the heterogeneity of different types of entities in the network, and we can further improve the work by supplementing different types of relationships into the network.

Another limitation of this study is that we evaluated the methods on a limited dataset. And that is related to the inherent problem of human annotations. Manual annotation is time-consuming and laborious. In the future, we would like to explore alternative evaluation methods with less human efforts and use larger dataset for validation.

6. Conclusions

Online health social websites have established platforms where consumers can share healthcare information and provide social support for each other. However, the relationships between consumers in online health social websites are built on weak social ties. Social networking activities are usually based on common interests rather than long-term relationships. Therefore, it would be difficult for consumers to find similar users efficiently when they need useful informational or emotional support. In this work, we proposed to consider an online health social website as a heterogeneous healthcare information network, and explored methods for constructing such a network. After that, we introduced three different methods for measuring user similarity in an online health social website, including a content-based method and two structural methods.

As demonstrated in the experiments, content-based method performed better than structural methods for inactive user group. Inactive users that posted a very small set of messages usually have very focused interests, which could be effectively captured by content-based methods. On the other hand, structural methods yielded better performance for users with more than ten messages. This is because active users are connected to more entities in the network, so richer structural information would be available for structural methods to measure the similarity. The results also showed that global similarity method outperformed local similarity method. Local similarity only considers direct connections between nodes in the network, while global similarity takes into account indirect connections. In this case, global similarity could deal with sparse networks and capture the implicit similarity between two users. We also tested the performance of structural methods using different weight schema, and it was shown that leverage performed the best for both local and global similarity. Finally, we used logical operator to combine different methods. We demonstrated that different methods could capture different aspects of user similarity, and we could get a more comprehensive measurement by combining different methods together.

For future study, we will develop more structural methods for measuring similarity between nodes in a heterogeneous healthcare information network. In addition, the structural methods introduced in this work not only can be used to measure the similarity between users, but also can be extended to address other problems. For example, we could represent threads by entity nodes and recommend threads to users who might have similar interests.

Acknowledgement

This work was supported in part by the National Science Foundation under the Grant IIS-1650531 and DIBBs-1443019.

