ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00601

User Influence in Online Health Communities: Integrating PageRank and Comment Sentiment Analysis Postprint

Authors: Dong Wei, Tao Jinhu

Date: 2023-04-01T16:02:51+00:00

Abstract

[Purpose/Significance] Effective identification of high-influence users in online health communities assists health information seekers in discovering valuable health information, which is of significant importance for reducing health information search costs and enhancing the effectiveness of health behavior decision-making. [Method/Process] From the perspectives of user interactivity and comment sentiment orientation, this study constructs a measurement methodology for user influence in online health communities utilizing algorithms such as PageRank and SVM. Taking Yixiang Network as the experimental case, and from the viewpoint of the utility value of published content, it further calculates the comprehensive influence of users within this community and conducts analysis on case users. [Results/Conclusion] The analysis results demonstrate that the proposed algorithm possesses certain rationality and can optimize the influence calculation outcomes of the PageRank algorithm; simultaneously, employing TF-IDF and mutual information algorithms reveals that the information content published by users with high comprehensive influence is fundamentally consistent with the content themes of other user groups in the community, and such users exert a certain guiding role on the community's thematic direction. Therefore, the method constructed in this study can effectively identify high-influence users, facilitate health information seekers to timely and accurately discover required information, improve the effectiveness of health information utilization, thereby enriching the theoretical and practical research on user information behavior in online health communities.

Full Text

Abstract

[**Purpose/Significance**] The effective identification of high-impact users in online health communities helps health information seekers discover valuable health information, which is significant for reducing health information search costs and improving the effectiveness of health behavior decision-making.

[**Method/Process**] From the perspectives of user interactivity and comment sentiment tendency, this study constructs a measurement method for user influence in online health communities using PageRank and SVM algorithms. Taking Yixiang.com as the experimental subject and from the perspective of the utility value of published content, we further calculate the comprehensive influence of users in this community and analyze case users.

[**Result/Conclusion**] The analysis results demonstrate that the algorithm has certain rationality and can optimize the influence calculation results of the PageRank algorithm. Meanwhile, using TF-IDF and mutual information algorithms reveals that the information content published by high comprehensive influence users is basically consistent with the content themes of other user groups in the community, and such users play a certain guiding role in the thematic direction of the community. Therefore, the method constructed in this study can effectively identify high-impact users, helping health information seekers discover required information in a timely and accurate manner, improving the effectiveness of health information utilization, and thus enriching the theoretical and practical research on user information behavior in online health communities.

Keywords: PageRank; Sentiment Tendency; Online Health Community; User Influence

1 Introduction

The "Internet + Healthcare" development strategy is a product of its time and an essential path toward intelligent healthcare. Users can not only make online appointments and consult materials but also receive answers from experts in their intended fields or learn from the experiences and discussions of fellow patients, significantly reducing the time cost of traditional medical consultations and greatly improving user engagement and treatment efficiency. According to medical statistics, in 2018, over 990,000 health institutions nationwide received 3.38 billion total consultations [1], and in 2019, the total number of online consultations reached 560 million, with a continued upward trend expected [2]. Meanwhile, the "Opinions on Promoting 'Internet + Healthcare' Development" [3] encourages online health communities to utilize internet-related technologies to accelerate resource sharing, information exchange, and telemedicine services, continuously improving the integrated "Internet + Healthcare" service system

and strengthening effective communication among hospitals, doctors, and patients.

Currently, major domestic online health communities with large-scale health discussions include Yixiang.com, 39 Health Forum, and Haodf.com. These online communities have numerous users, rapid knowledge dissemination, and generate massive amounts of information and data, providing valuable health information for health information seekers. Some highly active users in online communities can attract attention and interaction from other users, thereby influencing other users' information behaviors and health decisions to a certain extent and playing a strong guiding role in information dissemination across the entire online community. However, user activity level is not directly related to the utility value of the information they publish. For example, some users have strong interactive influence and high community activity, frequently seeking help and expressing emotions, and their published information receives considerable attention, but other users' evaluations of them are not high, reflecting limited utility value of their information to some extent. There are also users who, despite low interactive activity, receive more positive evaluations for their published information, which has good application value.

