Zhihu Information Origin Model and Credibility Assessment Postprint

Authors: Zhang Ting, Qi Xianghua

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] This study constructs a PROV data provenance model for the Zhihu information dissemination process and develops user credibility evaluation metrics to quantify Zhihu information credibility, thereby enriching and improving methods for assessing information credibility on social Q&A community platforms. [Method/Process] Taking Zhihu as the research object, we introduce the concept of data provenance from the perspective of information dissemination processes to evaluate information credibility. By establishing a PROV data provenance model for Zhihu, we trace and record the source and dissemination process of Zhihu information, combining it with user credibility scores involved in the information dissemination process to calculate quantitative results for Zhihu information credibility. [Results/Conclusion] Through evaluating Zhihu information credibility, this study further refines information credibility assessment methods and provides new insights for optimizing community information quality.

Full Text

Preamble

ChinaXiv Cooperative Journal
Vol. 63, No. 9, May 2019

Zhihu Information Provenance Model and Credibility Evaluation

Zhang Ting, Qi Xianghua
School of Economics and Management, Shanxi University, Taiyuan 030006

Abstract

[**Purpose/Significance**] This study constructs a PROV data provenance model and user credibility evaluation metrics for the Zhihu information dissemination process, quantifies the credibility of Zhihu information, and enriches

methods for assessing information credibility on social Q&A community platforms. [**Method/Process**] Taking Zhihu as the research object and adopting a data provenance perspective to evaluate information credibility from the standpoint of information dissemination, this paper establishes a PROV data provenance model for Zhihu to trace and record the origin and dissemination path of information. Combined with credibility scores of users involved in the dissemination process, quantitative results for Zhihu information credibility are calculated. [**Result/Conclusion**] Through evaluating Zhihu information credibility, this research further improves information credibility assessment methods and provides new insights for optimizing community information quality.

Keywords: data provenance; PROV model; Zhihu information; credibility
Classification Number: G252
DOI: 10.13266/j.issn.0252-3116.2019.09.009

Introduction

With the development of the Internet, the fundamental direction of information flow has changed amid explosive and disordered information growth. Users' purposefulness in actively acquiring information has strengthened, and their methods are no longer limited to search engines but now include online questioning and community discussions to obtain experience, information, and knowledge from peers and experts. As a social Q&A community website where users share knowledge, experience, and insights, Zhihu had over 100 million registered individual users and 26 million daily active users as of September 2017, with an average daily visit duration of one hour and monthly page views reaching 18 billion [1]. According to the 41st "Statistical Report on China's Internet Development" released by CNNIC, Zhihu's user adoption rate increased from 7.0% in 2016 to 8.8% by December 2017 [2].

Information credibility generally refers to the degree to which information or information sources are trusted by recipients, interpreted here as users' perception or evaluation of the trustworthiness of information sources or dissemination media and the authority of information publishers. Currently, scholars' research on information credibility assessment in Zhihu and other Q&A communities primarily focuses on two independent perspectives—users and information text content—and on two aspects: information quality feature extraction based on classification features and information quality assessment impact modeling [3]. However, research evaluating information credibility from the perspective of information generation and dissemination processes is lacking. Data provenance can help track data sources, record dynamic information generated during data processing, assess data quality and credibility, and enable users to clearly understand the full story of information to decide whether to adopt it based on their needs.

The uncontrolled changes in community content have seriously impacted

Zhihu's requirements and positioning for "high-quality" communities. Meanwhile, Zhihu's multi-dimensional insights and multi-angle analyses on hot topics have made it a breeding ground for public opinion and an important node for secondary dissemination of viewpoints, exerting certain influence on the depth of user discussion and the direction of event development. The sheer number of users is not necessarily the main cause of declining content quality; rather, information overload caused by high-speed information flow is a key factor leading to uneven content quality. High-quality information content can provide high-value knowledge resources for users and even the entire Internet. Therefore, how to identify credible answers from vast amounts of Zhihu information has become an urgent problem, making research on the credibility of Zhihu Q&A information crucial.

This study takes Zhihu as the research object, employs data provenance methods to evaluate Zhihu information credibility, assesses online Q&A community information credibility from the perspective of information dissemination processes, and provides new ideas for related research. By analyzing Zhihu information dissemination scenarios and constructing a PROV data provenance model, the study captures network behavior data of Zhihu users involved in the information dissemination process, estimates information providers' credibility by assigning trust scores, and calculates Zhihu information credibility scores based on dissemination paths. Based on these trust scores, users can make more informed decisions about whether to use the information.

