Enriching User Experience in Online Health Communities Through Thread Recommendations and Heterogeneous Information Network Mining

Christopher C. Yang and Ling Jiang

***Abstract*— Online health communities (OHCs) provide health consumers with platforms for discussing medical conditions and sharing a personal experience. Although a wealth of healthcare information is available in OHCs, consumers find it challenging to locate information of interest efficiently due to the information overload. The lack of medical knowledge and searching skills makes it even harder for consumers to retrieve demanded information from a popular OHC with hundreds of thousands of threads. Therefore, effective thread recommendation is critical for OHCs to enhance user experience and engage the users in the community. In this paper, we proposed to recommend threads to users in OHCs by exploiting heterogeneous healthcare information network mining. We first constructed a heterogeneous healthcare information network from OHCs data. Unlike bipartite graphs studied in most existing works, which only consider user nodes and item nodes, a heterogeneous healthcare information network retains the rich context information of users and threads. We extracted features from the network to capture basic network metrics, thread–thread relationship, and user–user relationship, and utilize the features to train a binary classification model for thread recommendation. Experiments were conducted using a data set collected from MedHelp. The proposed approach was proven to be effective in measuring user interests in online discussion threads. In addition, by testing our approaches using different settings, we found that the local similarity achieved better performance than the global similarity in heterogeneous information network. By incorporating thread–thread relationship and user–user relationship, it can achieve the best performance.**

***Index Terms*— Online health community (OHC), recommendation, social media analytics.**

I. Introduction

THE dramatic development of social media has boosted the growth of online health communities (OHCs), such as MedHelp and PatientsLikeMe. These OHCs established communication platforms for social interactions such as discussion forums and online social groups. Health consumers discuss medical conditions and treatments with peer health consumers through these platforms. In addition, they share personal experiences and provide social support for those who are suffering from medical conditions. Social support has been found to be critical in helping patients to cope with stressful health conditions. Evidence showed that social support is beneficial for health outcomes by enhancing patient adherence to medical treatment [1]. The most common social support usually found in OHCs is informational support and emotional support [2]. Informational support helps consumers to reduce uncertainty by offering facts or knowledge, including advice, information referral, insight from personal experiences, or opinions [2], [3]. Liu *et al.* [4] developed the CARE framework to incorporate global and local context to extract sentences containing patient experience. Through the extracted patient experience, consumers can see what their peers are doing or experiencing by joining the discussion in OHCs, and thus enables automated selection of "relevant information" [5]. More importantly, consumers could receive emotional support from other consumers. OHCs help consumers to find emotional resonance by social networking with similar consumers. 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 [6]. Therefore, consumers could benefit from OHCs in terms of satisfying both informational and emotional needs.

OHCs empower health consumers to actively participate in their own healthcare and promote communication and collaboration between people. Nevertheless, will these OHCs combat the "Law of Attrition" (the phenomenon that users lose interest and stop using online health applications over time) [7]? OHCs capture an enormous amount of evolving consumer-contributed healthcare content that, however, comes with inherent challenges. It is no easier than looking for a needle in a haystack for consumers to locate relevant information in an OHC with hundreds of thousands of threads on various health-related topics, not to mention most of them are not skilled information searcher, who are familiar with the search engine mechanism and OCH architecture. On the one hand, consumers use very different languages from professional terminologies to describe their healthcare issues [8], and this language gap directly results in poor query formation [9]. On the other hand, consumers usually cannot fully understand their health conditions due to the lack of medical

Manuscript received May 24, 2018; revised October 18, 2018; accepted October 24, 2018. Date of current version December 3, 2018. This work was supported by the National Science Foundation under Grant NSF-1741306, Grant IIS-1650531, and Grant DIBBs-1443019. *(Corresponding author: Christopher C. Yang.)*

C. C. Yang is with the College of Computing and Informatics, Drexel University, Philadelphia, PA 19104 USA (e-mail: chris.yang@drexel.edu).

L. Jiang was with the College of Computing and Informatics, Drexel University, Philadelphia, PA 19104 USA. She is now with The Washington Post, Washington, DC 20071 USA.

Digital Object Identifier 10.1109/TCSS.2018.2879044

knowledge, which would further impede effective information searching [9]. These issues make it much harder to keep users engaged in OHCs. In order to encourage consumers to actively participate in OHCs, it is imperative to facilitate their access to demanded information, as well as connection with peers who are interested in similar healthcare topics. Asking for informational support from similar consumers could be much more efficient than relying on browsing or using the website search function. Moreover, social connectivity with similar consumers not only provides a shortcut to relevant healthcare information, but also helps consumers to find emotional resonance. Therefore, effective recommendation systems are critical for OHCs to enhance user experiences and encourage them to stay longer.

In [10], we explored the user recommendation problem in OHCs. By introducing similar peers with common health concerns to consumers, a user recommendation system could lead consumers to the pool of information that is most likely of interest to them. However, consumers still need to explore in this pool and make further selections, which could be difficult for some consumers with limited medical knowledge and searching skills. In addition, consumers could miss lots of valuable information if they only focus on this small pool of information. Therefore, user recommendation is mainly focused on encouraging the connections between users. It leads consumers to those who have common health concerns or experience so that they could seek informational or emotional support from their peers. Enhancing social networking activities between users is beneficial for information propagation, but we should not rely on it for that purpose. Hence, a thread recommendation system is needed to specialize in recommending threads to consumers in OHCs. Thread recommendation finds the information that could be of most interests to consumers and push the information to them directly. It can make it much easier for consumers to efficiently identify relevant information from hundreds of thousands of threads. In this way, consumers could participate in the discussion more actively and stay engaged in the communication on the topics with their peers. In this paper, we investigate the thread recommendation problem in OHCs.

Existing recommendation techniques usually fall into two categories: content-based recommendation and collaborative filtering recommendation [11]. Content-based approaches [12]–[15] analyze the textual content features of items, and recommend items that are similar to those previously preferred by users. Collaborative filtering approaches [16]–[19] recommend items that are liked by similar users in the past. Most of the current commercial recommender systems are built on collaborative filtering approach. In order to deal with the cold star problem, content-based approaches are often incorporated with collaborative filtering to boost the recommendation performance [20], [21].

Nevertheless, traditional recommender systems could not be effectively applied to thread recommendations in OHCs, which is more challenging due to the following reasons. First, rating information, which is the typical feature used for recommendation in electronic commerce websites, is usually not available in OHCs. For example, users are always asked to rate a product after purchasing it on Amazon.com, whereas OHCs rarely request thread ratings from their users. Consumers' interests in threads can only be implied by their participation and activeness in the discussion. Second, timeliness is a critical factor to consider in OHC thread recommendation, since consumers are usually attracted to new threads. As a large number of new threads are created every day, most of them would not be surfaced again after a short period of time partly due to the limitation of the browsing and searching capabilities of the system. Traditional recommender systems do not perform well on fresh new threads, known as cold start problem, but when they collect enough replies to make good prediction, the threads would have been inactive for a while already. Last but not least, massive lurkers and threads without participants make the user–thread matrix much sparser than user–item matrix in electronic commerce websites. Due to the abovementioned characteristics of OHCs, a more effective thread recommendation approach is desirable.

We propose to mine heterogeneous healthcare information network for OHC thread recommendations in this paper. The vast volume of consumer content in OHCs forms a huge healthcare information network. A healthcare information network is a network that captures the associations among the healthcare entities being discussed in the OHCs. The healthcare entities include disease, symptoms, drugs, adverse drug reactions (ADRs), treatments, patients, and more. These healthcare entities are extracted from the discussion threads, and their associations are measured based on the frequency of each entity and the co-occurrence frequency of a pair of entities in the heterogeneous network. Hidden in this huge health information network is the key to answering important questions. We need to explore the power of links in this network to reveal the hidden knowledge [22].

