Content Recommendation by Analyzing User Behavior in Online Health Communities

Research-in-Progress

Hangzhou Yang **Zhijun Yan**

Abstract

Online health communities (OHCs) are the platforms for patients and their care-givers to search and share health-related information, and have attracted a vast amount of users in recent years. However, health consumers are easily overwhelmed by the overloaded information in OHCs, which makes it inefficient for users to find contents of their interest. This study proposes a framework for content recommendation by analyzing user activities in OHCs that utilizes social network analysis and text mining technology. We model users' activities by constructing user behavior networks that capture implicit interactions of users, based on which closely related users are detected and user similarities are calculated. Text analysis are performed using topic model to select the threads for final content recommendation. Based on the data collected from a famous Chinese OHCs, we expect that our model could achieve promising results.

Keywords: Content recommendation, online health communities, user behavior, social network analysis

Introduction

OHCs have seen a substantial development over the last decades. Health consumers participates in various discussions to exchange health-related information with peers (Yan et al. 2016; Yang and Gao 2018). Social support is generated during social networking with each other, which has been found essential for patients to cope with stressful health conditions. Research has shown that social support contributes to health outcomes by enhancing patients adherence to medical treatment (DiMatteo 2004). Informational support help patients reduce uncertainty by providing information and knowledge, such as experiences, advices and opinions (Kagashe et al. 2017; Yan et al. 2016). Emotional support help patients to maintain a positive attitude towards healthcare issues and obtain better health outcomes (Zhang et al. 2013). Patients can discuss concerned topics with people who have similar conditions and know what their peers are experiencing in OHCs (Jiang and Yang 2017).

However, users can be easily overwhelmed by the ocean of information as tremendous contents are generated continuously in OHCs, and it is tough for users to find contents of interest efficiently. Most patients cannot fully understand their health conditions due to the lack of medical knowledge. In addition, patients usually use very different expression style from healthcare professional and cannot precisely express their health issues, which lead them to poor information searcher. Furthermore, new threads posted by users can be easily swamped in discussion board. Consequently, an effective content recommendation system is desired to find the topics that are of interest to users in OHCs.

Existing recommendation systems usually focus on explicit relationship between entities in social media websites such as Facebook and Twitter, where social ties are basically built based on real-world connections. In OHCs, however, users contact with each other based on common interests or concerns

rather than long-term relationships (Jiang and Yang 2017). Therefore, new topics that are potentially of interest to users can be detected by finding the users that share similar behaviors in OHCs. In addition, most social network-based recommendation approaches ignore the usage of contents information in social media, which depict users' interest and can be utilized to measure user similarity.

Based on the above observations, this study proposes a framework of content recommendation in OHCs using social network analysis and text mining. Implicit user behavior networks are constructed according to user activities traits, and users who share common interest can be detected based on the networks. Topic model is utilized at last to find the contents for final recommendation. According to homogeneity theory, users share similar interest tend to connect to each other and closely connected users tend to share interest. A thread for recommendation is the one that is not currently engaged by a focal user, but is participated by other similar users. In addition, users are more likely to interest in the threads that are similar in terms of topics of their discussion history in OHCs.