1. Research Status

Unlike traditional search engine-based Q&A communities, Zhihu introduces social relationships, with interactive Q&A at its core to further promote knowledge sharing and interaction, making it a true social Q&A community platform. To date, domestic research on social Q&A community platforms has mostly focused on platform characteristics, development models, and user behavior, with limited research on platform information quality and credibility. Scholars such as Jiang Wen, Wang Ping, and Li Baozhen [4-6] have reviewed and summarized domestic and international research on network information quality evaluation.

Based on a review of relevant literature, Q&A community information quality assessment primarily studies two aspects: information quality feature extraction based on classification features and information quality assessment impact modeling.

1.1 Information Quality Feature Extraction Based on Classification Features

Related research using this method includes: J. Jeon [7] analyzed 1,700 answers from Naver Q&A using maximum entropy algorithms, examining answerer adoption rates, expertise, answer length, quantity, and copy frequency. E. Agichtein [8] used stochastic gradient boosted decision trees to extract features from 6,665

questions and 8,366 answers on Yahoo!. C. Shah et al. [9] employed logistic regression to analyze features of 120 questions and 600 answers on Yahoo! Quest. Q. Tian et al. [10] used random forests to analyze 103,793 questions and 196,145 answers on Stack Overflow. Lai She'an [11] used SVM algorithms to extract features from 12,707 questions and 463,114 answers on Baidu Zhidao. Wang Wei [12] used logistic regression, support vector machines, and random forests to construct a feature system based on 20 topics, 200,000 questions, 924,266 answers, and 158,007 user data points from Zhihu.

1.2 Information Quality Assessment Impact Modeling

S. Oh et al. [13] invited people with different functions to evaluate accuracy, completeness, relevance, objectivity, source reliability, and other indicators of health-related answers on Yahoo! Answers under selected evaluation criteria. Domestic scholar Jia Jia et al. [14] used questionnaires to compare and score answer quality on Zhihu and Baidu Zhidao platforms. P. Fichman [15] manually compared answer quality on four Q&A sites—Wikipedia Reference Desk, WikiAnswers, Yahoo! Answers, and Askville—from three aspects: accuracy, completeness, and verifiability, concluding that high-traffic Q&A sites do not necessarily correlate with answer quality. S. Kim et al. [16] analyzed user comments when selecting best answers on Yahoo! Answers, finding that evaluation patterns differ by topic category and that users primarily select best answers based on social emotion, content, and utility. Li Jing [17] and Sun Xiaoning [18] both used structural equation modeling to construct information quality models for Baidu Zhidao and SQA system answer quality models. Cao Gaohui et al. [3] used questionnaires to study users' perception of answer information quality on social Q&A platforms, constructing an external model of answer quality perception. Shi Guoliang [19] took Zhihu as the research object, proposed hypotheses from answer characteristics and answerer characteristics, tested them using content analysis and regression models, and concluded that answer length, timeliness, and answerer influence positively affect answer acceptance, with answerer characteristics having greater impact than answer characteristics.

In summary, information quality feature extraction based on classification features is a research method suited to the era of big data and represents the trend of automated computer evaluation, offering high efficiency and strong analytical power but struggling to handle subjective user emotional attitudes. Information quality assessment impact modeling compensates for machines' difficulty in understanding subjective issues by obtaining evaluation factors from different populations to establish information quality models. However, this method only reflects conditions of small data samples, and assessment results are influenced by subjective factors such as evaluators' information literacy, problem awareness, and expertise.

2. Data Provenance Model

PROV is a provenance standard document collection consisting of 12 documents defined by the W3C Provenance Incubator Group, with the PROV Conceptual Data Model (PROV-DM) at its core. It aims to trace source data needed on the web, describing dynamic information such as the generation of "entities," the occurrence of "activities," and the responsibility of "agents" to achieve data traceability and standardized expression. The core architecture includes three types—entity, activity, and agent—and seven relationships and derivation relationships among them. Based on activity scenario characteristics and needs, it provides further description methods for relationships between entities, activities, and agents. Specific extension clauses are detailed in the PROV documentation.