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 [23]. Healthcare information network is one typical heterogeneous network. In an online healthcare information network, nodes can be consumers, professionals, diseases, drugs, ADRs, etc. Links can be drug-treat-disease relationships, drug-cause-adverse reactions relationship, consumer-have-disease relationships, or consumer-take-drug relationships. Although lots of studies have been done on homogeneous information network, heterogeneous information networks can better represent real-world objects. 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 [23].

Social networks such as friendship networks and trust networks have been studied for recommendation in previous studies [24]. However, OHCs are very different from traditional social websites. There are not explicit friendship networks in OHCs, so it is difficult to directly apply social network-based approaches in OHCs. Also, some existing studies on recommendation represent user–item interactions as bipartite graphs for recommendation [25], [26]. A bipartite graph only contains

two sets of nodes: user nodes and item nodes. When a new user posts a new thread in an OHC, the user node and the thread node will not be connected to any nodes in a bipartite graph because the user is not making comments on a particular commercial item such as book, movie, or an electronic device, but discussing a healthcare issue or offering informational and social supports. Therefore, a bipartite graph cannot effectively handle such situations in OHCs. By representing the healthcare social media data as a heterogeneous information network, we are able to keep rich context information about threads and users as well as construct a relationship network for analyzing similarity between different types of objects. Heterogeneous information networks provide us with rich context information about a node, which could be critical in unveiling some underlying patterns. By harnessing the heterogeneous information network, we can make better prediction on user preference in thread recommendations.

The remainder of this paper is organized as follows. We will discuss related works in Section II, and then introduce the proposed methods in Section III. We present the experiments in Section IV, and conclude this paper in Section V.

II. Related Work

A. Recommendation Techniques

Typically, most existing recommendation systems use two major approaches: content-based approach and collaborative filtering approach.

Content-based approach has originated from information retrieval [27], [28] and information filtering studies [29]. Content-based recommendation systems analyze the content of textual information of user and items, and calculate similarities of user interests and items features for recommendation [12]–[15]. However, the limitation of content-based approach is that there needs to be enough content information for analysis. Otherwise, features need to be either automatically extracted from the systems or manually assigned to users or items [12]. Some domains may need expert knowledge or ontologies for extracting features, while some others have inherent problems with automatic feature extraction, such as multimedia data [11]. Content-based approach suffers from new user problem. When a new user becomes a new member of the system, there would not be enough information for analyzing the user preference and it would be difficult to compute the similarity between user interests and item profiles.

In contrast to content-based approach, collaborative filtering approach predicts a user preference on an item by utilizing the preferences of the user's neighbors on this item. The underlying assumption is that if two users have similar preferences or tastes, they will rate the same item similarly [30]. Collaborative filtering approaches can be further divided into memory-based and model-based methods [30]. Memory-based methods are the most popular methods, widely applied in many commercial recommender systems [17]. Memory-based methods include user-based [18], [19] and item-based [16], [17] approaches. User-based approaches predict a user rating on an item by aggregating the ratings of N most similar users of the target user. And, the similarity between users is calculated based on the user ratings of previously rated items. On the other hand, item-based approaches leverage the ratings of the similar items rated by the user in the past for predicting the user preference on the target item. The problem with memory-based method is that it cannot deal with the new user or new item problem. In addition, memory-based method cannot achieve good performance on sparse data, since it rely on similarity values between users or items. In model-based approaches, statistical or machine learning techniques, such as clustering [31], [32], latent semantic analysis [33], and matrix factorization [34]–[36], are utilized to learn model from the rating data. The model will then be used for prediction. Since collaborative filtering approach mainly relies on the user–item matrix for prediction, it suffers from severe sparsity problem. The cold star problem occurs when a new user or a new item enters the system, and a collaborative filtering system will fail to provide good recommendations in such cases.

In order to overcome the sparsity problem, many studies proposed hybrid recommendation approaches that combined different approaches, including content-based, collaborative filtering, and knowledge-based to boost the performance. Melville *et al.* [20] presented a content-boosted collaborative filtering method. They used a naive Bayesian text classifier to learn user profiles from the content information of rated movies and used the learned profiles to predict ratings for unrated movies. Then, the content-based predictor was combined with a collaborative filtering predictor to improve the recommendation performance. In addition to local information such as web page content, some researchers used external knowledge such as Wikipedia [21] to support collaborative filtering and improve predictions.

B. Bipartite Graph for Recommendation

As user/item neighborhood is critical for collaborative filtering to extract user/item similarity, many studies represent user–item interactions as bipartite graphs to build neighborhood models [25], [26]. Such a bipartite graph contains two types of node: user nodes and item nodes. Links in the bipartite graph only exists between nodes of different types [37], [38]. In a bipartite graph, the underlying relationship between users and items can be modeled by the graph structure even if they are not directly connected to each other. Many diffusion-based recommendation algorithms have been introduced in bipartite graphs [37], [39]–[41]. Huang *et al.* [39] used associative retrieval techniques and related spreading activation algorithms to generate transitive associations in a bipartite graph and then used the transitive associations in collaborative filtering to address the sparsity problem. Zhang *et al.* [42] proposed a recommendation algorithm based on an integrated diffusion in user–item–tag tripartite graphs. With the bipartite graph representation, the recommendation problem can be viewed as a link prediction problem. Huang *et al.* [43] summarized six linkage measures adapted for collaborative filtering recommendation, including common neighbors, Jaccard's coefficient, Adamic/Adar, preferential attachment, graph distance, and Katz. Some researchers also used graph-based features in machine learning techniques to construct recommendation models. Reddy *et al.* [44] utilized

a graph-based clustering algorithm to group similar users in bipartite graphs, and then used the generated user groups to improve recommendation. Li and Chen [45] proposed a kernel-based recommendation approach. They define a graph kernel on the user–item pair's context and use its graph structure to predict if there would be a link between them.

C. Recommendation in Social Media

The importance of users' social connections for recommendation has drawn more and more attention recently., Ziegler and Golbeck [46] demonstrated a significant correlation between the trust and their similarity based on the recommendations they made in the system. Studies showed that friendship and trustworthiness between users could effectively improve the recommendation performance, especially when the user–item matrix is very sparse [24], [47], [48]. Ma *et al.* [47], [49] introduced a factor analysis approach based on probabilistic matrix factorization, and used both user social network information and rating records to solve the data sparsity problem. Pham *et al.* [32] first performed a clustering algorithm on the social network of users, and then used the generated clusters as the neighborhoods for user-based recommendations. However, these methods can only be applied in scenarios where there exists a friendship or trust network between users.

In recent years, the recommendation techniques have been applied to online forums to predict thread participation. Inspired by the idea of Zipf's law, Fung *et al.* [50] proposed the pfidf score and used a weighted nonnegative matrix factorization to calculate similarity matrix for user–thread relationship. Zhao *et al.* [48] proposed to make use of the reply relationships among users and thread contents to learn a model of user–thread relationship in Digg.com. Castro-Herrera [51] proposed a hybrid recommendation system organizer and promoter of collaborative ideas (OPCI) in online forums. OPCI uses both content-based and collaborative-based methods. The content-based part recommends similar topics to a user with the content of the discussion threads, and the collaborative filtering part generates additional recommendations by identifying users with similar interests. Yang *et al.* [36] recommended threads in massive online open courses (MOOC) forums using an adaptive feature-based matrix factorization framework. They argued that one important property of MOOC thread recommendation is that each time a student logs into the forum, they are more likely to participate in recently posted threads, which makes the task different from traditional product recommendation. In order to address the problem, they defined a time window that moves along the course weeks, and only used the data during each time window to train the model. Tang and Yang [52] and Tang *et al.* [54] utilized topic detection of threads and identified user interest for personalized recommendation in social media. Instead of matching a thread with a user, a statistical model was developed to learn the topics in threads and the user interests. The recommendations were then made based on the model.

