Evolutionary Analysis of Topics and Topic Clusters in Informal Information Exchange from a Conversation Analysis Perspective (Postprint)

Authors: Wang Xiao, Ma Chao, Zhai Shanshan

Date: 2023-04-01T16:02:55+00:00

Abstract

[Purpose/Significance] Aiming at the limitations of current research on topic evolution in informal information exchange regarding analytical hierarchy and measurement indicators, this study proposes a generalizable evolutionary analysis method to investigate the characteristics and patterns of topic evolution from micro and meso perspectives. [Method/Process] By introducing conversation analysis theory and taking Sina Weibo and Zhihu as examples, this study reveals the evolutionary characteristics and patterns of informal information exchange through analyzing the operational processes of topics and topic clusters from two dimensions—conversation content and discussion mode. Simultaneously, it designs a computational determination method for topic persistence to enrich the measurement criteria for topic evolution. [Results/Conclusion] The topic evolution analysis results demonstrate that opinion groups on Sina Weibo and Zhihu exhibit distinct biases in their posting topics, and reveal the primary perspectives through which these groups engage in discussions on social focal events. The topic cluster evolution analysis discovers that opinion groups on Sina Weibo tend to diverge and explore diverse topics within a certain scope, while those on Zhihu consistently focus on core topics. The differences in conversation content and discussion mode between opinion groups on the two social media platforms signify the distinct roles that Sina Weibo and Zhihu play in informal information exchange within the online environment.

Full Text

Preamble

Evolutionary Analysis of Topics and Topic Clusters in Informal Information Exchange from the Perspective of Conversation Analysis

Wang Xiao[1], Ma Chao[2], Zhai Shanshan[1] [1] School of Information Management, Central China Normal University, Wuhan 430079 [2] College of Economics and Management, Zhejiang Normal University, Jinhua 321004

Abstract: [Purpose/Significance] Aiming at the limitations of current informal information exchange topic evolution research in both analysis level and measurement indicators, this study proposes a universal evolution analysis method to explore topic evolution characteristics and patterns from micro and meso perspectives. [Method/Process] Introducing conversation analysis theory and taking Sina Weibo and Zhihu as examples, this paper reveals the evolution characteristics and patterns of informal information exchange from two dimensions: conversation content and discussion style through analyzing the running processes of topics and topic clusters. Meanwhile, a computational method for topic continuity determination is designed to enrich the measurement criteria for topic evolution. [Result/Conclusion] Topic evolution analysis shows that opinion groups on Sina Weibo and Zhihu exhibit obvious biases in posting topics, indicating the main perspectives through which these groups engage in discussions of social focus events. Topic cluster evolution analysis reveals that Weibo opinion groups tend to diverge and explore diverse topics within a certain range, while Zhihu opinion groups consistently focus on core topics. The differences between the two social media platforms in conversation content and discussion style suggest their distinct roles in informal information exchange within the online environment.

Keywords: informal information exchange; topic evolution; topic cluster; conversation analysis **Classification Number:** G203 **DOI:** 10.13266/j.issn.0252-3116.2021.17.009

With the development of internet technology and the impact of the COVID-19 pandemic, informal information exchange has largely migrated to and become increasingly active on social media platforms. The vast amount of user-generated content (UGC) that is massively produced, virally disseminated, and rapidly updated on social media contains diverse topics that are often used to characterize users' content preferences [1]. Based on the trace data objectively recorded by social media, comprehensively depicting topic evolution trends and deeply exploring topic evolution patterns can help accurately grasp the characteristics and patterns of informal information exchange. When applied to specific contexts, this can provide service references for intelligent public opinion monitoring and personalized content recommendation.

Conversation analysis research provides a sociological theoretical foundation for informal information exchange analysis, but its application in social media is still in its infancy. Based on this, this paper introduces conversation analysis theory to parse the evolution processes of topics and topic clusters, combined with the conceptual connotation, measurement, and determination criteria of topic continuity, aiming to deeply explore the evolution process and characteristic patterns of informal information exchange from micro and meso levels, and to provide references for optimizing communication strategies in network pub-

lic opinion management on social media platforms. In the empirical analysis, taking Sina Weibo and Zhihu platforms as data sources, UGC is regarded as an asynchronous conversation process based on social media for analysis.