Changes between entities are represented through the property `prov:wasDerivedFrom`. To more detailedly describe relationships between entities in different scenarios, it further provides general attributes, such as three sub-properties of derivation: `prov:wasQuotedFrom` indicates quoting larger books, atlases, or networks to create new entities where the new entity repeats part or all of the original entity [20]; `prov:wasRevisionOf` indicates that the derived entity contains substantial content from the original entity, such as different versions of a book; `prov:hadPrimarySource` indicates that the new entity's content mainly comes from experience or knowledge in a certain topic or original entity. `prov:specializationOf` and `prov:alternateOf` describe relationships at an abstract level—the former connects entities with detailed content or greater explanatory power to simple, summarized general entities, while the latter represents linking entities that supplement other attributes of the same thing or topic. The property `prov:wasInfluencedBy` associates entities, activities, or agents that are influenced or affect their characteristics. The generation time and invalidation time of entities are represented by properties `prov:generatedAtTime` and `prov:invalidatedAtTime`, respectively.

Activities have temporal sequences, represented by the property `prov:wasInformedBy`; their start and end times are described by properties `prov:wasStartedBy` and `prov:wasEndedBy`. Relationships between agents use the property `prov:actedOnBehalfOf`. There are two relationships between entities and activities: `prov:used` and `prov:wasGeneratedBy`, indicating that an entity was used by an activity and that an entity was generated by an activity. The attribution relationship between an agent and an entity is expressed by the property `prov:wasAttributedTo`; the association relationship between an agent and an activity is expressed by the property `prov:wasAssociatedWith`.

[Figure 1: see original paper] PROV Core Architecture

3. Zhihu Information Credibility Assessment Analysis

Zhihu's core content comes from its Q&A framework, which is topic-based and question-centered—all answers must focus on the question itself. Therefore,

Zhihu employs an "agree-disagree" mechanism for user judgment. Users can interact through agreeing, thanking, collecting, commenting, following, private messaging, tipping, sharing, and disagreeing. Based on post-evaluation scores, answers with higher recognition are placed in more prominent positions, with each answer displaying its number of likes to provide users with intuitive quality judgment. For inappropriate content, Zhihu verifies through "system-assisted determination + manual review," using AI anti-spam systems named "Wali" and "Wukong" for real-time screening of erroneous content, invalid answers, false information, and spam advertising. For content that algorithms cannot handle, manual review determines whether the answer addresses the question, and off-topic or non-compliant content is collapsed and typically displayed at the bottom of the answer list.

The above represents intuitive judgments based on users, community managers, and algorithms. If an answer originates from an authoritative organization, its credibility is higher; if from an ordinary user, how should credibility be judged? If a user modifies and supplements their answer based on others' comments, can that answer be considered more credible? If a user's answer content has been reposted multiple times, and reposted content typically contains others' judgments and cognition, is that answer more credible? Therefore, recording information sources, changes, and dissemination processes becomes valuable. This paper utilizes the PROV provenance model's focus on how activities affect resources [22] to trace information sources and different versions of content changes by describing information chains of all activities creating content and users. Since answer generation is a process where information publishers externalize their tacit knowledge based on their understanding, information publishers are directly related to answer quality. Thus, two important factors affecting "high quality" in content communities are the information dissemination process and user relationships. Based on these identified influencing factors, corresponding evaluation methods are selected:

(1) For the information dissemination process, capture provenance information of selected Zhihu information, store and verify it, and establish a PROV model to determine information dissemination paths.

(2) Evaluate the credibility of users involved in the dissemination process, determine user credibility assessment metrics based on Zhihu platform characteristics and user behavior features, and calculate quantitative results.

(3) Combine dissemination paths with user credibility values to calculate Zhihu information credibility scores.

3.1 Zhihu Information Dissemination Path Analysis

3.1.1 Zhihu Information Dissemination Scenario Analysis Based on the above analysis, Zhihu has the following scenario (see [Figure 3: see original paper]): Select a question from Zhihu Roundtable Topic1. User user_A

answered answer_a at 12:18 on February 5, 2018. Meanwhile, user user_B provided answer_b by citing an article they previously published, attaching a link to the cited article at the end. Users user_C and user_D gave comments comment_a and comment_b on answer_b, both selected as featured comments, while user user_D liked answer_b. On February 6, 2018, user user_E provided answer_c with a different viewpoint from previous users, excerpted from relevant professional papers. User user_F gave comment_c, and user user_G provided answer_d based on the question and comment_c. User user_A edited and modified answer_a based on comment_a and comment_b, producing answer_{a1}, and further edited the answer based on answer_d on February 11, 2018, producing answer_{a2}.