Related Works

Extensive studies have been performed to recommend personalized contents for users in recent years. We provide a short literature review here due to the limit of paper length. Content-based approaches (Cami et al. 2019; Lash and Zhao 2016) recommend articles or books that are similar to items previously preferred by a specific user. Methods such as information retrieval and machine learning technologies can be used to analyze the contents of items and generate recommendations. Collaborative filtering-based systems (Geuens et al. 2018; Sahoo et al. 2012) recommend items for users based on opinions of other people who share similar interest . Collaborative filtering-based systems can be divided into user-based approach that recommend items liked by similar users, and item-based approach that recommend items that are similar to those previously preferred by them. Due to inability of finding sufficient similar neighbors in sparse dataset, social relationships are emerging as another improvement facet for recommendation systems (Lu et al. 2015). Recommendation systems that based on social network (Crespo et al. 2011; Fang and Hu 2018; Martens et al. 2016; Zhang et al. 2016) make recommendation by analyzing user connections with each other. Context aware (Binucci et al. 2017; Panniello et al. 2016) approaches incorporate context information that can be used to characterize the situation of a target to enhance recommendation in certain circumstances. Computational intelligence-based approaches (Ghoshal et al. 2015; Guo et al. 2018) utilize technologies such as association rules mining, deep learning and Bayesian technologies to construct recommendation models. Hybrid recommendation systems (Li et al. 2005; Xu et al. 2017) that combine multiple technologies could achieve better performance by overcoming the shortages of traditional approaches. Content recommendation is of great value to both online health consumers and OHCs yet remain not fully investigated. As discussed, both text contents and user network are potentially critical in thread recommendation. Content-based methods make recommend based on text analysis but ignore the user network knowledge, while collaborative filtering-based systems fail to make full use of text information. In sight of this, this paper integrates social network analysis and content analysis technologies to recommend items of interest to users in the context of online communities.

Methodology

The designed framework for content recommendation in OHCs contains four steps: implicit network generation and normalization, user communities' detection, user similarity analysis, and content topic analysis, as shown in Figure 1. At first, we build and normalize two implicit networks, namely undirected user behavior network and directed user behavior network to model users' activities based on a training dataset collected from an online health community. Second, hierarchical user communities are detected based on the undirected user behavior network by recursively performing a modularity based community detection algorithm to find closely connected users. Meanwhile, the similarities between the users in the directed user behavior network are calculated using an adapted SimRank algorithm. The threads that are posted by similar users within the same sub-community are selected as candidate contents for recommendation. Lastly, content analysis is performed using topic model to find the most related threads with contents that are posted by the focal user as final threads recommendation. The performance of proposed model is then evaluated based on a test dataset.