Although many approaches exist for recommendation, traditional content-based and collaborative filtering methods cannot be directly applied in thread recommendation in OHCs due to the lack of rating information, the short life span of threads, and the sparsity problem. In addition, OHCs are very different from traditional social media websites. There are usually explicit friendship networks or trust networks in the latter. In OHCs, social networking activities are mainly based on common health concerns rather than explicit friendship connections. The social ties are much weaker in OHCs [55], and this feature of OHCs makes it difficult to make use of user social connections to improve recommendation performance. In this paper, we introduce a novel approach for thread recommendation in OHCs. We first represent the social media data in OHCs as a heterogeneous healthcare information network, which contains much richer contextual information than homogeneous networks or bipartite graphs. Then, we use a supervised learning technique on the heterogeneous healthcare information network to predict user–thread participations.

III. Recommending Threads in Online Health Communities

A. Heterogeneous Healthcare Information Network

Most real-world data can be represented as a heterogeneous information network. An OHC is a typical example. Besides user nodes and item nodes considered in studies on bipartite graphs [25], [45], other important entities could also be represented as nodes in the network. Threads in OHCs contain textual content generated by health consumers, discussing medical conditions and treatments. However, it would be impractical to extract every keyword in the text as nodes. Entities representing medical concepts are what consumers care the most, including but not limited to diseases, drugs, and ADRs. Here, ADR is defined as "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product" [56]. In this paper, we extracted these three key entities, along with user nodes and thread nodes to construct the network.

Heterogeneous information network has been adopted in recommendation systems [57]. The heterogeneous information network captures the comprehensive information and rich semantics to enhance the recommendations. For example, Yu *et al.* [58] used user implicit feedback data and meta-path latent features to make global and personalized entity recommendation for movies. Yang *et al.* [59] proposed an SVM-rank based method in heterogeneous information network for scientific collaboration recommendation. Shi *et al.* [60] introduced meta-path-based similarity measure to evaluate the similarity of users or items and proposed matrix factorization-based framework for movie recommendation. However, none of the previous work has investigated recommendation in the heterogeneous healthcare information network.

In this paper, we define a heterogeneous information network as an undirected graph $G = (\mathcal{V}, E; T, R)$ with an entity type mapping: $\varphi : \mathcal{V} \rightarrow T$ and a link type mapping: $\emptyset : E \rightarrow R$. Vertex $v \in \mathcal{V}$ is an entity, and an edge $e = \langle v, u \rangle \in E$

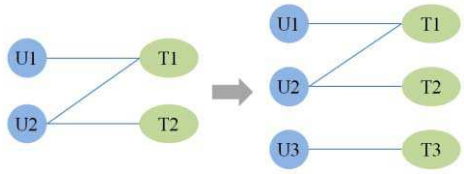

Fig. 1. User–thread bipartite graph.

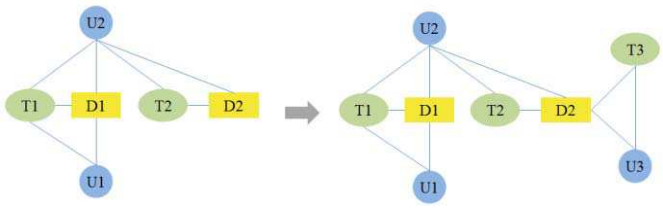

Fig. 2. Heterogeneous healthcare information network.

represents a relationship between v and u, where $v, u \in \mathcal{V}$. Type $t_i \in T = \{t_1, t_2, \ldots, t_n\}$ is an entity type, and $\varphi(v) \in T$. Relation $r_j \in R = \{r_1, r_2, \ldots, r_m\}$ is a type of relationship, and $\emptyset(e) \in R$. The number of the types of entities $|T| > 1$ and the number of types of relations $|R| > 1$. All vertexes $\mathcal{V} = \{V_1 \cup V_2 \ldots \cup V_n\}$ can be partitioned into n mutually exclusive subsets. All edges $E = \{E_1 \cup E_2 \ldots \cup E_m\}$ can be partitioned into m mutually exclusive subsets. In a weighted network, $w(e)$ stands for the weights of $e = \langle v, u \rangle \in E$.

A bipartite graph only contains two sets of nodes: user nodes and item nodes. As shown in Fig. 1, we have user nodes and thread nodes in a bipartite graph. If a new user U3 posted a new thread T3, it would be hard to predict the preference of U3 on other threads in this bipartite graph. However, if we add drug entities into the network (Fig. 2), we can see that U1 and U2 talked about a drug D1 in T1 and U2 discussed D2 in T2. When the new user U3 posted a new thread T3 and mentioned drug D2, we can connect U3 to the network through node D2. By constructing a heterogeneous healthcare information network from OHCs, we are able to retain the rich context information that can help us to address the sparsity problem.

B. Problem Formulation

Unlike the consumer products in electronic commerce, there are no explicit rating scores of threads given by users in OHC. However, if a user participated in the discussion of a thread, the user is showing some of his/her interest in the thread through discussing the effectiveness of drugs/treatments for a health condition/disease they are experience and/or other concerns. Normally, a user would have a range of interests, and he/she would choose a thread to participate based on the interests that are relevant to their health conditions. Therefore, the threads a user joined in the past can represent his/her interests to a great extent. If a new thread is very similar to the threads the user participated in, it could be a potential recommendation for the user. In addition, users with common interests would be very likely to join the same thread since they are interested in the same topics. Based on these observations,

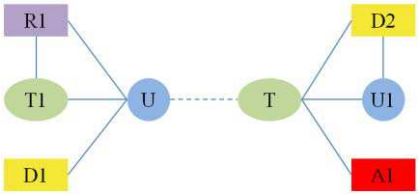

Fig. 3. Example of a user–thread pair.

we make three assumptions for thread recommendation in OHCs.

1) If a user posted comments at least once in a thread, he/she is interested in the thread.
2) If a thread is similar to those a user participated in previously, he/she is likely to be interested in the thread.
3) If most users who posted in a thread are similar to a user, he/she should be interested in joining the thread as well.

With the assumptions, we propose to train a supervised classification model for predicting users' participation in threads, using rich features extracted from constructed heterogeneous network. The problem is defined as follows:

Given a heterogeneous healthcare information network $G = (\mathcal{V}, E; T, R)$, let $U = \{u_1, u_2, \ldots, u_n\}, U \in \mathcal{V}$ be the collection of user nodes and $T = \{t_1, t_2, \ldots, t_m\}, T \in \mathcal{V}$ represent the collection of thread nodes. $D \in \mathcal{V}, R \in \mathcal{V}$, and $A \in \mathcal{V}$ stand for the collection of disease nodes, drug nodes, and ADR nodes. If a user u_i replied a thread t_j, then there is an edge between u_i and t_j. And, both u_i and t_j are connected to all the disease, drug, and ADR nodes appeared in the thread. The recommendation problem can be formulated as link prediction between a pair of unconnected user–thread nodes $\langle u, t \rangle$. In a heterogeneous healthcare information network, we label a connected $\langle u, t \rangle$ pair as a positive instance, and negative otherwise.