Therefore, from the perspective of information utility value, how to identify the comprehensive influence of online health community users by combining user activity and interactive sentiment tendency is important for helping users utilize health information conveniently and effectively and make objective health behavior decisions. This study proposes to exploratorily construct a comprehensive influence measurement algorithm for online health community users based on the fusion of user interactive activity and comment sentiment tendency, and conduct experiments and result analysis in corresponding online health communities to provide methods and references for effectively mining influential users and valuable health information in online health communities.

2 Related Research

User influence analysis and measurement is one of the important research directions in the fields of online social media and online communities. Current research on user influence mainly adopts eigenvalue statistical analysis methods, social network analysis methods, and PageRank methods.

Eigenvalue Statistical Analysis Methods primarily calculate user influence by statistically analyzing relevant eigenvalues that reflect the active characteristics of online community users and setting certain indicators and weights. For example, Wang Jiamin et al. [4] mainly counted influence indicators and activity indicators when analyzing user influence, where influence indicators included four eigenvalues: number of followers, number of reposts, number of comments, and verification status, while activity indicators included two eigenvalues: number of posts and number of followings. Zhao Fazhen et al. [5] used eigenvalues such as blog citation quantity, reply quantity, and internal/external link counts

to model user influence. Dong Wei et al. [6] also identified active users and analyzed their influence in the community by obtaining and analyzing relevant eigenvalues reflecting personal and interactive dimensions, such as user retention time, post quantity, and number of followers.

Social Network Analysis Methods primarily calculate the importance of each network node in the network through attribute values in the relationship network structure, such as network density, degree centrality, betweenness centrality, and closeness centrality. Chen Yuan et al. [7] mined user influence in online communities by analyzing centrality and structural holes indicators of social networks. Xie Yingxiang et al. [8] analyzed user influence in virtual communities using centrality analysis in social network analysis and MDS methods, further revealing the existence of opinion leaders in the community. S. Jonnalagadda et al. [9] comprehensively analyzed centrality indicators such as degree centrality, betweenness centrality, and closeness centrality to discover opinion leaders with significant influence in medical online communities.

PageRank Algorithm posits that interactive relationships such as likes, reposts, and comments among users in social networks are very similar to links between web pages, so analysis methods for link structures between web pages can also be used to analyze interactive relationships such as reposts and comments among social network users [10]. The PageRank algorithm has been increasingly applied by scholars to the analysis and measurement of user influence in online communities. For example, Liu Ling et al. [11] and Zhang Junhao et al. [12] incorporated indicators such as repost rate, comment rate, number of posts, and time intervals into user behavior on the basis of the PageRank algorithm to explore core contributors and high-influence users in information dissemination in Weibo communities. X. Song et al. [13] proposed a comprehensive influence algorithm by combining the novelty of information provided by users with PageRank. Yuan Liling et al. [14] considered weighted social network-related factors on the basis of the PageRank algorithm to improve it and explore user influence. Xiao Yu et al. [15] further considered the degree of interaction between users and users' willingness to share on the basis of PageRank, proposing the Weibo-Rank algorithm for calculating user influence.

In summary, current research on user influence analysis mainly focuses on analyzing user interaction indicators and interactive network structure attributes, but most studies evaluate and analyze user influence in online communities from a single perspective, which reduces the effectiveness of user influence measurement to a certain extent. Although eigenvalue statistical analysis methods and social network analysis methods can measure community user influence to varying degrees, the former relies too heavily on feature scores and ignores real interactive influence, while the latter focuses more on small networks and direct relationship measurement. The PageRank algorithm can calculate interactive influence and incorporate more feature scores, offering good integration and more objectively and comprehensively reflecting user comprehensive influence. In online health communities, the sentiment tendency of interactive information

among users can effectively judge whether the information in the community has good utilization value, but most current studies ignore such subjective factors. Therefore, it is necessary to combine user interactive behavior and comment sentiment tendency to further improve and develop calculation and evaluation methods for user influence in online health communities. Hence, this study starts from the perspective of integrating interactivity and sentiment tendency, first uses the PageRank algorithm to rank the interactive influence of all users in online health communities, then identifies and finds the optimal machine learning sentiment classification model to recognize user comment sentiment tendency, and further fuses interactive influence and sentiment tendency to calculate and identify user comprehensive influence.

