User Influence Analysis and Key Opinion Leader Mining on the Zhihu Platform (Postprint)

Authors: Guo Bo, Xu Haodi, Lei Shuiwang

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] With the rapid development of Internet technologies, the Zhihu platform has gradually become a vehicle for discussing popular social public topics and sharing knowledge and experience. Therefore, analyzing the influence of key users on the Zhihu platform and mining key opinion leaders therein is of great significance in studying information dissemination pathways in social networks. [Method/Process] By proposing improved PageRank and HITS algorithms, user influence mining models are constructed based on Zhihu's user social network and question-and-answer network respectively, which can accurately and objectively identify key users and opinion leaders within the platform. [Results/Conclusion] Experimental results demonstrate that the proposed PageRank and HITS algorithms can effectively identify key opinion leaders with notable characteristics on the Zhihu platform, and the algorithms demonstrate rapid convergence, reusability, and transferability. Through processing and effective analysis of Zhihu platform user datasets, user influence and key opinion leader mining models are successfully established; simultaneously, validation is conducted on specific topics. Therefore, it can be inferred that this model possesses significant application value and promising commercial promotion prospects.

Full Text

Preamble

Vol. 62 No. 20, October 2018
ChinaXiv Cooperative Journal

Analysis of User Influence and Identification of Key Opinion Leaders on the Zhihu Platform

Guo Bo[1], Xu Haodi[2], Lei Shuiwang[3]
[1] Meizu Technology Co., Ltd. Beijing Branch, Beijing 100872

[2] Hong Kong University of Science and Technology, Hong Kong 999077
[3] Library of Xuzhou College of Industrial Technology, Xuzhou 221140

Abstract

[**Purpose/Significance**] With the rapid development of internet technology, the Zhihu platform has gradually become an important carrier for discussing public social issues and sharing knowledge and experiences. Therefore, analyzing the influence of key users and mining key opinion leaders on Zhihu plays a crucial role in studying information dissemination pathways in social networks. [**Method/Process**] By proposing improved PageRank and HITS algorithms, this study constructs user influence mining models based on Zhihu's social network and question-answering network to accurately and objectively identify key users and opinion leaders. [**Result/Conclusion**] Experimental results demonstrate that the proposed PageRank and HITS algorithms can effectively extract key opinion leaders with prominent characteristics from the Zhihu platform, featuring fast convergence speed, reusability, and transferability. Through processing and analyzing the Zhihu user dataset, we successfully established a model for evaluating user influence and mining key opinion leaders, which was validated on specific topics. Consequently, it can be inferred that this model possesses significant application value and commercial promotion prospects.

Classification Number: TP393
Keywords: Zhihu, user influence, key opinion leader, PageRank algorithm, HITS algorithm

Introduction

With the development of social networks, platforms such as Weibo and Zhihu have gradually become common tools for elites across various industries to exchange professional information, specialized knowledge, and life experiences [1]. In particular, Zhihu's Q&A platform not only satisfies users' basic social needs (mutual following and visiting) but also fulfills human needs for establishing personal prestige and self-actualization through sharing and Q&A processes, thereby attracting numerous authoritative figures including industry elites, experts, internet celebrities, and socialites. Key opinion leaders refer to "active individuals" who provide information to others while exerting influence in interpersonal communication networks [2]. They serve as both authoritative sources and primary disseminators in the information propagation process, forming information cascade transmission in social networks. In recent years, extensive research has shown that key opinion leaders play vital roles in online knowledge dissemination [3], word-of-mouth effects [4], and online communication [5], holding significant meaning for network intervention [6], online marketing [7-9], and network structure analysis [10].

To comprehensively identify key users in the Zhihu network, we constructed a multi-level comprehensive evaluation system based on user interactions including social connections and Q&A behaviors, forming a social Q&A network with multi-agent direct interactivity, long-term engagement, and near-simultaneity according to their relationships. Most existing literature on identifying key opinion leaders employs social network analysis methods, which offer distinct advantages over other approaches [11]. We established a two-layer Zhihu network structure using social relationships, Q&A relationships, and their interconnections, as illustrated in Figure 1 [Figure 1: see original paper]. In Figure 1, each circular node represents a user, square nodes represent answers, and connections between them represent social and Q&A relationships.