Fig. 3 demonstrates an example of a $\langle u, t \rangle$ pair in a heterogeneous healthcare information network. As we can see, currently the $\langle u, t \rangle$ pair is not connected to each other. However, there exist two positive instances $\langle u, t_1 \rangle$ and $\langle u_1, t \rangle$ in the network. In addition, both u and t are connected to other different types of nodes, and we can make use of the structural information to estimate the user's preference in the thread node t. In order to predict the likelihood of connection between $\langle u, t \rangle$, we propose to extract all positive and negative $\langle u, t \rangle$ pairs from the network, identify network-based features for each pair, and then use all the pairs to train a binary classification model for prediction.

C. Feature Extraction

We extract both node-based and path-based features for each $\langle u, t \rangle$ pair in the network. For node-based features, we calculate typical social network metrics to capture the characteristics of nodes in a network. For example, the node frequency implies the activeness of a node in the network, the degree centrality may suggest a node's popularity, and a node with high betweenness centrality would play a critical

role of bridging different groups in the network. These characteristics of a user node u and a thread node t could help us to capture some clues for predicting the probability of connection between the $\langle u, t\rangle$ pair. Path-based features consider the relationship between two different nodes. If two nodes are strongly associated, we may infer that they could have very similar neighbors and have activities in the same subnetwork. For example, two similar user nodes might connect to a lot of common thread nodes or common drug nodes, and two thread nodes may be linked to the same group of user nodes if they are very similar to each other. Therefore, we use the path-based features to measure the relationship between different nodes, and then predict the connection of $\langle u, t\rangle$ pairs based on the relationship.

1) Node-Based Features: We calculate node frequency, node degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality [61] for both the user node u and the thread node t in a $\langle u, t\rangle$ pair. Node frequency measures the number of times a node appears in a network. A user who posted in a great amount of threads would get a high node frequency, meaning he/she may have a wider range of interests than less active users. We could further infer that this user would be more likely to join a recommended thread than the others. Degree centrality is defined as the number of links connected to the node, which reflects the user or thread's popularity in a network. Popular threads usually tend to attract more attention. Betweenness centrality measures the frequency of a node falling on the shortest path connecting other pairs of nodes. It is a useful index that quantifies a node's potential power of bridging communication. Closeness centrality is calculated by summing the length of all the shortest paths between a node and all other nodes in a network. Eigenvector centrality is a natural extension of degree centrality, and it reflects the influence of a node. The underlying concept is a node that is important if it is linked to by other important nodes in a network. Specifically, a node is scored in a way that links to high-scoring nodes contribute more to the node's score than links to low-scoring nodes do. We estimate a user or thread's importance and influence in the network by collecting these metrics as node-based features.

2) Path-Based Features:

a) Thread–thread relationship: Based on the previous assumptions, users are inclined to participate in threads similar to what the users showed interests in before. If most of the threads a user has previously participated in are about depression, a thread talking about the effectiveness of Effexor (a drug indicated for depression) would be of more interest to the user than a thread discussing dental crown inflammation is. As a result, thread–thread relationship should be a critical indicator for users' preferences in a given thread.

Fig. 4 illustrates the role of thread–thread relationship in predicting user interests. The similarity between a given thread t and a user u's history approximates the possibility of u joining t. Hence, we extracted thread–thread relationship between each $\langle u, t\rangle$ pair as a feature.

Given a user–thread pair $\langle u, t\rangle$, let $P(u, t)$ denote the user u's interest in thread t, and $T^u = \{t_1^u, t_2^u, \ldots, t_N^u\}$ stand for the set of threads that user u has joined previously, we can then

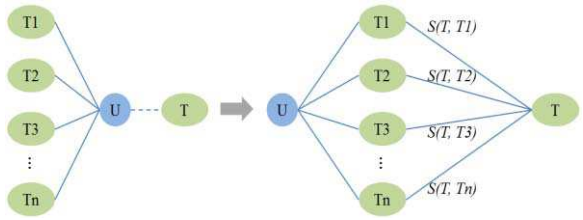

Fig. 4. Thread–thread relationship.

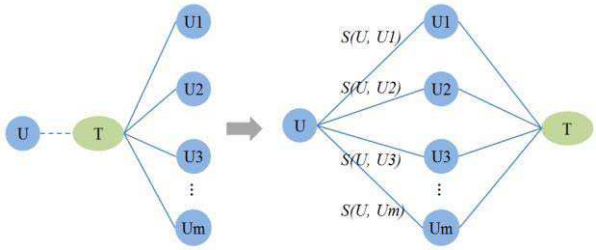

Fig. 5. User–user relationship.

get $P(u, t)$ by calculating the similarity between t and T^u. We use a simple averaging approach here for aggregating the similarity

$$P(u, t) = \frac{1}{N} \sum_{t_i^u \in T^u} s(t_i^u, t).$$

b) User–user relationship: Likewise, a user would very likely participate in a thread if most of the users who commented in the thread are similar to the user. Many studies took into account of user social connections to make more accurate recommendations [24], [47], [48]. There is not an explicit social network defining the friendship among users in an OHC. However, we can still measure the similarity between users in a heterogeneous healthcare information network [10]. Although such user–user relationship in a heterogeneous information network is not the same as the links in a friendship network, a higher degree of similarity between two users in a heterogeneous information network generally implies higher possibility that they share common interests. As illustrated in Fig. 5, for a $\langle u, t\rangle$ pair, we can predict the user's preference on the thread based on the similarity between the user and all users joined the target thread. So, we also extract user–user relationship as a feature for each $\langle u, t\rangle$ pair.

Given a user–thread pair $\langle u, t\rangle$, let $P(u, t)$ be the preference of user u in thread t, we will have a set of users $U^t = \{u_1^t, u_2^t, \ldots, u_M^t\}$ that joined thread t, and $P(u, t)$ can be calculated by averaging the similarity between u and all users in U^t

$$P(u, t) = \frac{1}{M} \sum_{u_i^t \in U^t} s(u_i^t, u).$$

3) Path-Based Feature Quantification: In terms of calculating similarity of user nodes and thread nodes, we utilize the path information in the network to extract the thread–thread and user–user relationship features. We propose three different approaches in computing the similarity between two nodes

in a heterogeneous network, namely, content similarity, local similarity (ProfileNet), and global similarity (HealthRank).

a) Content similarity: In an OHC, users participate in discussions on topics of interest. The messages authored by a user can best represent his/her interests, and the topic of a thread can also be captured by its content. Thus, the problem of node similarity can be transformed into the problem of content (text) similarity. The problem is formulated as follows.

Given a set of nodes of the same entity type (either user nodes or thread nodes in this paper) $V = \{v_1, v_2, \ldots, v_n\}$ and a collection of messages $M = \{m_1, m_2, \ldots, m_n\}$, node v_i can be represented by the messages m_i. If v_i is a user node, v_i is the author of message m_i. If v_i is a thread node, then the message m_i is the content of v_i. Each node v_i can be represented by a term vector $\overrightarrow{t_i} = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$, which are all terms in m_i. t_{ij} is the TF-IDF value of the term in M. In order to measure the similarity of two nodes v_i and v_j, we adopt the cosine similarity to calculate the similarity between vector $\overrightarrow{t_i}$ and $\overrightarrow{t_j}$

$$S(v_i, v_j) = \frac{\overrightarrow{t_i} \cdot \overrightarrow{t_j}}{\| \overrightarrow{t_i} \| \| \overrightarrow{t_j} \|}.$$

Since content similarity relies on the textual content, the content quality is crucial. 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.

b) Local similarity (ProfileNet): In order to overcome the above discussed shortcomings of content similarity, we propose a local similarity approach called ProfileNet utilizing nodes' structural information in the network.