3.1.2 Zhihu's PROV Data Provenance Model According to data provenance model description rules, provenance resources are identified by unique URIs, and namespaces are identified by URIs. This paper uses the example namespace `http://example.org/` to identify resources. Therefore, entities, activities, and agents involved in Zhihu information dissemination are represented with abbreviated namespace prefixes.

(1) **Entity (entity)**: In the PROV provenance model, this refers to objectively existing, conceptual things of various types. For Zhihu information dissemination, entities mainly include answers, articles, comments, and reposted information links.

(2) **Activity (activity)**: In the PROV provenance model, this describes how entities maintain their state and generate new entities by changing entity attributes, such as actions and processes. Activities in Zhihu information dissemination mainly include answer publishing, editing, commenting, and liking.

(3) **Agent (agent)**: In the PROV provenance model, this bears responsibility for activities, entities, and roles. It can be individuals, organizations, or inanimate objects like software. An activity may be associated with several agents, and related entities also have causal relationships with agents. Agents in Zhihu information dissemination refer to Zhihu users.

(4) **Derivation and Revision (derivation and revision)**: In the PROV provenance model, this describes relationships between entities where one entity derives another, with the latter's content, attributes, and existence originating from the former. In Zhihu, users provide new answers based on previous answers, related comments, and their own experience or expertise. Revision is a special type of derivation—as time passes, users make multiple edits to previous answers. In the PROV provenance model, each edited version is considered a new entity.

(5) **Time (time)**: In the PROV provenance model, this describes activity generation and end times, further explaining relationships between activities and entities. In Zhihu information dissemination, this mainly manifests

as answer content publication start times and editing end times. More recent answers may be modified or relatively new ideas, with relatively higher credibility.

Descriptions of objects involved in Zhihu information dissemination are shown in .

Descriptions of Objects Involved in Zhihu Information Dissemination Process

Based on the Zhihu information dissemination scenario, a complete data provenance description can be obtained using the PROV provenance model, shown in [Figure 4: see original paper]. Entities are represented by ellipses, activities by rectangles, and agents by pentagons according to W3C data provenance model description specifications. Agent credibility is indicated by `prov:userTrust`. Arrows and unboxed text represent and describe usage and generation relationships among the three, with arrow directions pointing from future to past. Agents are responsible for activities or entities, with arrow directions pointing from activities to agents. Connections between agents and activities use `wasAssociatedWith`, and connections between agents and entities use `wasAttributedTo`.

[Figure 4: see original paper] Zhihu Information Dissemination PROV Provenance Model

3.2 User Credibility Assessment

Zhihu's core users provide rare experiences and unique problem-solving approaches that constitute high-quality content, representing an important force in enhancing Zhihu's core competitiveness. Therefore, Zhihu requires users to provide complete personal information and adopts real-name authentication. It increases answer credibility by counting personal achievements such as agreements, thanks, and collections obtained during usage. Due to social attributes between users, the higher the frequency of users posting answers, ideas, and comments, the faster their information exchange rate with other users, and the higher their personal trustworthiness. Users with more followers can have their answers read, agreed with, and commented on by more people, increasing their interaction capability. Strong interaction capability between users and followers increases credibility with influence. Users can also follow other users to timely obtain useful information and enrich and improve their own knowledge reserves and answer quality. Therefore, this paper evaluates user credibility from four aspects: user information completeness, user authentication and achievements, user activity, and user social breadth [23].

3.2.1 User Information Completeness This metric specifically includes profile picture, location, education information, occupation information, and personal introduction in the Zhihu community. Higher profile completeness leads to greater public trust. User basic information is represented by vector Y = (x_1, x_2, …, x). The formula for calculating user information completeness is:

x（i）= 0，& x（i）为无效信息 1，& x（i） 为有效信息 UI（u）= i=1 x（i）

Formula (1) indicates whether the i-th item contains valid information. UI(u) represents the user information completeness function.