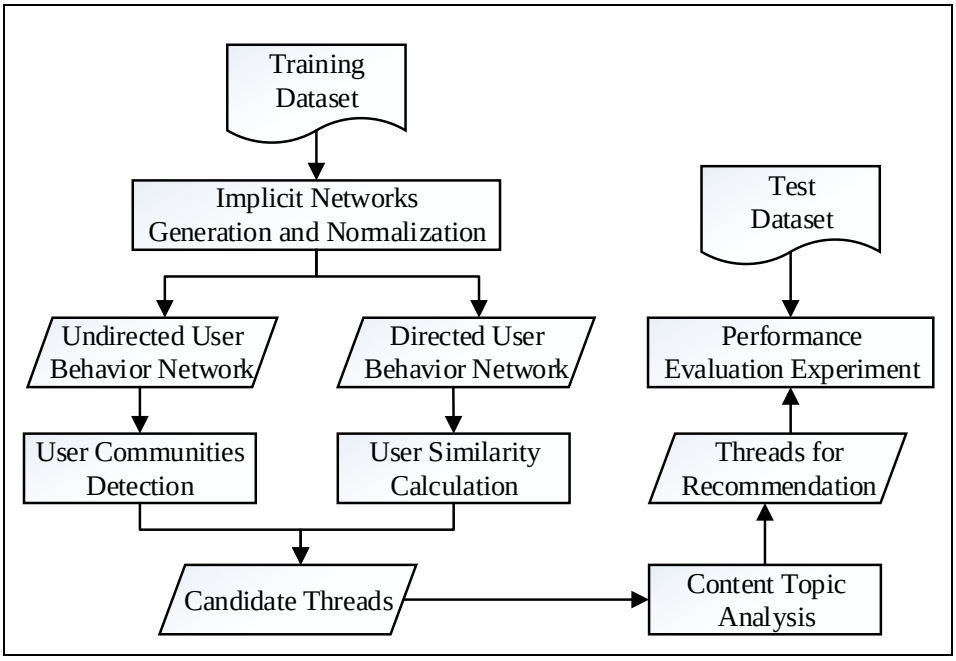

Figure 1. The Framework of Proposed Content Recommendation System

Implicit Networks Generation and Normalization

In this section, two implicit networks, namely undirected user behavior network and directed user behavior network are constructed to depict the characters of users' activities in OHCs. The detail of network generation and normalization are described as follows. In OHCs, users exchange health-related information by posting and replying to each other in discussion threads. The undirected implicit user behavior network (denoted as ..) is built based on the idea that two users are considered to share similar interest if they join in the discussion of the same threads, and are more similar if more such connections exist. Consider a simple example shown in Table 1, both user A and user B joined the discussion of thread T1, so A and B are connected in B, as shown in Figure 2(a).

Table 1. An Example of Online Discussion Dataset

Threads	Users	NU
T1	A,B	2
T2	A,D	2
T3	A,B	2
T4	B,C,D	3

*NU: number of users in each discussion thread.

Three aspects are considered to assign the weight of links in B:

- The more threads that two users share, the higher the weight of the link between them.
- The more popular a thread, the lower the weight it contributes.
- The more threads that a user has discussed, the lower the weight of the links that connect the user should be.

In OHCs, users tend to enroll in the discussions that of their interest, so the more threads that two users get involved simultaneously, the more similar interest they share. For the second aspect, if a thread is quite popular and attracts wide discussions among users in OHCs, the thread usually provides little information for interest similarity calculation between users. In this study, we simply use the inverse of

the number of participants: $1/NU$ to weight a thread that two users share. Based on the first two aspects, we weight the link between two users in B as:

$$w_{ij} = \sum_{t \in threads(u_i) \cap threads(u_j)} \frac{1}{NU_t}, \tag{1}$$

where w_{ij} is the weight of the edge between users u_i and u_j in B, $threads(v_i)$ and $threads(v_j)$ represent the thread sets that users u_i and u_j have participated in respectively, and NU_t is the number of participants in the thread t.

Users normally have various aspect of interest and likely to join in many discussions, so these users share more common threads with other users in the network, compared with those users that are less active in OHCs. In this way, active users could obtain much larger edge weights than inactive users in the built weighted network, and these higher weighted edges may dominate the analyses of the network. Therefore, the third aspect of assigning edge weight is proposed. We use the number of threads that a user has participated: f_i as the node weight of the user users u_i, as shown in Figure 2(a). In the study, we normalize the network B to B_n in two steps (Zhang et al. 2016). We first normalize the edge weight w_{ij} between two users in B as:

$$w_{ij}' = \frac{w_{ij}}{f_i * f_j}, \tag{2}$$

where f_i and f_j are the number of threads that users u_i and u_j have participated in respectively.

Afterwards, all w_{ij}' are normalized as:

$$w_{ij}'' = \frac{w_{ij}'}{\max_{\forall(i,j)}\{w_{ij}'\}}, \tag{3}$$

The above two steps of normalization are shown in Figure 2(b) and 2(c), respectively.