3 Research Design

3.1 Research Framework

The research framework of this study mainly includes four steps. First, use data crawlers to crawl relevant information from online communities, preprocess the data, and store the final available data in a database, including user and comment information. Second, calculate the comprehensive influence of users, which mainly includes three sub-algorithms: Use the PageRank algorithm to calculate user interactive influence; Through selecting the optimal sentiment classification model, classify and analyze comment information sentiment, and further calculate the sentiment tendency value of comment information; Fuse the results of the above two components according to a specific formula and compare through case analysis. Third, further explore the relationship between the information content published by high comprehensive influence users and the content theme direction of other user groups in the community using TF-IDF and mutual information algorithms, and conduct comparative analysis through visualization. Finally, summarize the research process and methods of this study and propose corresponding research prospects. As shown in Figure 1 [Figure 1: see original paper].

3.2 Data Acquisition and Preprocessing

This study takes the information published by users in health communities and their comment information as analysis objects, uses Python language to build a multi-threaded crawler tool, uses Cookie parameters and header information as user and browser characterization tools, and obtains relevant content from user communication in the community by parsing the DOM tree, including user nicknames, post content, and corresponding reply information. Furthermore, relevant data preprocessing is conducted, such as word segmentation, construction of user encoding mapping tables, construction of user comment mapping tables, construction of user commenter mapping tables, and handling of abnormal users. This study intends to take the user-generated content of the Yixiang.com community as an example and collect corresponding data for relevant experiments and analysis.

3.3 Analysis Process and Techniques

Traditional PageRank algorithms mainly consider websites or users' interactive relationships and weights without analyzing their quality. Therefore, this study combines user interactive relationships and user comments to analyze, on the one hand, to discover potential network user influence rankings, and on the other hand, to identify user sentiment tendency and fuse both for comprehensive exploration.

3.3.1 Calculation of User Interactive Influence This study extracts the mapping relationship between users and commenting users, sorts out the multi-co-occurrence relationship between posters and commenters, converts the specific interactive network into an interaction matrix, and uses the PageRank algorithm to obtain users with high interactive influence in the interactive network. The specific algorithm is as follows:

The basic PageRank algorithm idea is shown in Formula (1), where O represents nodes pointing to node A, PR(O) represents the PR value of nodes pointing to node A, L(O) represents the out-degree of nodes pointing to node A, PR represents the next iteration PR value of the corresponding node, and m represents the number of iterations when the model converges. N represents the total number of user nodes, each user node's initial PR value is 1/N, and the final PR value is the interactive influence score of these users.

$$PR' = \sum \frac{PR(O)}{L(O)} \quad \text{Formula (1)[16]}$$

For convenience of calculation, Formula (2) (equivalent conversion of Formula (1)) is generally used for calculation, where M is the transition matrix formed by the user interaction network in this study, as specifically shown in Formula (3), and M(u_i, u_j) represents user j's out-link to user i, i.e., user interaction situation. PR is the result of the previous iteration of PR .

$$PR' = M \cdot PR \quad \text{Formula (2)}$$

$$M = \begin{pmatrix} M(u_1,u_1) & M(u_1,u_2) & \cdots & M(u_1,u_N) \\ M(u_2,u_1) & M(u_2,u_2) & \cdots & M(u_2,u_N) \\ \vdots & \vdots & \ddots & \vdots \\ M(u_N,u_1) & M(u_N,u_2) & \cdots & M(u_N,u_N) \end{pmatrix} \quad \text{Formula (3)}$$

However, the above calculation method is powerless to explain the PR value of certain in-link nodes themselves, causing the PR value of nodes to shift and become erroneous, with the PR value of in-link nodes ultimately becoming 1 and other nodes' PR values becoming 0. To solve this problem, Formula (4)

is introduced for correction, where β is the damping coefficient with a value of 0.85, mainly used to solve trap and isolated point problems.

$$PR' = \beta \cdot M \cdot PR + \frac{1-\beta}{N} \quad \text{Formula (4)[16]}$$

The iteration stop condition is set as the next PR value being equal to the previous PR value, and $\sum_{p=1}^{N} PR'_p = 1$. This algorithm can effectively discover key figures in the interaction network and assign higher PR values to these users, thereby discovering users with high interactive influence.