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

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.

If two thread nodes have very similar ego-centered networks, it means that they both contain a lot of the same entities such as drugs, diseases, or ADRs. And, there is a high probability that the two threads are about very similar topics. Likewise, two user nodes that have similar ego-centered networks would very likely have common interests. Thus, we calculate the similarity of two nodes as follows:

$$S(v_i, v_j) = \sum_{i=1}^{n} \frac{\alpha_{t_i} P_{v_i}^d(\overrightarrow{w_{t_i}}) \cdot P_{v_j}^d(\overrightarrow{w_{t_i}})}{\| P_{v_i}^d(\overrightarrow{w_{t_i}}) \| \cdot \| P_{v_j}^d(\overrightarrow{w_{t_i}}) \|}$$

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 $P_{v_i}^d$ of a node v_i is defined as a vector of weights between v_i and its neighbor nodes within distance d

$$P_{v_i}^d = \{\overrightarrow{w_{t_1}}, \overrightarrow{w_{t_2}}, \ldots, \overrightarrow{w_{t_n}}\}$$

and

$$P_{v_i}^d(\overrightarrow{w_{t_i}}) = (w(e_{v_i u_1}), w(e_{v_i u_2}), \ldots, w(e_{v_i u_k}))$$

where $u_i \in \{u_1, u_2, \ldots, u_k\}$ are all the nodes of type t_i in the union of ego-centered networks of node v_i and v_j, which means the type of $u_i \varphi(u_i) = t_i$. $w(e_{v_i u_j})$ is the weight between node v_i and node u_j, and it is calculated by the following equation:

$$w(e_{v_i u_j}) = w(e_{v_i x_1}) \times w(e_{x_1 x_2}) \times \cdots \times w(e_{x_n u_j})$$

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

Since different types of links carry different semantic meanings, the profile is organized into several separate vectors based on the relationship types. For example, two users may have similar profiles because they both talked about the same drug, or because they are interested in the same disease. In another word, users can be similar in different ways. By taking into account the heterogeneity of the network, we could set different weights for different relationship types and provide more personalized recommendation.

c) Global similarity (HealthRank): ProfileNet measures similarity of two nodes from a local perspective, considering two given nodes' common neighbors within a distance in the network. But if two nodes do not share neighbors in certain distance, ProfileNet cannot calculate their similarity. For example, if one user talked about "Prozac" while another user talked about "Zoloft," the local similarity would consider them dissimilar in terms of drug. Nevertheless, even if they are not connected to the identical drug entity, there should be some degree of similarity between them because both "Prozac" and "Zoloft" are popular drugs for depression. To deal with this problem, we should estimate the similarity between nodes from a global point of view.

HealthRank is proposed based on a well-known global similarity algorithm SimRank [62]. The fundamental concept is that two objects are similar if they are referenced by similar objects. SimRank only takes in-links into account when computing similarity as it was proposed for a homogeneous directed graph. The situation is more complex in a heterogeneous information network. It makes no sense to measure the similarity of two nodes from different entity groups. For example, computing the similarity between a consumer and a disease is meaningless. The global similarity should be aggregated from the similarities between nodes of the same type. In addition, we consider an undirected network in this paper. Therefore, we propose HealthRank specifically for measuring global similarity in a heterogeneous healthcare information network.

Given two nodes v and u in an undirected heterogeneous information network, $N(v)$ and $N(u)$ are the sets of neighbors

TABLE I
EXTRACTED FEATURES

Type	Features
Node-based	Node frequency, degree centrality, betweenness centrality, closeness centrality, eigenvector centrality
Path-based	Thread-thread relationship (content/local/global)
	User-user relationship (content/local/global)

of v and u, respectively. Individual neighbors are represented by $N_i(v)$, where $1 \leq i \leq |N(v)|$. $\varphi(v)$ stands for the entity type of v. The similarity between two nodes is computed as follows:

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

$$s_{k+1}(v,u) = \begin{cases} 0, & (\text{if } \varphi(v) \neq \varphi(u)) \\ \dfrac{C}{|N(v)||N(u)|} \displaystyle\sum_{i=1}^{|N(v)|} \sum_{j=1}^{|N(u)|} s_k(N_i(v), N_j(u)), \\ \quad (\text{if } \varphi(v) = \varphi(u)). \end{cases}$$

Unlike ProfileNet, HealthRank considers the similarity of every 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.

In summary, we extract features based on nodes and paths of a heterogeneous information network. Table I presents the features we use in the thread recommendation learning model.

D. Thread Recommendation Learning Model

As mentioned above, we formulate thread recommendation as a binary classification problem. Each $\langle u, t \rangle$ pair was labeled as either positive or negative instance depending on whether they are already connected in the network. Then, we used the extracted features to build the model. In this paper, we employed logistic regression to minimize the following cost function:

$$J(\theta) = -\frac{1}{m}\left(\sum_{i=1}^{m}(y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)})))\right) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

where

- m the number of training example (pairs of user–thread node);
- n the number of features;
- $x^{(i)} \in \mathbb{R}^{n+1}$ an $n+1$ dimensional vector including a constant 1 and n features;
- $\theta \in \mathbb{R}^{n+1}$ an $n+1$ dimensional vector of parameters associated with constant 1 and each of the n features;

$h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ where $g(z) = (1/(1+e^{-z}))$ is called sigmoid function or logistic function and θ^T is the transpose of θ;

For each training example (a pair of user–thread nodes $\langle u^{(i)}, t^{(i)} \rangle$), $y^{(i)} = 1$ if the user joined the thread, and $y^{(i)} = 0$ otherwise;

$(\lambda/2m)\sum_{j=1}^{n}\theta_j^2$: regularization term for the purpose of preventing overfitting problem where λ is the regularization parameter.

IV. EXPERIMENT

A. Data Set and Network Construction

To validate the proposed approach, data set was collected from a popular OHC MedHelp.[1] Millions of health consumers have been using MedHelp for exchanging healthcare information and opinions with peer consumers. They post threads discussing a variety of healthcare concerns in the hundreds of medical support communities in MedHelp. Each of these communities is dedicated to a specific medical concern, such as heart disease, lung cancer, and diabetes. Our data set was collected from four active communities: heart disease, depression, breast cancer, and dental health.

First, we constructed a heterogeneous healthcare information network from the data set. As mentioned before, we consider five types of entities in our network, including consumers/users, threads, drugs, diseases, and ADRs. Consumers and threads were identified by their IDs from the data set. In terms of the other three types of entities, we employed dictionary-based approach for recognizing them from the data set. DrugBank[2] was used to build dictionary for drugs by collecting their generic names along with their brand names and brand mixture names. For diseases, all "Disease or Syndrome" concepts with lexical variants are collected from UMLS to build the dictionary. Finally, we used SIDER[3] to develop a dictionary for ADRs. Extracting ADRs is more challenging. Consumers describe ARDs in many different ways, because ADR is a reaction which could be a symptom or a sign and there are usually no formal names for ADRs. In addition, consumers use different vocabularies from professional terms [8], [63], [64]. For instance, the medical terminology for "hair loss" is "alopecia"; most consumers are not familiar with the latter and tend to use the former. So, we proposed to use the consumer health vocabulary (CHV) to handle this issue. CHV is a collection of expressions describing medical concepts likely to be recognized by most consumers [65]. The CHV terms are more likely to be used by consumers and can be used for expanding the ADR dictionary.