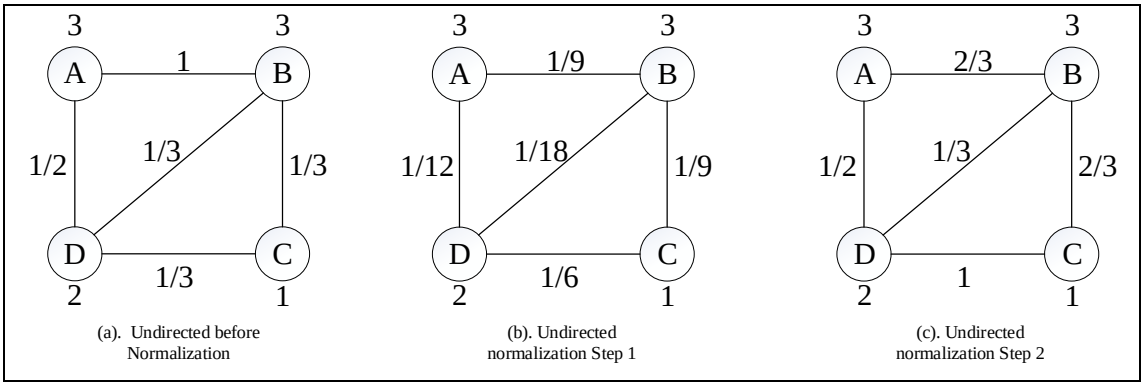

Figure 2. The Normalization of Undirected User Behavior Network

The obtained normalized user behavior network B_n can successfully capture the relative strength between users. However, B_n fails to distinguish the influence from the perspective of each user, which means the affinity of user u_i to user u_j could be different. To address this problem, we define an directed implicit user behavior network (denoted as $\vec{B}$). The weight of both edges between two users in $\vec{B}$ are equal before normalization, which are the same as in B, namely $w_{d(i,j)} = w_{d(j,i)} = w_{ij}$, as shown in Figure 3(b). To capture the local information of a user, we normalize the directed network $\vec{B}$ to $\vec{B}_n$ as follows:

$$w_{d(i,j)}' = \frac{w_{d(i,j)}}{f_i}, \tag{4}$$

where f_i is the number of threads that user u_i has participated in. The generated normalized directed network is shown in Figure 3(c).

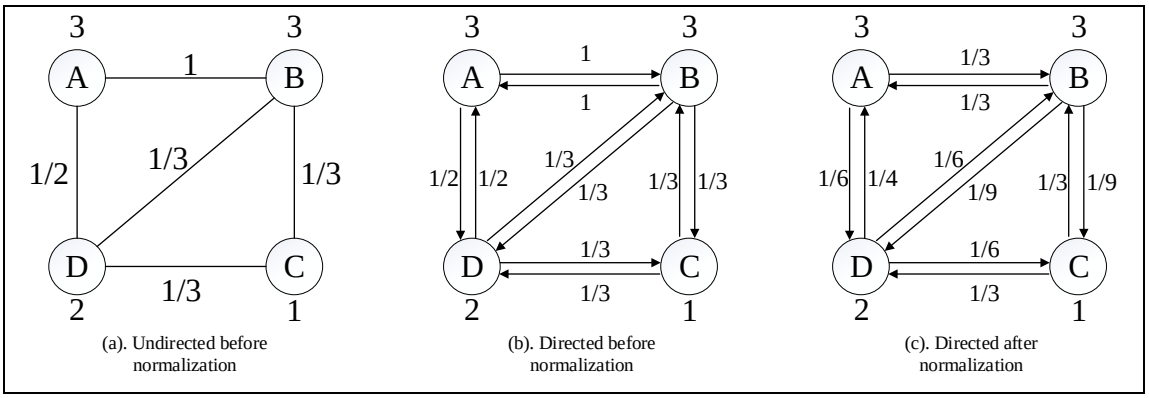

Figure 3. The Normalization of Directed User Behavior Network

User Communities Detection

In this section, we detect user communities based on the constructed undirected user behavior network B_n. The detected communities are considered as closely connected users who potentially share similar activity modes and have similar interest in OHCs. Consider a thread t that is engaged in by user u_i but is not engaged in by user u_j, we assume that user u_j is more likely interested in the thread t if users u_i and u_j share similar interest. Therefore, to find the threads that are potentially of interest for users who are not engaged in these threads but are engaged in by other close ones, we need to detect the users that are closely related to each other. The built user behavior networks can capture the interest relationship among users in OHCs by incorporating both thread information and user activities. Users that show similar interest are expected to be connected closely in the user behavior network, and community detection technologies can be utilized to detect users that are closely connected to each other. Hierarchical community detection is used for the normalized undirected implicit user behavior network B_n to detect user communities, the approach is based on the well-known community detection algorithm named modularity maximization (Newman 2006). Modularity (Q) measures the quality of a partition in a network, and higher modularity values indicate better partition results. The calculation of modularity is as follows:

$$Q = \frac{1}{2m} * \sum_{ij} \left[A_{ij} - \frac{k_i * k_j}{2m} \right] \delta\left(C_i, C_j\right). \tag{5}$$

where m represent the number of edges in a network; A_{ij} is an element of the adjacency matrix of network and denotes the edge weight between nodes u_i and u_j; $\delta\left(C_i, C_j\right)$ indicates whether nodes u_i and u_j are in the same community; k_i and k_j denote the node degree of nodes u_i and u_j respectively; and

$$A_{ij} = \begin{cases} w_{ij}'' & \text{if node } u_j \text{ connects to node } u_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

$$\delta\left(C_i, C_j\right) = \begin{cases} 1 & \text{if nodes } u_i \text{ and } u_j \text{ belong to the same community}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

$$k_i = \sum_{u_h \in B_n, h \neq i} w_{hi}^{"}. \tag{8}$$

The community detection algorithm is run recursively to divide B_n and further divide the detected big communities into smaller ones, and it stops when the size of all detected communities are smaller than 50.

User Similarity Calculation

After detecting the peers that behave closely to a user, we need to select the most similar users that are related to each user. This section calculates the exact behavior similarity among users in OHCs based on the constructed directed user behavior network. We adapt the SimRank algorithms (Jeh and Widom 2002) to measure the similarities among the users in the user behavior network $\vec{B}_n$. SimRank is an algorithm that can measure the similarity between each pair of nodes in a network based on the intuition that two objects are similar if they are related to similar objects (Jeh and Widom 2002). We calculate the similarity of two nodes u and v in the network $\vec{B}_n$ as follows:

$$s_0(u,v) = \begin{cases} 0 & if\ u \neq v, \\ 1 & if\ u = v. \end{cases} \tag{9}$$

$$s_{k+1}(u,v) = \frac{c}{|I(u)||I(v)|} \sum_i^{|I(u)|} \sum_j^{|I(v)|} s_k(I_i(u), I_j(v)) * w_{d(I_i(u),u)}^{'} * w_{d(I_j(v),v)}^{'}. \tag{10}$$

where c is a constant between 0 and 1; $I(u)$ and $I(v)$ represents the set of in-neighbors of nodes u and v, and $s(u,v) = 0$ when $I(u) = \varnothing$ or $I(v) = \varnothing$; $w_{d(I_i(u),u)}^{'}$ and $w_{d(I_j(v),v)}^{'}$ are the edge weight in $\vec{B}_n$ from one of the in-neighbors to the nodes u and v respectively. We run the adapted SimRank algorithm recursively to calculates the similarities between nodes by updating the similarities values of each pair of nodes in the network. The top n most similar users who are in the same sub-community with a focal user are selected as the final close users.

Content Topic Analysis

All the threads that the close users of a focal user have involved but not involved by the focal user are considered as the candidate contents for recommendation. This section chooses the threads for final recommendation from the candidate contents pool. To find the contents that are of interest to a focal user, we use a topic model based approach to detect the threads that are most similar to the contents posted by the focal user for final selection. This study uses Latent Dirichlet Allocation (LDA) (Blei et al. 2003). LDA is a generative statistical model that posits each document a mixture of various topics, and assumes that each word's presence is attributable to one of the document's topics. The contents of text i can be represented by a topic distribution $T_i = \{T_{i,1}, T_{i,2}, ..., T_{i,K}\}$ where $T_{i,k}$ is the weight of the k^{th} topic and $\sum_{k=1}^{K} T_{i,k} = 1$. We define the similarity between two collections of text using normalized cosine similarity as:

$$sim(t_i, t_j) = \frac{T_i \bullet T_j}{2\|T_i\|\|T_j\|} + \frac{1}{2} = \frac{\sum_{k=1}^{K} T_{i,k} T_{j,k}}{2\sqrt{\sum_{k=1}^{K} (T_{i,k})^2}\sqrt{\sum_{k=1}^{K} (T_{j,k})^2}} + \frac{1}{2} \tag{11}$$

The resulting similarity values range between 0 and 1, where a bigger value indicates higher similarity between the pair of text. We choose the top m most similar threads as the final recommendation contents to the focal user.