Based on this two-layer structure, we first employed the Analytic Hierarchy Process (AHP) to construct comprehensive Zhihu user evaluation indicators and establish a user influence evaluation model. This model yields influence weights for each Zhihu user across AHP indicators, which are then integrated with traditional PageRank and HITS algorithms to create an improved comprehensive evaluation model for mining key users and opinion leaders on the platform. The research process and framework are shown in Figure 2 [Figure 2: see original paper].

Literature Review

Current methods for identifying key opinion leaders and user influence primarily include Analytic Hierarchy Process and social network mining approaches. AHP utilizes key evaluation indicators to characterize opinion leader features, establishing corresponding multi-layer indicator systems for scoring and ranking. For instance, in Douban opinion leader identification, features such as centrality, activity, attractiveness, and contagion were used to evaluate users [12]. In Weibo user influence assessment, numerous studies have employed indicators including follower count, post count, reposts, comments, popularity, interaction rate, and originality to construct hierarchical structures and obtain final results through weighted averaging [13]. AHP can establish comprehensive evaluation systems based on extensive subjective research experience, yielding structural multi-objective system evaluation models.

Social network mining methods typically construct corresponding social networks based on user relationships or forwarding/Q&A interactions, then analyze user importance using network structure-based algorithms or metrics to produce final rankings. Examples include using PageRank to mine important users in Twitter social networks [14], applying HITS to identify opinion leaders in Weibo forwarding networks [15], and obtaining more accurate importance rankings through hybrid evaluation models like SALSA or weighted evaluation algorithms [16]. Social network-based opinion leader mining offers greater objectivity and accuracy compared to indicator-based approaches. Therefore, this study combines both methods, using AHP to evaluate user activity, credibility, and influence, and applying these results to inform social network algorithms.

The commonly used social network mining algorithm is PageRank. The original

PageRank model was proposed by Google for evaluating webpage importance. Recent research has focused on improving traditional PageRank for mining key users in social networks [17], demonstrating its high efficiency and structural stability. However, on Zhihu's Q&A platform, users' repeated questioning and answering behaviors reflect their activity and information diffusion capacity without dispersing influence contributions. Therefore, we also employ the HITS algorithm for multi-angle measurement of user social and Q&A behaviors, weighting both algorithms' results for comprehensive ranking.

The HITS algorithm, initially proposed by Kleinberg, is a structural mining algorithm for webpage ranking [18]. In this study, HITS utilizes a mutually reinforcing relationship to identify key users and information disseminators that meet requirements. We first establish social and Q&A networks from the dataset, then dynamically mine user influence using improved PageRank and HITS algorithms. Subsequently, ranking algorithms produce user influence rankings across various indicators, which are weighted and averaged for final influence ordering. Finally, we construct a topic-based key opinion leader mining model to analyze characteristics and validate algorithm effectiveness.

Dataset and Network Construction

Dataset

The Zhihu dataset used in this paper includes basic information from 26,000 registered Zhihu users, comprising features such as follow count, follower count, upvotes received, thanks received, favorites received, answer count, question count, and article count. Additionally, corresponding follow relationships and Q&A relationships were obtained. Two active subsets were selected for analysis: users with thanks received $>$ 10,000 (1,607 users) and users with thanks received $>$ 50,000 (398 users), designated as Net10K and Net50K. Their specific network characteristics are shown in Table 1 .

Social Network and Q&A Network Construction

Based on Zhihu's user and content structure, we constructed a two-layer network (Figure 1). Mutual following relationships between users constitute the first layer. For example, U2 and U3 follow each other, represented by a bidirectional arrow; U1 follows U3, represented by a unidirectional line from U1 to U3. The Q&A network comprises questions posed by users and their answers. In Figure 1, users U2 and U5 posed questions Q1 and Q2 respectively, while answers A1 and A2 responded to Q1, represented by unidirectional lines from A1 and A2 to Q5. We abstract and model this structure accordingly.