Multiple entities are usually mentioned in an online post, and relations can be drawn between them. We find out the relationship between entities using co-occurrence analysis. The basic idea is that the co-occurrence of two entities usually implies an underlying relationship between them. If entities were mentioned in the same message, there should exist relationships between them. The more frequent two entities appear together, the stronger the relationship. The relations extracted by co-occurrence analysis are undirected.

[1] http://www.medhelp.org/
[2] http://www.drugbank.ca/
[3] http://sideeffects.embl.de/

A thread in OHCs usually consists of several messages. Topic digression is often observed as the discussion unfolds, especially in long threads with tens of messages. Consumers keep bringing new concepts in the conversation and messages in a thread might be about totally different diseases or drugs. In this case, extracting relations using co-occurrence analysis with the whole thread as the analysis unit would create a large number of false connections. A message is more likely to be focused on a single topic, so the co-occurrence of entities in the same message can speak more for their underlying relationship. Therefore, we decided to extract entities and relations by considering message as an analysis unit in this paper. In terms of node frequency and link frequency, we consider multiple appearances of entities or relations in one message as only one occurrence.

In a heterogeneous information network, the weight schema of the edges is crucial for calculating node similarity. The most straightforward way is treating all edges equally important. Nevertheless, this is not the case for most real-world information network. Ignoring the difference among the messages carried by each edge could lead to information loss. Using two nodes' co-occurrence frequency as the edge weight is an alternative option. However, co-occurrence frequency could favor the nodes with high frequency over those with much lower frequency without normalization. Hence, we propose to use association measurements to estimate the association strength between the two endpoints of an edge. In our previous experiments [10], we found that leverage yields better results. Therefore, we use leverage as edge weight in this paper

$$\text{leverage} = \text{support}(e_{vu}) - \text{support}(v) \times \text{support}(u)$$

where

$$\text{support}(e_{vu}) = \frac{\text{freq}(e_{vu})}{\text{total count}}$$
$$\text{support}(v) = \frac{\text{freq}(v)}{\text{total count}}$$

where $\text{freq}(e_{vu})$ is the edge frequency, namely, the co-occurrence frequency of nodes v and u. $\text{freq}(v)$ denotes the frequency of node u, and total count is the total number of messages in the data set. $\text{support}(e_{vu})$ is the actual probability of co-occurrence of node v and node u in the data set. $\text{support}(v) \times \text{support}(u)$ is the probability of their co-occurrence if v and u are absolutely independent. Leverage measures the difference between the actual co-occurrence probability and the theoretical co-occurrence probability if the two nodes are independent.

B. Experimental Settings

The collected data set consists of 701 threads in total from January 1, 2010 to March 31, 2010, containing 3759 messages. The final constructed network is composed of 2439 nodes and 19971 edges. The 93018 pairs of user–thread nodes were discovered from the network, and there were 1437 positive instances versus 91581 negative instances. An instance of user–thread node is labeled as 1 (positive instance) if there exists a link between them, which means the user commented in the thread before. Otherwise, we label an instance as 0 (negative instance) if the user never participated in the thread before. We did not include instances in which the user initiated the thread, since it does not make much sense to recommend to a user his/her own thread. As the data set is extremely imbalanced, we utilized cost sensitive technique during the training process. Then, we used tenfold cross validation for evaluation. We conducted experiment in three groups as listed in Table II.

TABLE II
COMPARISON OF DIFFERENT MODELS

Group	Model	Included Features
T-T (Thread-Thread Relationship)	Baseline	Node-based features
	Baseline + Content	Node-based features + thread-thread content similarity
	Baseline + Local	Node-based features + thread-thread local similarity
	Baseline + Global	Node-based features + thread-thread global similarity
	Baseline + All Features	Node-based features + thread-thread similarity (content + local + global)
U-U (User-User Relationship)	Baseline	Node-based features
	Baseline + Content	Node-based features + user-user content similarity
	Baseline + Local	Node-based features + user-user local similarity
	Baseline + Global	Node-based features + user-user global similarity
	Baseline + All Features	Node-based features + user-user similarity (content + local + global)
T-T + U-U (Thread-Thread + User-User Relationship)	Baseline	Node-based features
	Baseline + Content	Node-based features + thread-thread + user-user content similarity
	Baseline + Local	Node-based features + thread-thread + user-user local similarity
	Baseline + Global	Node-based features + thread-thread + user-user global similarity
	Baseline + All Features	Node-based features + thread-thread + user-user similarity (content + local + global)

First, we combine basic node-based features with thread–thread relationship, and then compare the performance of different node similarity approaches. The second group combines node-based features with user–user relationship, and the last group mixes both thread–thread and user–user relationships for prediction. The three groups use the same baseline model, which is the model that only includes the node-based features. Then, we add path-based features from different similarity algorithms. Finally, we combine all features together.

C. Experimental Results and Discussion

Figs. 6–8 demonstrate the precision, recall, and F1 score, respectively. As we can see, baseline model performed the worst and got an F1 score of 0.2. And, we were able to improve the performance by combining thread–thread relationship and user–user relationship with the baseline model.

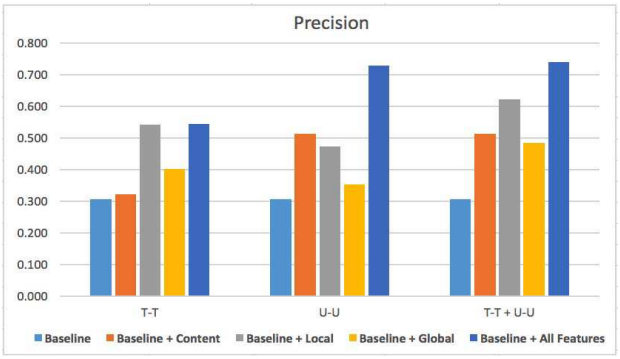

Fig. 6. Precision.

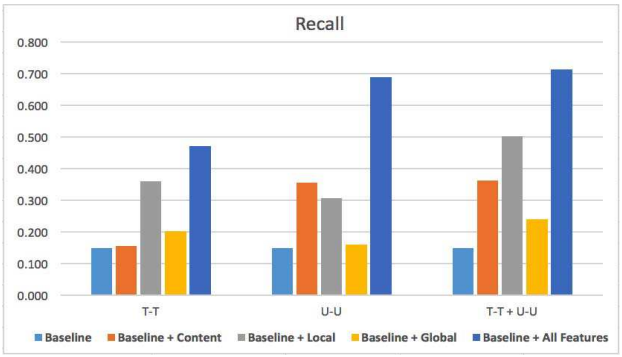

Fig. 7. Recall.

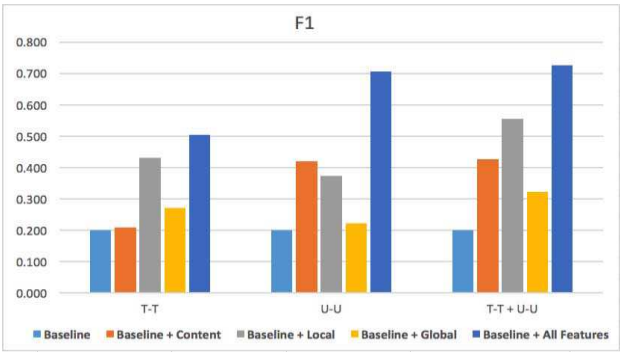

Fig. 8. F1 score.