3.2.2 User Authentication and Achievements The Zhihu platform has certain requirements for user information authenticity, which is also an important condition for evaluating user credibility. Zhihu has established blue authentication mechanisms for individuals and organizations. With the influx of large numbers of users and increasing answer volumes, to help users more efficiently identify experts and content in various fields, Zhihu has implemented an "Excellent Answerer" identification (orange badge) for users who have created substantial professional content under specific topics [24]. Excellent answerers are calculated by the system through topic weight, and the only way to increase one's weight in a Zhihu topic is to publish high-quality answers in that field. Compared with blue authentication, orange authentication better represents user credibility. Achievements refer to agreements, thanks, and collections from other users, calculated as follows:

UL=fj+A（approve（n）+thanks（n）+collection

Different authentication types have different credibility rating scores, as shown in Formula (3) (user authentication levels and achievement weight coefficients are shown in), where f is the authentication level score for type j users, n is the total number of agreements, thanks, and collections from other users on all answers, and A is the influence weight coefficient of user achievements on credibility, set using the product scaling method.

User Authentication Level and Achievement Weight Coefficients

3.2.3 User Activity This refers to the number of likes, answers, ideas published, and participation in community public editing within a certain period. Public editing is users' active participation in improving and optimizing Zhihu's public content. Therefore, higher user activity indicates higher community participation and relatively higher credibility. The formula is:

i=1（publish（i）+approve（i）+idea（i））+Bed

In Formula (4), publish(i), approve(i), and idea(i) respectively refer to the number of questions answered, agreements given, and ideas published by Zhihu users on day i. T is a set time period. ed refers to the number of times Zhihu users participated in public editing, with trust weight coefficient B set at 0.44.

3.2.4 User Social Breadth The more followers a user has, the greater the possibility of information diffusion and influence. If followers have higher reputation, the user's credibility is higher. Therefore, follower count affects user social breadth more than the number of users they follow.

UD=Clev（fans）+Dlev（followers）

In Formula (5), fans refers to the number of users following the user; followers refers to the number of users they follow; Lev(n) represents the authentication level of users' followers or followees, where larger n indicates higher level; C and D are trust weight coefficients, set at 0.32 and 0.24 respectively.

Based on the above four user indicators, user credibility (UserTrust) is calculated as follows:

```
UserTrust = UI + UL + UC + UD
```

3.3 Zhihu Information Credibility Calculation

By tracking the origin and content changes during dissemination through the PROV data provenance model, the information dissemination path can be clearly obtained. Combined with user credibility values at nodes on the dissemination path, Zhihu information credibility is calculated.

Formula (7) calculates Zhihu information credibility, where n is the number of users involved in the dissemination path of an answer, userTrust(i) is the credibility value of the i-th user, and Trust(answer) is the Zhihu information credibility.

```
Trust(answer) = i=1 userTrust
```

4. Experiment Design and Analysis

Based on the previously described Zhihu information dissemination scenario, corresponding provenance information was collected and described according to RDF statement specifications required by the provenance model, then stored and verified to ultimately establish a PROV provenance model. User credibility in the dissemination process was analyzed based on previously established metrics, and Zhihu information credibility evaluation values were calculated by combining the data provenance-based Zhihu information dissemination process.

4.1 Provenance Information Preprocessing

Resource Description Framework (RDF) is a W3C standard for describing web resources and has effectively become the standard description method for PROV models [25]. This paper describes collected information using RDF triples. Partial RDF descriptions are as follows:

```
ex:publishprov:wasAssociatedWithex:user_A
ex:compileprov:wasAssociatedWithex:user_A
ex:commentprov:wasAssociatedWithex:user_C
ex:approveprov:wasAssociatedWithex:user_D
ex:commentprov:usedex:answer_b
ex:approveprov:usedex:answer_b
ex:approveprov:usedex:answer_c
ex:answer_a2prov:wasGeneratedByex:compile
```

ex:answer_bprov:wasGeneratedByex:publish

By collecting and recording Zhihu information dissemination path information, the Octopus Collector was used to collect personal information and behavioral dynamics of seven evaluation users involved in dissemination path nodes. The collection content was based on the four user credibility evaluation indicators and information that could clearly reflect or indirectly 挖掘 user credibility. Among them, the number of questions answered, agreements given, and ideas published were collected within one month, as shown in .