The social network can be defined as an unweighted directed graph $G_{SN} = (V, E)$, where G_{SN} represents the Zhihu social network, V is the vertex set with each independent user v_i V, and E is the edge set where $<v_i, v_j>$ E if user v_i follows user v_j. This structure is shown in Figure 3 Figure 3: see original paper.

The Q&A network for specific topics is defined as a weighted directed graph G_{QA} = (V, E, W, P), where G_{QA} represents the Zhihu Q&A network, V is the vertex set representing users, E is the edge set where <v_i, v_j>   E indicates user v_i answered a question posed by user v_j, W is the edge weight vector with w_{ij} representing the number of Q&A relationships between users v_i and v_j, and P represents vertex strength. Since each user has different influence in real networks, p(i) is defined based on the comprehensive evaluation model determined by AHP. This structure is shown in Figure 3(b) [Figure 3: see original paper].

Key Opinion Leader Mining Model

AHP-Based Evaluation Indicator Model

AHP is a decision-making method combining quantitative and qualitative analysis through hierarchical evaluation indicator structures. Based on Zhihu dataset characteristics, we established three primary indicators and eight secondary indicators, as shown in Table 2 .

Assuming equal weights for secondary indicators under the same primary indicator, the primary indicator formulas are defined as:

$$A = (A_1 + A_2 + A_3), \quad Q = (V + T + P), \quad I = (F_1 + F_2)$$

Secondary indicators are normalized from corresponding real data using:

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

The comprehensive evaluation model for user i is:

$$U(i) = w_A A(i) + w_Q Q(i) + w_I I(i)$$

where A, Q, and I represent activity, credibility, and influence, with weight vector w = (w_A, w_Q, w_I). The algorithm is summarized in Table 3 .

PageRank-Based Key Opinion Leader Mining Model

PageRank is a common algorithm in data mining that uses page authority values to assess website importance. A page's authority is defined as the sum of authority values allocated by other pages linking to it, with final rankings obtained through iterative calculation [19]. Assuming users have corresponding interactions in both social and Q&A networks, we combine Zhihu Rank values with PageRank to create an improved algorithm.

For the unweighted social network, vertex i's social network ranking is defined as:

$$SR(i) = (1 - \alpha) + \alpha \sum \frac{SR(j)}{d_j}$$

where d_j is node j's out-degree and α is the damping coefficient (typically 0.85 [20]). To address authority stagnation in users with many followers but few followings, we introduce the random surfer model. The algorithm is summarized in Table 4 , with convergence parameter e set to 10^-20.

The Q&A network is weighted, with edge weight defined as:

$$w_{ij} = p(i) \cdot N_{ij}$$

where p(i) is the comprehensive influence indicator from AHP and N_{ij} is the Q&A relationship count between users i and j. Considering these weights, vertex i's Q&A network ranking QR(i) is:

$$QR(i) = (1-\alpha) + \alpha \sum \frac{QR(j) w_{ji}}{\sum w_{jk}}$$

The comprehensive Zhihu ranking combining both networks is:

$$ZR(i) = w_1 SR(i) + w_2 QR(i)$$

where w = (w_1, w_2) is the weight vector. After ranking all users, the top N are selected as key opinion leaders. The Q&A network PageRank algorithm is summarized in Table 5 .

HITS-Based Key Opinion Leader Mining Model

The HITS algorithm, originally for webpage ranking, classifies pages into hubs and authorities. Authorities are important pages recognized for specific topics, while hubs point to many relevant authorities. They mutually reinforce each other: good authorities are pointed to by many good hubs, and good hubs point to many good authorities [22]. In Zhihu's context, authority values represent opinion leaders while hub values represent information disseminators.

For the unweighted social network, each iteration t is defined as:

$$a_i^{(t+1)} = \sum h_j^{(t)}, \quad h_i^{(t+1)} = \sum a_j^{(t)}$$

where a_i is user i's authority value and h_i is hub value. Iteration continues until convergence with parameter e = 10^-20. The algorithm is summarized in Table 6 .