3.3.2 Calculation of Sentiment Tendency Recognition for User Comment Text User-generated content comments in online health communities have obvious sentiment tendencies, which can serve as one of the important indicators for evaluating user influence and content quality [17]. This study uses supervised machine learning models for sentiment tendency recognition based on extensive feature extraction from text, including Random Forest algorithm, Logistic algorithm, SGD algorithm, SVM algorithm, and Naive Bayesian algorithm. Performance evaluation indicators include accuracy and F1 value, where the F1 value is related to both recall and precision and is generally considered a comprehensive indicator for evaluating model quality. The calculation methods are shown in Formulas (5) and (8), where TP represents the number of positive classes predicted as positive, TN represents the number of negative classes predicted as negative, FP represents the number of negative classes predicted as positive, and FN represents the number of positive classes predicted as negative. Sentiment classification mainly involves three aspects: Support, marked as 1; Discussion/negotiation, marked as 0; Opposition, marked as 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Formula (5)[18]}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{Formula (6)[18]}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Formula (7)[18]}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad \text{Formula (8)[18]}$$

Based on the above evaluation indicators, an appropriate model is selected for prediction, and the results are sorted out. As shown in Formula (9), this calculation idea can overcome interference caused by different proportions and orders of magnitude of data. In the formula, AV represents the sentiment tendency value, Up represents a certain user among all users, w belongs to class 0 or 1

(i.e., non-negative class), len(w) is used to measure the number of specific categories, and r represents the sentiment category of comments received by this user. To further reduce interference caused by order of magnitude relationships, each user's sentiment tendency value is placed in a list and normalized through a standardization function.

$$AV = \sum_{w\in[0,1],r\in[0,1,2]} Up(len(w)) \cdot 2^{if(w==r)} \quad \text{Formula (9)}$$

3.3.3 Analysis of Comprehensive Influence Fusing Two Algorithms This study fuses the above PageRank algorithm and sentiment tendency recognition results to explore a comprehensive evaluation of health community user influence from both interactivity and comment sentiment aspects. On the basis of the PageRank algorithm, the sentiment tendency value is fused, i.e., the user's sentiment tendency value is used as the corresponding user's weight and fused with interactive influence to form a new comprehensive influence value for the user, as shown in Formula (10), where p represents a certain user among n users, UR represents comprehensive influence, PR _p represents interactive influence, and AV_p represents sentiment tendency value.

$$UR_p = PR'_p \cdot AV_p, \quad p \in [1, 2, 3, \ldots, n] \quad \text{Formula (10)}$$

3.3.4 Analysis of Text Content from Comprehensive Influence Users To further explore the impact of high comprehensive influence users on the thematic direction of health communities, i.e., whether the information text published by these users represents or influences the content theme direction of the community to a certain extent, this study further uses TF-IDF and mutual information algorithms to construct co-word matrices of content generated by different user groups for analysis and comparison. First, TF-IDF is used to calculate high-frequency words in content generated by high comprehensive influence users and all users in the community, and then the mutual information algorithm is used to extract the most relevant terms to high-frequency words, thereby forming co-word matrix networks for high comprehensive influence users and other user groups, respectively, for further comparison to explore the relationship between content published by high comprehensive influence users and content published by other users in the community.

(1) TF-IDF Calculation. TF-IDF is a weighted algorithm whose advantage lies in filtering out common but meaningless words in text while retaining words that truly affect the text. Therefore, TF-IDF is more accurate and objective than ordinary word frequency statistics. The specific algorithm is as follows:

$$TF\text{-}IDF = \frac{N_{i,j}}{\sum_{k=1}^{N} N_{k,j}} \cdot \log \frac{D}{D_i + 1} \quad \text{Formula (11)[19]}$$

Where N_{i,j} represents the frequency of occurrence of keyword i in document j, $\sum_{k=1}^{N} N_{k,j}$ represents the total number of words in the article corresponding to k keywords (i.e., the first half of the calculation is called TF, representing the frequency of keyword i in document j). D represents the total number of documents in the corpus, and D_i represents the number of documents containing keyword i among D documents. To avoid the situation where no documents contain the word, 1 is added to the denominator.