Both thread–thread relationship and user–user relationship boosted the performance. However, thread–thread content similarity contributed less than user–user content similarity, while thread–thread structural similarity performed better than user–user structural similarity.

A user's interest usually evolves over time. For example, a user may be interested in the diagnosis of a disease at the beginning, then moved to the treatment or surgery for curing the disease, and finally shifted to discussion on recovery from the treatment. The content similarity may not be able to capture this evolvement and would consider a thread about recovery irrelevant to the user that mostly read treatment for the disease in the past. Under this circumstance, structural similarity abstracts users' interests onto concept level, such as disease, drugs, and ADRs, which enables us to reduce the dimensionality of the data, only focuses on the key concepts and makes it easier to capture the users' general interests. On the other hand, users who joined the same thread usually have very similar medical concerns, even similar evolving path in reading history. In this case, content similarity may be more effective in finding relevant threads for a user.

As for the two kinds of structural similarity, the local approach yields better results than the global approach in all scenarios. The local approach calculates similarity with nodes' ego-centered networks and considers two nodes' common neighbors. On the other hand, the global approach tries to measure two nodes' similarity from a global perspective by making use of the whole network's structural information. The results imply that global approach may bring in some noisy information when considering relationships in long distance, while local approaches can focus on users' main interests. As a result, the local approach has a better performance in predicting users' interests in threads.

The performance was greatly improved for both thread–thread and user–user group after we integrated all similarity approaches. One may speculate that each of the similarity approaches measures different dimensions of the relationship between two nodes. They behaved differently when employed separately, but they could complement each other and work better together, leading to a better performance.

We could further improve the results by integrating thread–thread and user–user relationships and achieved an F1 score of over 0.7. This observation agrees with our assumptions that users' connections and thread similarity have great influence on a user's preference. Thread–thread relationship employs users' reading history to make inference about users' preference in threads. User–user relationship represents the social aspect of OHCs and how this social factor impacts users' interests. Although there is not an explicit friendship network between users in OHCs, we can still estimate the similarity between users' interests by making use of the network information. These two types of relationship work well individually, but can reciprocally improve each other's predictive ability.

V. Conclusion

In order to assist health consumers in acquiring relevant information in OHCs, we investigated the thread recommendation problem in this paper. We found that structural information captured by a heterogeneous healthcare information network is valuable for predicting users' preferences in threads. The heterogeneous healthcare information network captures the rich information about the medical concepts and the health consumers involved in the online discussions. Such heterogeneous network not only represents the medical concepts that a health consumer is interested in, but also represents the relationships of these medical entities based on the discussions in OHCs. We captured basic network metrics, thread–thread relationship, and user–user relationship through extracting features from the heterogeneous healthcare information network. We utilized the extracted features to train a binary classification model for thread recommendation. Both structural (local and global) approaches proposed in this paper

achieved better performance than the content-based approach in capturing thread–thread relationship. Moreover, it was found that the local approach outperformed the global approach, which means that local similarity had a better predictive ability in terms of user preferences. We also demonstrated that considering both thread–thread relationship and user–user relationship could achieve better predictive performance than using either one individually.

One limitation of our work is that we used dictionary-based methods for constructing the heterogeneous healthcare information network. We can achieve high precision in this way by making sure that the constructed network is in high quality. However, we may miss some important information since we are dealing with social media data and health consumers use different languages from professional vocabularies. Although we used CHV to expand the vocabularies, there is still a huge gap between professional vocabularies and consumer language. In the future, we will explore effective approaches for information extraction from OHCs that will capture the new vocabularies used by health consumers and mapping with the professional ontologies. We shall also explore topic modeling [53], [66] to identify the user interest for supporting the recommendation that cannot be captured in keyword matching through CHV.

In addition, temporal factor was not considered in the prediction in this paper. Health consumers' interests may change over time as their health conditions may change at different stages. A thread that is of interest to a user a few months ago may be irrelevant to the current user interest. We have shortened the span of data set to eliminate the impact of temporal factor in this paper. In the future, we will analyze how users' preferences evolve over time and integrate the temporal attribute of threads into recommendation.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M. R. DiMatteo, "Social support and patient adherence to medical treatment: A meta-analysis," *Health Psychol.*, vol. 23, no. 2, p. 207, 2004.

[2] K. Chuang and C. C. Yang, "Social support in online healthcare social networking," presented at the iConf., Urbana, IL, USA, 2010.

[3] C. K. Coursaris and M. Liu, "An analysis of social support exchanges in online HIV/AIDS self-help groups," *Comput. Human Behav.*, vol. 25, no. 4, pp. 911–918, 2009.

[4] Y. Liu, Y. Chen, J. Tang, and H. Liu, "Context-aware experience extraction from online health forums," in *Proc. Int. Conf. Healthcare Inform. (ICHI)*, Oct. 2015, pp. 42–47.

[5] G. Eysenbach, "Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness," *J. Med. Internet Res.*, vol. 10, no. 3, p. e22, 2008.

[6] S. D. Hodges, K. J. Kiel, A. D. Kramer, D. Veach, and B. R. Villanueva, "Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy," *Personality Social Psychol. Bull.*, vol. 36, no. 3, pp. 398–409, 2010.

[7] G. Eysenbach, "The law of attrition," *J. Med. Internet Res.*, vol. 7, no. 1, p. e11, 2005.

[8] L. Jiang and C. C. Yang, "Using co-occurrence analysis to expand consumer health vocabularies from social media data," presented at the IEEE Int. Conf. Healthcare Inform., Philadelphia, PA, USA, Sep. 2013.

[9] Y. Zhang, P. Wang, A. Heaton, and H. Winkler, "Health information searching behavior in MedlinePlus and the impact of tasks," in *Proc. Proc. 2nd ACM SIGHIT Int. Health Inform. Symp.*, 2012, pp. 641–650.

[10] L. Jiang and C. C. Yang, "Determining user similarity in healthcare social media using content similarity and structural similarity," presented at the 15th Conf. Artif. Intell. Med. (AIME), Pavia, Italy, 2015.

[11] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[12] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*. Berlin, Germany: Springer, 2007, pp. 325–341.

[13] S. Debnath, N. Ganguly, and P. Mitra, "Feature weighting in content based recommendation system using social network analysis," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 1041–1042.

[14] I. Cantador, A. Bellogín, and D. Vallet, "Content-based recommendation in social tagging systems," in *Proc. 4th ACM Conf. Rec. Syst.*, 2010, pp. 237–240.

[15] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2011, pp. 73–105.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.

[17] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, Jan./Feb. 2003.

[18] R. Jin, J. Y. Chai, and L. Si, "An automatic weighting scheme for collaborative filtering," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2004, pp. 337–344.

[19] H. Ma, I. King, and M. R. Lyu, "Effective missing data prediction for collaborative filtering," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 39–46.

[20] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *Proc. AAAI/IAAI*, 2002, pp. 187–192.

[21] G. Katz, N. Ofek, B. Shapira, L. Rokach, and G. Shani, "Using Wikipedia to boost collaborative filtering techniques," in *Proc. 5th ACM Conf. Rec. Syst.*, 2011, pp. 285–288.

[22] J. Han, "Mining heterogeneous information networks by exploring the power of links," in *Proc. Discovery Sci.*, 2009, pp. 13–30.

[23] Y. Sun and J. Han, "Mining heterogeneous information networks: Principles and methodologies," *Synth. Lectures Data Mining Knowl. Discovery*, vol. 3, no. 2, pp. 1–159, 2012.