For the weighted Q&A network, we incorporate weights from the previous definition. The improved iterative formulas are:

$$a_i^{(t+1)} = \sum h_j^{(t)} \cdot w_{ij}, \quad h_i^{(t+1)} = \sum a_j^{(t)}$$

The resulting authority and hub values are denoted as AUTH_{QA}(i) and HUB_{QA}(i). Comprehensive rankings are calculated as:

$$ZR\_AUTH(i) = w_1 AUTH\_SR(i) + w_2 AUTH\_QA(i)$$

$$ZR\_HUB(i) = w_1 HUB\_SR(i) + w_2 HUB\_QA(i)$$

Experiments and Analysis

Network Structure

Zhihu's social network consists of large-scale mutual following relationships, while the Q&A network comprises corresponding Q&A relationships. Due to computational constraints, we selected the active subsets Net10K and Net50K to mine TOP10 key opinion leaders. To analyze basic network characteristics, we plotted in-degree distributions on Net10K. Figures 4(a) and 4(b) [Figure 4: see original paper] show both networks are scale-free, following power-law distributions.

The social network's degree distribution concentrates between 500-2000, with fewer than 0.1% of users having in-degree > 7000. The Q&A network's distribution concentrates between 100-250 with power exponent $\gamma = 1.78$, indicating a relatively sparse adjacency matrix with few high in-degree nodes, suggesting dispersed distribution of authoritative users. Further statistics reveal that among 27,846 Q&A relationships, the top 100 answerers account for 49% while the top 100 questioners account for only 2.8%.

Table 7 presents basic statistics for Net10K users across multiple dimensions: upvotes, in-degree, eigenvector centrality, betweenness centrality, and closeness centrality. These metrics measure influence from various perspectives: upvotes represent authority and exposure; in-degree indicates cohesion and importance; eigenvector centrality reflects importance based on neighbor quality; betweenness centrality shows brokerage potential; closeness centrality represents proximity to other nodes.

PageRank Algorithm Results and Analysis

Based on the improved PageRank algorithm, Tables 8 and 9 present PageRank values from social networks (SR) and Q&A networks (QR) for Net10K and Net50K, with weighted comprehensive Zhihu Rank values (ZR). The tables show TOP10 key users by ZR value.

The Net10K results show substantial overlap between TOP10 key users and basic statistical TOP10, validating the PageRank algorithm's effectiveness in objectively reflecting influential users with information guidance and dissemination roles. Due to significant differences in node counts between Net10K and Net50K (average shortest path lengths of 2.31 vs. 1.86), Net50K exhibits a tighter attention circle. The algorithms successfully identify key opinion leaders across different network densities and structures, with some leaders maintaining high PageRank values in both datasets, indicating stable authority unaffected by network scope reduction.

Figure 5 [Figure 5: see original paper] plots various metrics for TOP40 key opinion leaders from Net10K. When N $>$ 25, influence-determining parameters stabilize; when N $<$ 10, significant fluctuations occur. Therefore, the optimal focus range for TOPN key opinion leaders is 10-20. TOP10 leaders consistently show high answer counts and follower numbers, confirming our analysis's validity.

HITS Algorithm Results and Analysis

Based on the improved HITS algorithm, Tables 10 and 11 present authority values (AUTH_{SR}, AUTH_{QA}) and hub values (HUB_{SR}, HUB_{QA}) from both networks, weighted to produce final Zhihu authority ranking ZR_{AUTH} and hub ranking ZR_{HUB}.

Table 10(a) shows that TOP5 authority users all appear in PageRank's TOP10, demonstrating that both algorithms effectively capture comprehensive influence and authority. Table 10(b) reveals that TOP5 hub users mostly don't appear in PageRank's TOP10 or even basic statistical TOP10, yet exhibit high betweenness and closeness centrality, indicating strong information dissemination capabilities.

Figures 6(a) and 6(b) [Figure 6: see original paper] characterize TOP5 leaders from Net10K. High-authority users typically show absolute advantages in single features (followers, answers, influence), while high-hub users demonstrate balanced performance across metrics, indicating strong multi-angle information propagation capacity.

Both algorithms show substantial overlap in identified leaders, as PageRank approximates an ordered superposition of HITS authority and hub values. For example, user Zhang Liang ranks TOP3 in PageRank while placing within TOP15 for both AUTH and HUB in Net10K, proving many users are both authoritative and disseminative.