(2) Mutual Information. Mutual information mainly refers to the degree to which uncertainty about another term is reduced when one term is known. Specifically, we need to first use JIEBA for word segmentation of user-generated content, then traverse the dependency measurement between high-frequency terms and other segmented words, and on this basis form a high-frequency word-mutual information network for visualization and comparative analysis. The basic algorithm is as follows:

$$M(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad \text{Formula (12)[20]}$$

Where p(x, y) represents the joint probability distribution of two words, i.e., the probability that term x and y co-occur in user-generated content, and p(x) and p(y) represent the probability distributions of term x and term y in user-generated content, respectively. Generally, the larger M(x, y), the closer the relationship between the two, and the greater the probability of co-occurrence, and vice versa.

4 Research Results

4.1 Experimental Data

Yixiang.com is one of the online health communities in China with many users and high credibility, supporting case database queries and online health problem Q&A, among which the gout circle community has frequent content interaction and relatively comprehensive discussions [21]. Therefore, this study sets Yixiang.com's gout circle as the data source, with data collection in February 2020. Based on relevant public content, a crawler program was designed for data extraction, mainly including user nicknames, post content, and reply content.

Further data preprocessing was conducted, including word segmentation using JIEBA for user text to conduct high-frequency word statistics and mutual information model construction; construction of user encoding mapping tables, i.e., unified encoding of all users such as User 1, User 2, etc.; user comment mapping tables, i.e., corresponding comment content published by users; user commenter mapping tables, i.e., constructing mapping tables of user comments such as Commenting User 1, Commenting User 2, etc.; and abnormal user handling, i.e., filtering out users whose comments or posts are irrelevant to the gout

circle, such as advertising push users. After final preprocessing, 2,560 valid interaction contents from 292 valid users were obtained.

4.2 Analysis Results

4.2.1 Results of User Interactive Influence Analysis Based on PageRank The calculation of user interactive influence is mainly conducted using the PageRank algorithm in Formula (4) of this study. The specific analysis results are shown in Figure 2 [Figure 2: see original paper]. From the overall distribution, most users have low interactive influence, while only a small portion of users have high influence levels, such as User 253, User 269, User 151, User 154, etc., indicating that these users receive more attention from other users and have certain influence.

However, the ranking based on the PageRank algorithm only considers the interaction mechanism between users for determination. Although it has certain practicality, it ignores the judgment of information utility value. That is, some users may have high interactive influence, but if the information they publish is questioned or negated by most other users, the effectiveness of that information will be affected. For example, comments on high interactive influence User 253 and User 269 include "Are you a TCM doctor?" "It's because it can't be cured" "Are you being fooled by a quack?" "You're ignorant yourself" and other negative or questioning comments, which will affect the user's interactive influence to a certain extent. Therefore, this study further explores the impact of comment sentiment tendency on user influence and explores fusing comment sentiment tendency into user interactive influence values to comprehensively discuss and analyze the utility value of user-generated content, thereby enhancing the objectivity and effectiveness of user influence measurement.

4.2.2 Analysis Results of User Comprehensive Influence Fusing Comment Sentiment Tendency (1) Selection and Analysis of Sentiment Tendency Classification Models. Based on the PageRank analysis results, this study fuses user sentiment tendency analysis to analyze relevant texts. To determine the optimal model for comment sentiment tendency analysis, this paper selects five classic machine learning algorithms for comparison: Random Forest, Logistic Regression, SGD (Stochastic Gradient Descent), SVM (Support Vector Machine), and Bayesian (Naive Bayes). First, two rounds of manual data annotation were conducted for sentiment tendency in the text, with consistency reaching over 95%. To optimize sentiment tendency recognition effectiveness, through multiple rounds of testing and debugging, the final main parameter settings were determined: Random Forest set min_{{{samples}}*{{leaf}}} to 1, min{{{samples}}*{{split}}} to 2, criterion to "gini" algorithm, and n{estimators} to 10; SGD set loss to "log" and max_{iter} to 100; SVM set kernel to "linear" and C to 1; Logistic and Naive Bayes both used default parameters for comparison and judgment, thereby selecting a model with higher comprehensive performance and more stable

predictive ability as the basis for fusion with the PageRank algorithm.