2 Related Research

2.1 Topic Identification and Evolution in Informal Information Exchange

Social media platforms provide an ideal environment for studying users, interaction relationships, and information flows in informal information exchange [2]. Therefore, research on informal information exchange in online environments often uses Sina Weibo, Zhihu, Twitter, Facebook, and online forums as information exchange carriers and empirical data sources. Topic identification research can be mainly divided into three categories based on representation methods: Word-based topic representation, represented by topic identification methods based on weighted algorithms, which combine word frequency statistics with part-of-speech [3], inverse document frequency [4], etc., to calculate word contribution and extract topic content through ranking and screening; Word-cluster-based topic representation, represented by text clustering-based topic identification methods, which mostly use Word2vec to construct feature word sets combined with K-means clustering algorithms [5-6] to extract text topics; Probability distribution-based topic representation, which uses topic models to identify text topics. Among them, the LDA model is applied in multiple studies to identify short text topics on social media due to its excellent data dimensionality reduction and implicit semantic mining capabilities [7-8]. Additionally, a large number of non-textual features on social media, such as user, geographic, interaction, and temporal features, have been introduced into topic models [9] or combined with text content features to build hypernetwork models [10] for joint topic mining.

Topic evolution research can be mainly divided into two categories based on evolution structure: linear topic structure evolution and nonlinear topic structure evolution. The former dominated early topic evolution research [11], mainly presenting the linear evolution of topic content or discussion intensity on the timeline to reveal the temporal change characteristics and patterns of text content represented by topics. The latter has gradually increased in recent years [12], using storyline analysis [13-14] to explore the evolution process of relationships between topics. Regarding the development stages of topic evolution, related studies typically propose various division methods such as three-stage [15], four-stage [16], and five-stage [17] theories based on life cycle theory. In terms of topic evolution analysis dimensions, different studies expand single temporal dimensions by introducing spatial [18], user [19] dimensions, or enrich them through multi-dimensional feature integration [20].

In summary, topic identification research is rich and provides methodological and technical support for text topic analysis. Topic evolution research has been

deepened and expanded in various aspects such as evolution structure, development stages, and analysis dimensions, but still has certain limitations: The analysis level is relatively macro, focusing on the content evolution of informal information exchange in specific contexts represented by topics, while lacking micro and meso-level analysis of the internal operation of individual topics and topic clusters composed of several closely related topics during evolution. The measurement indicators for topic evolution focus on discussing topic intensity quantification, neglecting the exploration of topic continuity from the perspective of survival time. Conversation analysis research provides a theoretical basis for revealing sociological patterns of human verbal communication based on information exchange data analysis, but its application in informal information exchange analysis based on social media is still in its infancy. Therefore, this study takes text topics as the core, combines conversation analysis with topic analysis, and aims to reveal the characteristics and patterns of informal information exchange from micro and meso levels based on the analysis of topic and topic cluster running processes. Simultaneously, it discusses the connotation of topic continuity from both coherent continuation and intermittent continuation perspectives, and establishes continuity judgment criteria based on relative discussion intensity to quantitatively analyze topic evolution.

2.2 Conversation Analysis Research

Conversation Analysis Theory belongs to both linguistics and sociology. In linguistics, conversation analysis emphasizes the analysis of language forms and functions such as grammar, discourse, turn-taking, and topics [21]; in sociology, it aims to explain the social patterns and orders behind human verbal communication by discovering its patterns and modes [22]. Related research data mainly comes from conversation records in informal information exchange processes, which can be divided into offline conversation data and online conversation data.

Offline conversation data mostly uses audio and video recordings of people's conversations in natural or semi-experimental environments, converted into text for analysis. Based on this, conversation analysis research focuses on sequence structure analysis to reflect characteristics and patterns in verbal communication, examining conversational exchanges in specific contexts such as foreign language teaching [23], doctor-patient communication [24], cross-cultural work [25], as well as specific behaviors like negotiation requests [26] and storytelling [27].