Experiment

The dataset was obtained from "TMJY" (bbs.tnbz.com), a famous Chinese online health community for diabetes. We crawled threads that are posted between July, 2016 and December, 2016 in the discussion board "Type 2 diabetes" at the end of May, 2018 using a python program. The collected thread information includes thread id, poster id, post time, title, message id, message time and post content. In addition, user information such as user id, gender, birth date and friends list were also collected. Finally, 27, 786 threads which contain 452, 988 messages that are generated by 6, 035 users were obtained. We use message as the basic unit in experiment, and the collected dataset was randomly split into two parts, with 90 percent as the training dataset and the rest 10 percent as the test dataset. The process was repeated 10 times, and each time with a different random training and test sets. The average of the 10 results generated from each dataset are regarded as the final performance, and Precision, Recall and F1-Measure are used as evaluation metrics. To evaluate the performance of the proposed model, four recommendation methods are selected as benchmarks. User-based and item-based collaborative filtering methods are selected due to their wide usage in recommendation systems. We take advantage of content-based method as the aim of this work is to recommend threads. Random selection method is incorporated to evaluate the exact performance of each method.

Conclusion

Online communities provide consumers ideal spots for health information exchange. Informational and emotional support are generated during the discussions among peers which facilitate consumers in various ways. However, the huge volume of user generated contents in OHCs make it threatening for consumers to find the contents they need efficiently. Users' behavior implicitly depicts their concerns and can be used for interest modeling. This study proposes a framework to recommend content for users in OHCs based on their previous activities traits. The model utilizes social network analysis and text mining to find the threads that are potentially of interest to user. A dataset is collected from an online health community to evaluate the proposed framework. We expect that our work contributes to the literature in the following aspects. First, the study designs a novel framework of content recommendation for users in online health websites from the perspective of user behavior analysis. In addition, user generated contents are analyzed to enhance final recommendation. Second, implicit user behavior networks are constructed based on user activities in online discussion boards in this work, which help to analyze user activities and find user communities that share similar interest. It provides researchers another perspective to find out users' attention points. The potential limitations of this work are as follows. First, this study captures users' activities in online health forums when they post or reply in thread, but fails to catch other types of activities such as browsing, which could result in inappropriate recommendation. Second, the proposed recommendation framework could be unfavorable to new users or users who are inactive in websites as little activity information can be captured.

Acknowledgements

This work was supported by National Key Research & Development Plan of China (Grant No: 2017YFB1400101), National Natural Science Foundation of China (Grant No: 71572013, 71872013, 71432002), and Beijing Municipal Social Science Foundation (Grant No: 18JDGLB040).

References

Binucci, C., De Luca, F., Di Giacomo, E., Liotta, G., and Montecchiani, F. 2017. "Designing the Content Analyzer of a Travel Recommender System," *Expert Systems with Applications* (87), pp. 199-208.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of machine Learning research* (3:Jan), pp. 993-1022.

Cami, B. R., Hassanpour, H., and Mashayekhi, H. 2019. "User Preferences Modeling Using Dirichlet Process Mixture Model for a Content-Based Recommender System," *Knowledge-Based Systems* (163), pp. 644-655.

Crespo, R. G., Mart ńez, O. S., Lovelle, J. M. C., Garc á-Bustelo, B. C. P., Gayo, J. E. L., and De Pablos, P. O. 2011. "Recommendation System Based on User Interaction Data Applied to Intelligent Electronic Books," *Computers in Human Behavior* (27:4), pp. 1445-1449.