[24] P. Massa and P. Avesani, "Trust-aware recommender systems," in *Proc. ACM Conf. Rec. Syst.*, 2007, pp. 17–24.

[25] J. Grujić, "Movies recommendation networks as bipartite graphs," in *Computational Science—ICCS*. Berlin, Germany: Springer, 2008, pp. 576–583.

[26] J.-G. Liu, T. Zhou, B.-H. Wang, Y.-C. Zhang, and Q. Guo, "Degree correlation of bipartite network on personalized recommendation," *Int. J. Modern Phys. C*, vol. 21, no. 1, pp. 137–147, 2010.

[27] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, vol. 463. New York, NY, USA: ACM, 1999.

[28] G. Salton, *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*. Reading, MA, USA: Addison-Wesley, 1989.

[29] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?" *Commun. ACM*, vol. 35, no. 12, pp. 29–38, 1992.

[30] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, Aug. 2009, Art. no. 421425. [Online]. Available: https://www.hindawi.com/journals/aai/2009/421425/abs/

[31] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender systems for large-scale E-commerce: Scalable neighborhood formation using clustering," in *Proc. 5th Int. Conf. Comput. Inf. Technol.*, 2002, pp. 1–6.

[32] M. C. Pham, Y. Cao, R. Klamma, and M. Jarke, "A clustering approach for collaborative filtering recommendation using social network analysis," *J. Univ. Comput. Sci.*, vol. 17, no. 4, pp. 583–604, 2011.

[33] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[34] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 713–719.

[35] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009. [Online]. Available: https://www.computer.org/csdl/mags/co/2009/08/mco2009080030-abs.html

[36] D. Yang, M. Piergallini, I. Howley, and C. Rose, "Forum thread recommendation for massive open online courses," in *Proc. 7th Int. Conf. Edu. Data Mining*, 2014, pp. 1–4.

[37] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 4, p. 046115, 2007.

[38] Z. Huang, D. D. Zeng, and H.-C. Chen, "A comparison of collaborative-filtering recommendation algorithms for e-commerce," *IEEE Intell. Syst.*, vol. 22, no. 5, pp. 68–78, Sep. 2007. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4338497

[39] Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 116–142, 2004.

[40] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, and F. Yang, "Recommendation model based on opinion diffusion," *Europhys. Lett.*, vol. 80, no. 6, p. 68003, 2007.

[41] J.-G. Liu, B.-H. Wang, and Q. Guo, "Improved collaborative filtering algorithm via information transformation," *Int. J. Modern Phys. C*, vol. 20, no. 2, pp. 285–293, 2009.

[42] Z.-K. Zhang, T. Zhou, and Y.-C. Zhang, "Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs," *Phys. A, Statist. Mech. Appl.*, vol. 389, no. 1, pp. 179–186, 2010.

[43] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proc. 5th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2005, pp. 141–142.

[44] P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao, "A graph based approach to extract a neighborhood customer community for collaborative filtering," in *Databases in Networked Information Systems*. Berlin, Germany: Springer, 2002, pp. 188–200.

[45] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decision Support Syst.*, vol. 54, no. 2, pp. 880–890, 2013.

[46] C.-N. Ziegler and J. Golbeck, "Investigating interactions of trust and interest similarity," *Decision Support Syst.*, vol. 43, no. 2, pp. 460–475, 2007.

[47] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 203–210.

[48] J. Zhao, J. Bu, C. L. Chen, Z. Guan, C. Wang, and C. Zhang, "Learning a user-thread alignment manifold for thread recommendation in online forum," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 559–568.

[49] H. Ma, H. Yang, M. R. Lyu, and I. King, "SoRec: Social recommendation using probabilistic matrix factorization," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 931–940.

[50] Y.-H. Fung, C.-H. Li, and W. K. Cheung, "Online discussion participation prediction using non-negative matrix factorization," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.-Workshops*, Nov. 2007, pp. 284–287.

[51] C. Castro-Herrera, "A hybrid recommender system for finding relevant users in open source forums," in *Proc. 3rd Int. Workshop Manag. Requirements Knowl. (MARK)*, Sep. 2010, pp. 41–50.

[52] X. Tang and C. C. Yang, "TUT: A statistical model for detecting trends, topics and user interests in social media," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 972–981.

[53] X. Tang, M. Zhang, and C. C. Yang, "User interest and topic detection for personalized recommendation," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, vol. 1, Dec. 2012, pp. 442–446.

[54] X. Tang, M. Zhang, and C. C. Yang, "Leveraging user interest to improve thread recommendation in online forum," in *Proc. Int. Conf. Social Intell. Technol. (SOCIETY)*, May 2013, pp. 11–19.

[55] K. B. Wright, S. B. Bell, and K. B. Wright, "Health-related support groups on the Internet: Linking empirical findings to social support and computer-mediated communication theory," *J. Health Psychol.*, vol. 8, no. 1, pp. 39–54, 2003.

[56] I. R. Edwards and J. K. Aronson, "Adverse drug reactions: Definitions, diagnosis, and management," *Lancet*, vol. 356, no. 9237, pp. 1255–1259, 2000.

[57] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, Jan. 2017.

[58] X. Yu *et al.*, "Personalized entity recommendation: A heterogeneous information network approach," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 283–292.

[59] C. Yang, J. Sun, J. Ma, S. Zhang, G. Wang, and Z. Hua, "Scientific collaborator recommendation in heterogeneous bibliographic networks," in *Proc. 48th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2015, pp. 552–561.

[60] C. Shi, J. Liu, F. Zhuang, P. S. Yu, and B. Wu, "Integrating heterogeneous information via flexible regularization framework for recommendation," *Knowl. Inf. Syst.*, vol. 49, no. 3, pp. 835–859, 2016.

[61] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[62] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 538–543.

[63] L. Jiang, C. C. Yang, and J. Li, "Discovering consumer health expressions from consumer-contributed content," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2013, pp. 164–174.

[64] L. Jiang and C. C. Yang, "Expanding consumer health vocabularies by learning consumer health expressions from online health social media," in *Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 2015, pp. 314–320.

[65] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 1, pp. 24–29, 2006.

[66] Z. Hai, K. Chang, J. J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 623–634, Mar. 2014.

Christopher C. Yang is a Professor with the College of Computing and Informatics and holds a courtesy appointment with the School of Biomedical Engineering, Science, and Health Systems, Drexel University, Philadelphia, PA, USA. He is also the Director of the Healthcare Informatics Research Lab, the Director of the Data Science Programs, and the Program Director of MS in Health Informatics. He has authored over 300 referred journal and conference papers in the *ACM Transactions on Intelligent Systems and Technology*, the *ACM Transactions on Management Information Systems,* the IEEE TRANSACTIONS ON KNOWLEDGE AND DATE ENGINEERING, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. His current research interests include healthcare informatics, healthcare data analytics, security informatics, social intelligence and technology, Web search and mining, knowledge management, and information visualization.

Dr. Yang is currently serving as the Editor-in-Chief of the *Journal of Healthcare Informatics Research* (Springer) and the Co-Editor of *Electronic Commerce Research and Applications* (Elsevier). He is also the Steering Committee Chair of the IEEE International Conference on Healthcare Informatics and an Editor of the CRC book series on healthcare informatics.

Ling Jiang received the Ph.D. degree from Drexel University, Philadelphia, PA, USA, in 2016.

She is currently a Senior Data Scientist with The Washington Post, Washington, DC, USA. Her current research interests include enhancing automated journalism using natural language processing and machine learning techniques.