Benefiting from internet technology development, the massive amount of objectively recorded communication data in online environments has promoted conversation analysis research based on online conversation data to keep pace with the times and grow increasingly. Scholars have explored information exchange characteristics of users in various social media such as academic virtual communities [28] and WeChat platforms [29] from multiple dimensions including content, relationships, and behavior. Additionally, conversation analysis meth-

ods in multimodal big data environments [30] and natural language processing methods based on conversation analysis [31], team decision support systems [32] have been improved and optimized.

3 Analytical Framework for Informal Information Exchange Topic Evolution Based on Conversation Analysis

This paper organizes topic running processes based on conversation analysis theory and defines various running states within them. It then combines the number of posts and participants to calculate the relative discussion intensity of topics, visualizing topic running processes and measuring continuity characteristics during topic evolution. Through topic running process analysis and continuity measurement, this study explores the bias characteristics of social media users in topic discussion content.

3.1 Topic Running Process Analysis

Within a time interval containing several time segments, the running process of a topic from initiation to termination may involve continuation, silence, regression, and other states. The discussion of these running states is limited to a certain scope, such as discussion time and participants. The specific descriptions of various states in the topic running process are as follows:

(1) **Topic Initiation**: Refers to a user posting new discussion topics and introducing related content, which may trigger participation from other users.

(2) **Topic Continuation**: Refers to after topic initiation, the user who introduced the topic or other users consecutively post to continuously express opinions related to the topic through in-depth exploration or extended expansion.

(3) **Topic Silence and Topic Regression**: These are complementary concepts, referring to the temporary cessation of topic-related posts in a certain time segment, but reappearing in one or several subsequent time segments within the overall selected time range.

(4) **Topic Termination**: Refers to the complete end of posts related to the topic, where all participants within the observation scope no longer post content related to the topic.

On social media platforms, different topics can be discussed simultaneously by the same user, and the same topic can be discussed simultaneously by different users. Posts about different topics on social media mostly present a parallel relationship, with each topic's running process being relatively independent. However, simultaneously, due to limited time and energy, users are selective in receiving, processing, and expressing information, leading to competitive relationships between different topics for more user attention and discussion in each

time segment. Therefore, exploring the running processes of multiple topics in a specific time interval can be achieved by calculating the relative discussion intensity of each topic, which simultaneously reveals the relationship of topics waxing and waning and analyzes changes in topic discussion focus and determines topic continuity characteristics, revealing users' bias in posting topics and its changes.

3.2 Topic Relative Discussion Intensity Calculation

When a topic occupies the largest proportion of user discussion posts in the current time segment, other topics may not be mentioned, or may be discussed limitedly by individual users through a small number of posts. If in the next time segment, another topic replaces the previous largest proportion topic and becomes the focus of user discussion in this time segment, the discussion status of this topic and other topics also has the aforementioned two possibilities. Therefore, calculating the relative discussion intensity of a topic in a single time segment mainly considers two indicators: the proportion of users discussing the topic and the proportion of posts. The specific calculation formula is as follows:

$$\text{Topic\_Strength}(TP_{tp}^{ti}) = \frac{\text{Count}(U_{tp}^{ti})}{\text{Total\_U}^{ti}} \cdot \alpha + \frac{\text{Count}(Post_{tp}^{ti})}{\text{Total\_Post}^{ti}} \cdot \beta \quad \text{(Formula 1)}$$

Where tp is the topic number, $tp = 1, 2, \ldots, n$; ti is the time segment number, $ti = 1, 2, \ldots, m$; $\text{Count}(U_{tp}^{ti})$ is the number of users participating in topic tp discussion in time segment ti, $\text{Count}(Post_{tp}^{ti})$ is the number of posts about topic tp discussion in time segment ti, Total_U^{ti} is the total number of users posting in time segment ti, Total_Post^{ti} is the total number of posts in time segment ti, and α and β represent the influence factors of the two indicators on topic relative discussion intensity. TP_{tp}^{ti} represents the relative discussion intensity of topic tp in time segment ti, with a value range of $[0, 1]$. If $TP_{tp}^{ti} = 1$, it means all posts by all users in time segment ti are related to topic tp; if $TP_{tp}^{ti} = 0$, it means no user posts content related to topic tp in time segment ti.