Actual Values of User Information Indicators

4.2 User Credibility Calculation

To reduce the negative impact of different measurement units and numerical magnitude differences on user credibility calculation results, this paper uses a standardization method for dimensionless preprocessing of raw data (actual indicator values) before calculating credibility-related indicators. The formula is:

xi−珈x

In Formula (8), x represents the actual indicator value, $\bar{x}$ is the mean indicator value, s is the standard deviation of indicators, and y represents the indicator evaluation value. After preliminary data processing and calculation of the four evaluation indicators, the quantitative results of evaluation indicators and user credibility values are obtained, as shown in .

User Credibility-Related Indicator Scores

4.3 Zhihu Information Assessment Results and Analysis

According to [Figure 4: see original paper], the Zhihu information dissemination PROV provenance model can track Zhihu information answer_{a2}. User user_A revised the existing answer answer_{a1} based on comments from user user_C and user_D on February 11, 2018, generating answer_{a2}, thus obtaining the complete generation process of this answer. Therefore, the credibility calculation formula for answer_{a2} is:

Trust(answer_a2) = i=1userTrust

By calculating user credibility values from four indicators—user information completeness, personal authentication and achievements, activity, and social breadth—and combining them with the PROV data provenance model, Zhihu information credibility is calculated through formulas. The credibility results sorted from high to low are shown in .

Zhihu Information Credibility Calculation Results

Although information has timeliness and lag, for this experimental sample, time has relatively less impact on information credibility compared to user credibility.

If time acts on users, causing their personal cognition to continuously optimize and improve with time and other users' influence, then time has a certain positive impact on information credibility. The publisher of answer_a, user_A, had neither personal nor organizational authentication from Zhihu platform, and their activity and social breadth were relatively low. The credibility of their published content answer_a was lower than content published by other users. However, as user cognition changed, time was continuously updated, and other users' influences led to two revisions, resulting in answer_{a2}. Its information credibility value increased from 0.348 to 0.560. This demonstrates that constructing a data provenance model to evaluate Zhihu information credibility is effective.

The calculation result ranking shows that information content credibility has a positive relationship with publishers' authentication and achievements on the Zhihu platform. As an authoritative and professional organization, user_B has the highest user credibility value. Their published information content answer_b references authoritative articles published internally by their organization. The calculation results show answer_b has the highest credibility, which matches actual conditions and aligns with the ranking given by Zhihu's algorithm and "agree-disagree" mechanism.

This study evaluates Zhihu platform information credibility from information dissemination process and user perspectives, constructing a credibility assessment framework. Using data provenance methods, relevant provenance information was collected based on identified Zhihu information dissemination scenarios to establish a PROV provenance model described using RDF. By calculating user node credibility on dissemination paths, Zhihu information credibility was obtained. The research results provide new ideas for improving information credibility assessment methods and optimizing social Q&A community information quality. This study is a preliminary attempt with certain limitations in Zhihu information selection scope and types. The complexity of user relationships and different topic types may cause evaluation differences. Future work can expand the quantity and types of evaluation objects to improve assessment accuracy. Additionally, the weighting scheme formulated based on Zhihu user characteristics has subjective factors; subsequent research can consider more influencing factors to optimize weighting and further improve information credibility assessment results. How to display and manage large amounts of provenance information using the PROV model will become a future research focus.

References

[1] Zhihu: As of September, Zhihu's total individual registered users exceeded 100 million [EB/OL]. [2018-12-24]. http://www.sohu.com/a/193351816_{812860}.
[2] CNNIC. 41st Statistical Report on China's Internet Development [EB/OL]. [2018-01-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf.
[3] Cao Gaohui, Hu Zijing, Zhang Yuxuan, et al. Research on social Q&A platform information quality perception model based on external clues [J].