Each training re-evaluated the training set data, with the test set accounting for 20% of the total data and the training set accounting for 80%, iterating 10 times respectively. The specific calculation results are shown in Table 1 . It can be found that the Logistic regression algorithm has a relatively low F1 value, indicating average model performance, while the SVM model based on linear function has the highest average F1 value (AVEG_{F1}) and average accuracy (AVEG_{ACC}), slightly better than other algorithms, with the smallest variance (S2_{F1}), having more stable predictive ability. Therefore, the SVM model is selected to identify and classify the overall data.

(2) Analysis Results of User Comprehensive Influence. On the basis of the above research, sentiment tendency and interactive influence are further fused for analysis, and the comprehensive influence distribution is obtained. The results are shown in Table 3 and Figure 4 [Figure 4: see original paper]. In Figure 4, the horizontal axis represents 292 users, and the vertical axis represents the result of fusing the sentiment tendency value with the user interactive influence PR value, i.e., comprehensive influence. Most users are within the range of 0 to 0.01, some users have relatively larger values above the 0.01 level, with the highest reaching about the 0.07 level. Among them, User 151's comprehensive influence has been greatly improved, reaching 0.0715, with 29% of other users having strong positive sentiment tendency toward its content, which has a significant impact on the user's comprehensive influence; User 154 reached 0.0312, with 33% of users holding positive sentiment tendency, but 13% having negative sentiment tendency.

Additionally, to further explore the transformation mechanism from different users' interactive influence to comprehensive influence, this study sorted out the specific indicators and comment data content of specific users. Taking the four typical users marked in the above figures (User 151, User 154, User 253, User 269) as case objects, the specific results are shown in Table 4 . Since User 151 and User 154 received comments mostly such as "Learned, thank you," "Thumbs up," "Thanks thanks" and other positive texts, these users have larger sentiment tendency values, have more important practical value and dissemination significance, and overall improve users' comprehensive influence. However, although User 253 and User 269 have high interactive influence, since their received comments are mostly questioning and negative, such as "Are you a TCM doctor?" "Can it really be cured?" "You're ignorant yourself," their comment sentiment tendency values are lower, thus affecting their comprehensive influence and causing it to decline.

4.2.3 Analysis of Text Content from Comprehensive Influence Users Based on Mutual Information To further explore the impact of high comprehensive influence users on the thematic direction of health communities, this study selected the top 20 high comprehensive influence users and other user comment content in the community for experiments. Through TF-IDF and mu-

tual information algorithms, a term co-occurrence network for this user group was constructed. To more intuitively and clearly display its association effects and overall structure, Vosviewer software was used for visualization analysis of the co-occurrence network, with specific results shown in Figure 5 [Figure 5: see original paper]. Among them, 15 nodes with larger radii such as pain, crystallization, blood disease, high uric acid, attack, joint, pain relief, and metabolism represent high-frequency vocabulary, while other nodes with smaller radii represent the most relevant several terms for each high-frequency word. From Figure 5 [Figure 5: see original paper], it can be learned that high comprehensive influence users mainly focus on several issues concentrated in three aspects: first, the manifestations when gout attacks, such as pain, before sleep, erosion, crystallization, precipitation, phospholipids, joints, nerves and other keywords all explain symptoms, timing and other content of disease occurrence from different aspects; second, medications used to treat gout, including colchicine, diclofenac sodium, acemetacin, dafalgan, colchicum, allopurinol, anti-inflammatory drugs, pain relief injections and other content; third, dietary therapy aids for better gout treatment, such as special attention to vegetables with high purine content like asparagus, spinach, mushrooms, fresh peas, as well as clams, animal organs, drinking more water, and less soup.