Algorithm Evaluation

Figure 7 [Figure 7: see original paper] compares convergence rates. Both algorithms converge relatively quickly, with PageRank achieving better fit rates. PageRank's fitting curve rises until 20% of training samples then stabilizes, while HITS converges more slowly, requiring more samples.

Figure 8 [Figure 8: see original paper] correlates PageRank values, AUTH, HUB, upvotes, followers, eigenvector centrality, betweenness centrality, and closeness centrality (labeled 0-7). Zhihu Rank shows moderate correlation with AUTH, eigenvector centrality, and closeness centrality, indicating these measure different influence aspects. Strong correlation with upvotes and followers validates PageRank's importance assessment, as users with high followers/upvotes have greater probability of being followed and answering.

Although AUTH and HUB show moderate correlation, TOP20 rankings exhibit substantial overlap, indicating mined key opinion leaders are both influential and information pathways.

Topic-Specific Key Opinion Leader Analysis

To validate algorithm effectiveness, we analyzed the topic "XX Technology" (company name anonymized for objectivity and neutrality). Tables 12 and 13 show TOP5 leaders for this topic from both algorithms.

The results show TOP5 leaders from both algorithms partially overlap. Users Li Nan, Huang Liang Yi Jiao, and Wong Xu appear in TOP5 for both PageRank and HITS (authority/hub), indicating they are both "authoritative users" and "information disseminators." Table 14 lists Zhihu-certified excellent answerers for this topic, including Li Nan, Hu Jie, and Deng Deng Da Ren, all identified by our algorithms as topic-dependent key opinion leaders, demonstrating strong practicality and accuracy.

Conclusion

This study constructs Zhihu's social and Q&A networks using user relationships, establishes a user influence evaluation model through AHP-based indicator hierarchies, and applies it to improved PageRank and HITS algorithms for mining key opinion leaders. By combining importance metrics from both networks, we overcome the one-sidedness of single-network approaches. The improved algorithms consider user activity, credibility, and influence, enhancing accuracy and objectivity.

Experimental results show the proposed algorithms effectively extract prominent key leaders from Zhihu with fast convergence, reusability, and transferability. Topic-dependent user studies reveal significant commercial value, suggesting the model could drive and guide marketing after continuous refinement. A limitation is the lack of consideration for "water army" [23] impacts, which represents future research focus.

References

[1] Jia Jia, Song Enmei, Su Huan. Answer Quality Assessment in Social Q&A Platforms: A Case Study of "Zhihu" and "Baidu Knows" [J]. Journal of Information Resources Management, 2013, 3(2): 19-28.
[2] Song Hao. Analysis of "Opinion Leader" Characteristics in the Weibo Era [J]. Today's Media, 2010, 11(41): 1.
[3] Wang Xiuli. Research on the Influence Mechanism of Opinion Leaders in Online Communities: A Case Study of the Social Q&A Community "Zhihu" [J]. International Press, 2014, 36(9): 47-57.
[4] Luo Xiaoguang, Xi Lulu. Research on Customer Word-of-Mouth Opinion Leaders Based on Social Network Analysis [J]. Management Review, 2012, 24(1): 75-81.
[5] Peng Lan. Interpersonal Communication in Networks [J]. International Press, 2001, 3(1): 47-52.
[6] Zeng Runxi. Research on Network Public Opinion Control Mechanisms [J].