DiMatteo, M. R. 2004. "Social Support and Patient Adherence to Medical Treatment: A Meta-Analysis," *Health psychology* (23:2), p. 207.

Fang, X., and Hu, P. J. H. 2018. "Top Persuader Prediction for Social Networks," *MIS Quarterly* (42:1), pp. 63-82.

Geuens, S., Coussement, K., and De Bock, K. W. 2018. "A Framework for Configuring Collaborative Filtering-Based Recommendations Derived from Purchase Data," *European Journal of Operational Research* (265:1), pp. 208-218.

Ghoshal, A., Menon, S., and Sarkar, S. 2015. "Recommendations Using Information from Multiple Association Rules: A Probabilistic Approach," *Information Systems Research* (26:3), pp. 532-551.

Guo, J., Zhang, W., Fan, W., and Li, W. 2018. "Combining Geographical and Social Influences with Deep Learning for Personalized Point-of-Interest Recommendation," *Journal of Management Information Systems* (35:4), pp. 1121-1153.

Jeh, G., and Widom, J. 2002. "Simrank: A Measure of Structural-Context Similarity," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 538-543.

Jiang, L., and Yang, C. C. 2017. "User Recommendation in Healthcare Social Media by Assessing User Similarity in Heterogeneous Network," *Artificial intelligence in medicine* (81), pp. 63-77.

Kagashe, I., Yan, Z., and Suheryani, I. 2017. "Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data," *Journal of medical Internet research* (19:9).

Lash, M. T., and Zhao, K. 2016. "Early Predictions of Movie Success: The Who, What, and When of Profitability," *Journal of Management Information Systems* (33:3), pp. 874-903.

Li, Y., Lu, L., and Xuefeng, L. 2005. "A Hybrid Collaborative Filtering Method for Multiple-Interests and Multiple-Content Recommendation in E-Commerce," *Expert systems with applications* (28:1), pp. 67-77.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. 2015. "Recommender System Application Developments: A Survey," *Decision Support Systems* (74), pp. 12-32.

Martens, D., Provost, F., Clark, J., and de Fortuny, E. J. 2016. "Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics," *MIS quarterly* (40:4).

Newman, M. E. 2006. "Modularity and Community Structure in Networks," *Proceedings of the national academy of sciences* (103:23), pp. 8577-8582.

Panniello, U., Gorgoglione, M., and Tuzhilin, A. 2016. "In Carss We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems," *Information Systems Research* (27:1), pp. 182-196.

Sahoo, N., Krishnan, R., Duncan, G., and Callan, J. 2012. "The Halo Effect in Multicomponent Ratings and Its Implications for Recommender Systems: The Case of Yahoo! Movies," *Information Systems Research* (23:1), pp. 231-246.

Xu, D. J., Benbasat, I., and Cenfetelli, R. T. 2017. "A Two-Stage Model of Generating Product Advice: Proposing and Testing the Complementarity Principle," *Journal of Management Information Systems* (34:3), pp. 826-862.

Yan, Z., Wang, T., Chen, Y., and Zhang, H. 2016. "Knowledge Sharing in Online Health Communities: A Social Exchange Theory Perspective," *Information & Management* (53:5), pp. 643-653.

Yang, H., and Gao, H. 2018. "Toward Sustainable Virtualized Healthcare: Extracting Medical Entities from Chinese Online Health Consultations Using Deep Neural Networks," *Sustainability* (10:9), p. 3292.

Zhang, K., Bhattacharyya, S., and Ram, S. 2016. "Large-Scale Network Analysis for Online Social Brand Advertising," *Mis Quarterly* (40:4).

Zhang, Y., He, D., and Sang, Y. 2013. "Facebook as a Platform for Health Information and Communication: A Case Study of a Diabetes Group," *Journal of medical systems* (37:3), p. 9942.