Similarly, we analyzed the high-frequency words and mutual information co-occurrence matrix generated by other users in the community, with specific results shown in Figure 6 [Figure 6: see original paper]. It can also be roughly divided into three main aspects: first, specific symptoms of the disease, such as keywords like erosion, swelling pain, severe pain, contracture, numbness, and redness; second, drug treatment, such as colchicine, acemetacin, allopurinol, dihydrochlorothiazide, hypoxanthine, indomethacin, ibuprofen, heat dissipation and other medications and methods; third, auxiliary treatment, where it can be found through keywords that on the one hand, foods like tofu skin, ham sausage, meat dishes should be avoided, while vegetables, fruits, wild rice stems, inorganic salts, and related alkaline foods can be consumed.

Overall, the content themes that high comprehensive influence users focus on have strong consistency with the themes that other user information in the community focuses on, i.e., the core content discussed by high comprehensive influence users is consistent with that discussed by most users. This indicates to a certain extent that the information published by the high comprehensive influence users identified in this study has strong utility value, and the thematic content they publish guides the information publication direction in the community to a certain extent, also indicating that the comprehensive influence identification and analysis method constructed has certain rationality and objectivity. At the same time, for health information seekers, effectively identifying high-influence users and their related information can better save seekers' information search costs, quickly understand the characteristics and corresponding thematic directions of the community, and help improve the efficiency of health behavior decision-making for seekers.

5 Conclusion and Future Work

5.1 Research Summary

This study starts from the two perspectives of interactive influence and sentiment tendency, establishes a sentiment recognition model based on the linear kernel function of the SVM algorithm, identifies and analyzes text effectiveness, and explores user comprehensive influence by combining interactive influence and sentiment tendency, drawing the following conclusions:

First, through the calculation of user interactive influence and combined with relevant cases, this study finds that interactive influence emphasizes interactive activity more but has certain deficiencies in revealing the effectiveness of user information resources. Therefore, it cannot completely and objectively reflect users' real influence, and it is necessary to introduce comment sentiment tendency values for further fusion calculation of influence.

Second, by comparing five main machine learning algorithms for sentiment calculation, the SVM algorithm is found to have the optimal effect for the comment sentiment tendency classification model constructed in this paper, providing technical support for effectively calculating user comprehensive influence.

Third, this study exploratorily fuses PageRank interactive influence and comment sentiment tendency for calculation, and further verifies high-influence users from the perspective of information content through corresponding case analysis, demonstrating to a certain extent that the comprehensive influence algorithm in this study has good rationality and applicability.

In addition, through the comparison of high-frequency word-mutual information matrices of high comprehensive influence users and other user groups, it is found that the similarity between the two is high and the basic thematic direction is consistent, which also indicates the necessity of finding high comprehensive influence users to a certain extent. It further shows that the user influence comprehensive calculation method in this study can more objectively identify users with higher influence who dominate the content direction of health communities, helping health information seekers obtain required valuable information from health communities in a timely and accurate manner, improving the utilization effect of health information, and thus enriching the theoretical and practical research on user information behavior in online health communities.

5.2 Future Work

This paper proposes a sentiment recognition model to explore the sentiment tendency of user-generated content, thereby constructing a research method for user comprehensive influence, and further discusses the impact of high comprehensive influence users on community direction through specific content analysis. However, there are still certain limitations:

(1) Optimization of Interactive Influence and Sentiment Analysis Algorithms. The user interactive influence in this study is mainly based on the PageRank algorithm. Although this method is widely used, there is still room for improvement in analyzing user influence. Future research can further optimize this algorithm by combining user behavior characteristics. In addition, the sentiment tendency analysis algorithm used in this study can be compared with more relevant algorithms and frameworks in future research to further improve calculation efficiency and accuracy.

(2) Further Enrichment of Research Data. This study is mainly based on data from the gout circle of Yixiang.com. In future research, the scope of data acquisition for health communities can be further expanded to continuously expand and verify the applicability of this study by comparing the distribution and characteristics of user comprehensive influence in different health communities.