Library and Information Service, 2009, 53(18): 79-82.
[7] Liu Zhiming, Liu Lu. Identification and Analysis of Opinion Leaders in Weibo Network Public Opinion [J]. Systems Engineering, 2011, 29(6): 8-16.
[8] Guo Bo, Zhao Junrui, Sun Yu. Statistical Characteristics and Dynamics of User Behavior in Social Q&A Communities: A Case Study of Zhihu [J]. Data Analysis and Knowledge Discovery, 2018, 2(4): 48-58.
[9] Guo Bo, Li Shouguang, Wang Hao, et al. Design and Implementation of a Comprehensive E-commerce Review Analysis System: Research and Application of Sentiment Analysis and Opinion Mining [J]. Data Analysis and Knowledge Discovery, 2017, 1(12): 1-9.
[10] YOGANARASIMHAN H. Impact of Social Network Structure on Content Propagation: A Study Using YouTube Data [J]. Quantitative Marketing and Economics, 2012, 10(1): 111-150.
[11] ZHAI Z, Xu H, Jia P. Identifying Opinion Leaders in BBS [C]//Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Piscataway: IEEE Computer Society, 2008: 398-401.
[12] Ding Hanqing, Wang Yaping. Analysis of "Opinion Leader" Characteristics in SNS Space: A Case Study of Douban [J]. Journalism & Communication, 2010(3): 82-91.
[13] Yang Changchun, Wang Tianyun, Ye Shiren. Research on Weibo Opinion Leader Influence Evaluation Index System: Based on Media Influence Perspective [J]. Journal of Intelligence, 2014, 33(8): 178-183.
[14] Kwak H, Lee C, Park H, et al. What is Twitter, a Social Network or a News Media? [C]//Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010.
[15] Xie Y, Huang TZ. A Model Based on Cocitation for Web Information Retrieval [J]. Mathematical Problems in Engineering, 2014: 1-6.
[16] Langville AN, Meyer CD. Google's PageRank and Beyond: The Science of Search Engine Rankings [M]. Princeton: Princeton University Press, 2011.
[17] Mao Guojun, Xie Songyan, Hu Dianjun. Improved PageRank Model and Weibo User Influence Mining Algorithm [J]. Computer Applications and Software, 2017, 34(5): 28-32.
[18] Miller JC, Rae G, Schaefer F, et al. Modifications of Kleinberg's HITS Algorithm Using Matrix Exponentiation and Web Log Records [C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2001: 444-445.
[19] Lin Y, Li H, Liu X, et al. Hot Topic Propagation Model and Opinion Leader Identifying Model in Microblog Network [C]//Abstract and Applied Analysis. London: Hindawi Publishing Corporation, 2013.
[20] Xing W, Ghorbani A. Weighted PageRank Algorithm [C]//Conference on Communication Networks and Services Research. Piscataway: IEEE, 2004: 305-314.
[21] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford: Stanford InfoLab, 1999.
[22] Xiong Tao, He Yue. Identification and Analysis of Opinion Leaders in

Weibo Forwarding Networks [J]. New Technology of Library and Information Service, 2013(6): 55-62.
[23] Guo B, Wang H, Yu Z, et al. Detecting Spammers in E-commerce Website via Spectrum Features of User Relation Graph [C]//Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference on. Piscataway: IEEE, 2017: 324-330.

Author Contributions

Guo Bo: Proposed research ideas, designed research framework, algorithm development, paper revision.
Xu Haodi: Conducted Zhihu key leader mining experiments and analysis, paper writing.
Lei Shuiwang: Paper revision.

Analysis of User Influence and Identification of Key Opinion Leaders Based on Zhihu Platform

Guo Bo[1], Xu Haodi[2], Lei Shuiwang[3]
[1] Meizu Telecom Equipment Co., Ltd. Beijing Branch, Beijing 100872
[2] The Hong Kong University of Science and Technology, Hong Kong 999077
[3] Xuzhou College of Industrial Technology Library, Xuzhou 221140

Abstract: [Purpose/significance] With the rapid development of network technology, the platform of Zhihu has become a significant carrier to discuss social public topics and share knowledge as well as specified experience. Therefore, it is of importance for studying the communication channels of social network information to investigate the influence of key users and dig out the key opinion leaders in the Zhihu platform. [Method/process] By the means of improved PageRank and HITS algorithms, this study constructed a model for evaluating user influence based on the social network and question-answering network of Zhihu platform, and identified the key users and opinion leaders accurately and objectively. [Result/conclusion] The experimental results show that PageRank and HITS algorithms in this paper could effectively extract several key opinion leaders with prominent features in Zhihu platform, the speed of the convergence is fast and with high reusability and mobility. By processing and analyzing the user dataset of Zhihu platform, we successfully build a model for evaluating the user influence and mining key opinion leaders. Along with the verification of specified topics, it can be inferred that this model has enormous application value and commercial promotion prospect.

Keywords: Zhihu, user influence, key opinion leader, PageRank algorithm, HITS algorithm

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