Information Science, 2016, 34(11): 122-128. [4] Jiang Wen, Xu Xin. Review of online Q&A community information quality evaluation research [J]. Modern Information, 2014, 30(6): 41-50. [5] Wang Ping, Cheng Qikai. Research progress and review of network information credibility evaluation [J]. Journal of Information Resources Management, 2013, 3(01): 46-52. [6] Li Baozhen, Wang Ya. Review of network information credibility evaluation research under social media environment [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(12): 1314-1321. [7] Jeon J, Croft W B, Lee J H, et al. A framework to predict the quality of answers with non-textual features [C]//Proceedings of the 2008 international conference on Web search and data mining (WSDM'08). New York: ACM, 2008: 183-194. [8] Agichtein E, Castillo C, Donato D, et al. Finding high-quality content in social media [C]//Proceedings of the 2008 international conference on Web search and data mining. New York: ACM, 2008: 183-194. [9] Shah C, Pomerantz J. Evaluating and predicting answer quality in community QA [C]//International ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2010: 411-418. [10] Tian Q, Zhang P, Li B. Towards predicting the best answers in community-based question-answering services [EB/OL]. [2018-09-15]. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6096/6334. [11] Lai She'an, Cai Zhongmin. Answer quality evaluation method based on similarity in Q&A communities [J]. Computer Applications and Software, 2013, 30(2): 266-269. [12] Wang Wei, Ji Yuqiang, Wang Hongwei, et al. Evaluation of answer quality in Chinese Q&A communities: A case study of Zhihu [J]. Library and Information Service, 2017, 61(22): 36-44. [13] Oh S, Worrall A, Yi Y J. Quality evaluation of health answers in Yahoo! Answers: a comparison between experts and users [J]. Proceedings of the American Society for Information Science & Technology, 2012, 48(1): 1-3. [14] Jia Jia, Song Enmei, Su Huan. Answer quality assessment on social Q&A platforms: A case study of "Zhihu" and "Baidu Zhidao" [J]. Journal of Information Resources Management, 2013, 3(2): 19-28. [15] Fichman P. A comparative assessment of answer quality on four question-answering sites [J]. Journal of information science, 2011, 37(5): 476-486. [16] Kim S, Oh J S, Oh S. Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective [C]//Proceedings of the 70th annual meeting of American Society for Information Science & Technology. Silver Spring: American Society for Information Science and Technology, 2012: 1-3. [17] Li Jing. Research on virtual community information quality modeling and perceived difference comparison [D]. Wuhan: Wuhan University, 2013. [18] Sun Xiaoning, Zhao Yuxiang, Zhu Qinghua. Construction of social search answer quality evaluation indicators based on SQA system [J]. Journal of Library Science in China, 2015, 41(4): 65-82. [19] Shi Guoliang, Chen Xu, Du Lufeng. Research on factors influencing answer acceptance in social Q&A websites: A case study of Zhihu [J]. Modern Information, 2016, 36(6): 41-45. [20] W3C. PROV-O: the PROV ontology [EB/OL]. [2018-01-11]. http://www.w3.org/TR/2013/REC-prov-o-20130430/. [21] Yan Hao. Serious people always exist: This may be the most sincere sharing about Zhihu [EB/OL]. [2018-03-28].

https://www.huxiu.com/article/147187/1.html?f=index_{{{feed}}}_{{article}}}. [22] Ni Jing, Meng Xianxue. Comparative study of data provenance description languages in linked data environment [J]. Modern Library and Information Technology, 2013(2): 18-23. [23] Liu Qingsong. Research on Chinese microblog information credibility analysis methods [D]. Beijing: Beijing Information Science and Technology University, 2015. [24] Zhihu. What is Zhihu's "Excellent Answerer" identification? [EB/OL]. [2017-12-23]. https://www.zhihu.com/question/48509984. [25] Ni Jing, Meng Xianxue. PROV data provenance model and Web applications [J]. Library and Information Service, 2014, 58(3): 13-19.

Author Contribution Statement

Zhang Ting: Determined the research topic, responsible for research framework design, data collection and organization, and paper writing.
Qi Xianghua: Responsible for guiding research content.

Credibility Evaluation and PROV Model of Zhihu Information

Zhang Ting, Qi Xianghua
School of Economics and Management, Shanxi University, Taiyuan 030006

Abstract: [Purpose/significance] This paper aims to construct a PROV provenance model and user credibility evaluation index for the information dissemination process, quantify the credibility of information, and enrich and improve methods for evaluating information credibility on social Q&A community platforms. [Method/process] The paper analyzed the credibility of data origin concept assessment from the perspective of information dissemination, traced and recorded the source and dissemination of information by establishing the relevant PROV data provenance model. The process, combined with the user credibility scores involved in the information dissemination process, was used to calculate the quantitative results of the credibility of the information. [Result/conclusion] Through the evaluation of the credibility of information, the information credibility evaluation method is further improved, which provides a new idea for optimizing the quality of community information.

Keywords: data provenance; PROV-O; Zhihu information; credibility

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