References

[1] Yang Zi. Latest! National Health Commission Releases National Medical Data [EB/OL]. [2021-04-27]. https://www.sohu.com/a/247593213_{439958}.

[2] Analysis of Development Status and Trends of China's Health Medical Big Data Industry in 2018 [EB/OL]. [2021-04-27]. http://www.chyxx.com/industry/201806/649591.html.

[3] General Office of the State Council. Opinions of the General Office of the State Council on Promoting the Development of "Internet + Healthcare" [EB/OL]. [2021-04-27]. http://www.pkulaw.cn/fulltext_{form}.aspx?Db=chl&Gid=37395b41f6f018e4bdfb&ke

[4] Wang Jiamin, Wu Peng, Chen Fen, et al. Empirical Research on Opinion Leader Identification and Influence in Emergencies [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(2): 169-176.

[5] Zhao Fazhen. Research on Network Community Influence Based on Link Analysis Method—Taking 30 Domestic Network Community Websites as Examples [J]. Modern Information, 2013, 33(6): 91-95.

[6] Dong Wei, Li Jianhong, Tao Jinhu. Identification of Active Users in Online Health Communities and Analysis of Their Interaction Types [J]. Journal of Documentation and Data Studies, 2020, 2(1): 89-101.

[7] Chen Yuan, Liu Xinyu. Research on Opinion Leader Identification Based on Social Network Analysis [J]. Information Science, 2015, 33(4): 13-19, 92.

[8] Xie Yingxiang, Feng Rui. Research on the Influence of Blog Network Position in Virtual Teacher Communities [J]. Modern Educational Technology, 2010, 20(1): 97-100, 110.

[9] Jonnalagadda S, Peele R, Topham P. Discovering Opinion Leaders for Medical Topics Using News Articles [J]. Journal of Biomedical Semantics, 2012, 3(1): 2.

[10] Chen Fen, Gao Xiaohuan, Peng Yue, et al. Fusion of Text Tendency Analysis for Weibo Opinion Leaders [J]. Information Network Security, 2015(6): 73-78.

[11] Liu Ling, Yang Changchun. A New Weibo Community User Influence Evaluation Algorithm [J]. Computer Applications and Software, 2017, 34(7): 212-216, 261.

[12] Zhang Junhao, Gu Yijun, Zhang Shihao. Weibo User Influence Assessment Based on PageRank and User Behavior [J]. Information Network Security, 2015(6): 73-78.

[13] Song X, Chi Y, Hino K, et al. Identifying Opinion Leaders in the Blogosphere [C]//ACM. Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. Lisbon: ACM, 2007: 971-974.

[14] Han Zhongming, Yuan Liling, Yang Weijie, et al. Important Node Discovery Algorithm in Weighted Social Networks [J]. Computer Applications, 2013, 33(6): 1553-1557, 1562.

[15] Xiao Yu, Xu Wei, Shang Zhaoxi. Weibo User Regional Influence Identification Algorithm and Analysis [J]. Computer Science, 2012, 39(9): 38-42.

[16] Ma Feng. Research on Journal Influence Based on PageRank Algorithm [J]. Information Magazine, 2014, 33(12): 103-108.

[17] Zhang Y. Determinants of Poster Reputation on Internet Stock Message Boards [J]. American Journal of Economics and Business Administration, 2009, 1(2): 114.

[18] Understanding of Precision, Recall, F1 score, Accuracy [EB/OL]. [2021-04-27]. https://blog.csdn.net/u014380165/article/details/77493978.

[19] Alam S, Yao N. Big Data Analytics, Text Mining and Modern English Language [J]. Journal of Grid Computing, 2019, 17(2): 357-366.

[20] Fei Hongxiao, Kang Songlin, Zhu Xiaojuan, et al. Research on Chinese Word Segmentation Based on Word Frequency Statistics [J]. Computer Engineering and Applications, 2005(7): 67-68, 100.

[21] Dong Wei, Tao Jinhu. Identification of User Interest Groups Based on Topic Preference in Online Health Communities [J]. Data Analysis and Knowledge Discovery, 2019, 3(11): 120-128.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.