References

[1] DiMatteo MR. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychol 2004;23(2):207.
[2] Chuang K, Yang CC. Social support in online healthcare social networking. In: Presented at the proceedings of iConference. 2010.
[3] Coursaris CK, Liu M. An analysis of social support exchanges in online HIV/AIDS self-help groups. Comput Hum Behav 2009;25(4):911–8.
[4] Coulson NS. Receiving social support online: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. Cyber Psychol Behav 2005;8(6):580–4.
[5] Davis DZ, Calitz W. Finding healthcare support in online communities: an exploration of the evolution and efficacy of virtual support groups. In: Handbook on 3D3C Platforms. Springer; 2016. p. 475–86.
[6] Jiang L, Yang CC. Using co-occurrence analysis to expand consumer health vocabularies from social media data. In: Presented at the proceedings of IEEE international conference on healthcare informatics. 2013.
[7] Zhang Y, Wang P, Heaton A, Winkler H. Health information searching behavior in MedlinePlus and the impact of tasks. Proceedings of the 2nd ACM SIGHIT international health informatics symposium 2012:641–50.
[8] Hodges SD, Kiel KJ, Kramer AD, Veach D, Villanueva BR. Giving birth to empathy: the effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. Pers Soc Psychol Bull 2010;36(3):398–409.
[9] Zhang Y, He D, Sang Y. Facebook as a platform for health information and communication: a case study of a diabetes group. J Med Syst 2013;37(3):1–12.
[10] Zhang M, Yang CC. Using content and network analysis to understand the social support exchange patterns and user behaviors of an online smoking cessation intervention program. J Assoc Inf Sci Technol 2015;66(3):564–75.
[11] Wright KB, Bell SB. Health-related support groups on the Internet: linking empirical findings to social support and computer-mediated communication theory. J Health Psychol 2003;8(1):39–54.
[12] MedHelp (March). About MedHelp. Available: http://www.medhelp.org/aboutus.htm.
[13] Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies. Synth Lect Data Mining Knowl Discov 2012;3(2):1–159.
[14] Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 workshop on biomedical natural language processing 2010:117–25.
[15] Patki A, Sarker A, Pimpalkhute P, Nikfarjam A, Ginn R, O'Connor K, et al. Mining adverse drug reaction signals from social media: going beyond extraction. Proc BioLinkSig 2014;2014:1–8.
[16] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(1):343.
[17] Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC. Exploring lexical forms: first-generation consumer health vocabularies. Proceedings of the AMIA 2006.
[18] Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011;44(6):989–96.
[19] Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of adverse drug reactions from user comments. Proceedings of the AMIA Annu Symp Proc 2011:1019–26.
[20] Pranoti Pimpalkhute M, Apurv Patki M. Phonetic spelling filter for keyword selection in drug mention mining from social media; 2014.
[21] Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O'Connor K, Sarker A, et al. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing 2014.
[22] Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015:ocu041.
[23] Bian J, Topaloglu U, Yu F. owards large-scale twitter mining for drug-related adverse events. Proceedings of the 2012 international workshop on Smart health and wellbeing 2012:25–32.
[24] Yang CC, Jiang L, Yang H, Tang X. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of ACM SIGKDD workshop on health informatics. 2012.
[25] Yang CC, Yang H, Jiang L. Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. ACM Trans Manage Inf Syst (TMIS) 2014;5(1):2.
[26] Liu X, Chen H. Identifying adverse drug events from patient social media. presented at the IEEE intelligent systems 2015.
[27] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval, vol. 463. New York: ACM press; 1999.
[28] Herlocker JL, Konstan JA, Borchers A, Riedl J. An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval 1999:230–7.
[29] Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Proceedings of the fourteenth conference on uncertainty in artificial intelligence 1998:43–52.
[30] Wang J, De Vries AP, Reinders MJ. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval 2006:501–8.
[31] Xiao J, Zhang Y, Jia X, Li T. Measuring similarity of interests for clustering web-users. Proceedings of the 12th Australasian database conference 2001:107–14.
[32] Pennacchiotti M, Gurumurthy S. Investigating topic models for social media user recommendation. Proceedings of the 20th international conference companion on world wide web 2011:101–2.
[33] Chen J, Geyer W, Dugan C, Muller M, Guy I. Make new friends, but keep the old: recommending people on social networking sites. Proceedings of the SIGCHI conference on human factors in computing systems 2009:201–10.
[34] Newman ME. Clustering and preferential attachment in growing networks. Phys Rev E 2001;64(2):025102.
[35] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Am Soc Inf Sci Technol 2007;58(7):1019–31.
[36] Jeh G, Widom J. SimRank: a measure of structural-context similarity. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining 2002:538–43.
[37] Lizorkin D, Velikhov P, Grinev M, Turdakov D. Accuracy estimate and optimization techniques for simrank computation. Proc VLDB Endow 2008;1(1):422–33.
[38] Li C, Han J, He G, Jin X, Sun Y, Yu Y, et al. Fast computation of simrank for static and dynamic information networks. Proceedings of the 13th international conference on extending database technology 2010:465–76.
[39] Pan J-Y, Yang H-J, Faloutsos C, Duygulu P. Automatic multimedia cross-modal correlation discovery. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining 2004:653–8.
[40] Symeonidis P, Tiakas E, Manolopoulos Y. Transitive node similarity for link prediction in social networks with positive and negative links. Proceedings of the fourth ACM conference on recommender systems 2010:183–90.
[41] Xiang B, Chen E-H, Zhou T. Finding community structure based on subgraph similarity. In: Complex networks. Springer; 2009. p. 73–81.
[42] Hsu WH, King AL, Paradesi MS, Pydimarri T, Weninger T. Collaborative and structural recommendation of friends using weblog-based social network analysis. Proceedings of the AAAI spring symposium: computational approaches to analyzing weblogs 2006:55–60.
[43] Zhao P, Han J, Sun Y. P-Rank: a comprehensive structural similarity measure over information networks. Proceedings of the Proceedings of the 18th ACM conference on Information and knowledge management 2009:553–62.

[44] Sun Y, Han J, Yan X, Yu PS, Wu T. Pathsim: meta path-based top-k similarity search in heterogeneous information networks. VLDB'11 2011.
[45] Yu X, Sun Y, Norick B, Mao T, Han J. User guided entity similarity search using meta-path selection in heterogeneous information networks. Proceedings of the 21st ACM international conference on information and knowledge management 2012:2025–9.
[46] Sun Y, Barber R, Gupta M, Aggarwal CC, Han J. Co-author relationship prediction in heterogeneous bibliographic networks. Proceedings of the advances in social networks analysis and mining (ASONAM), 2011 international conference on 2011:121–8.
[47] Sun Y, Han J, Aggarwal CC, Chawla NV. When will it happen?: relationship prediction in heterogeneous information networks. Proceedings of the fifth ACM international conference on Web search and data mining 2012:663–72.
[48] Denecke K. Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study. In: Presented at the the fourth workshop on building and evaluating resources for health and biomedical text processing. 2014.
[49] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;42(D1):D1091–7.
[50] DrugBank (July 20). About DrugBank. Available: http://www.drugbank.ca/about.
[51] UMLS (July 20). UMLS Metathesaurus. Available: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.
[52] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(1).
[53] SIDER (July 20). About SIDER. Available: http://sideeffects.embl.de/about/.
[54] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000;356(9237):1255–9.
[55] Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24–9.
[56] Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC. Exploring lexical forms: first-generation consumer health vocabularies. Proceedings of the AMIA annual symposium proceedings 2006:1155.
[57] Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213.