3.3 Topic Continuity and Its Determination

Topic continuity is manifested by the continuation of discussions related to the topic in UGC throughout the observed entire time interval, which can be defined from two perspectives: coherent continuation and intermittent continuation. First, in coherent continuation, topic continuity is manifested by topic-related posts spanning several time segments in the overall time interval, i.e., UGC involves the topic in multiple consecutive time segments. Second, in intermittent continuation, topic continuity is manifested by posting content entries related to the topic in several time segments of the UGC text stream, i.e., the number of time segments where topic-related posts exist exceeds a set threshold in

the overall time interval. Both perspectives measure user bias toward a certain topic from the time dimension, considering that users spontaneously filter topics they may contact and express opinions on due to limited time and energy. Through relative discussion intensity calculation in each time segment, longitudinal evolution analysis and horizontal comparative analysis of topics can be achieved.

This paper chooses to adopt the definition of coherent continuation from the perspective of survival time, combined with relative discussion intensity calculation to establish topic continuity determination criteria, i.e., the relative discussion intensity of the topic is greater than 0 in all time segments included in the overall time interval. Simultaneously, an exception is established: if a topic's relative discussion intensity is 0 occasionally, the topic should still be considered as having continuity characteristics. Relatively, topic non-continuity refers to topics being discussed briefly or frequently intermittently in the time interval, i.e., the topic's relative discussion intensity is greater than 0 only in some time segments, while being 0 in other time segments, and the situation where relative discussion intensity is 0 is non-occasional.

4 Informal Information Exchange Topic Cluster Identification Based on Semantic Association Filtering

A topic cluster composed of several semantically similar topics reflects social media users' expression of various yet implicitly related perspectives on a certain matter. Analyzing the composition structure, running process, and dominant topics of topic clusters can comprehensively reveal users' discussion style characteristics. Among them, the running state presented within the topic cluster can reveal the changing process of the topic cluster being stable and continuous, expanding and enriching, or converging and declining; the dominant topics determined by combining the continuity and relative discussion intensity comparison of various topics within the cluster can reflect users' discussion characteristics of being rich and diverse or focused and in-depth in posts. Therefore, to effectively extract topic clusters in informal information exchange, this paper first analyzes the applicability of commonly used topic similarity calculation methods, then designs association filtering conditions to determine candidate similar topic pairs, and finally discusses the complete topic cluster composition conditions.

4.1 Topic Similarity Calculation

To measure similarity between topics, it is necessary to first calculate the TF-IDF-based topic word set representing the core content of each topic, map it to the semantic space to obtain corresponding topic vectors for calculation. The main indicators include cosine similarity, KL divergence, symmetric KL divergence, JS divergence, etc.

Among them, the cosine similarity calculation method uses the cosine value of

the angle between topic vectors to measure their similarity, requiring both topics to be in the same semantic vector space. The cosine similarity between topic T_i and topic T_j, usually denoted as $\mathrm{Sim}(T_i, T_j)$, is calculated as follows:

$$\mathrm{Sim}(T_i, T_j) = \frac{T_i \cdot T_j}{|T_i| \times |T_j|} \quad \text{(Formula 2)}$$

The larger the cosine similarity $\mathrm{Sim}(T_i, T_j)$ value, the greater the similarity between topics. If two topics come from different vector spaces, KL divergence, symmetric KL divergence, JS divergence, etc., are usually chosen to measure similarity based on the distance of topic probability distributions, with the premise of having probability distributions of the same dimension. Assuming p is the probability distribution of topic T_i and q is the probability distribution of topic T_j, the probability distribution dimensions in p and q, i.e., the total vocabulary size, must both be n. If the vocabulary spaces of topics T_i and T_j are not the same, the source word lists of the topic probability distributions need to be merged and supplemented to ensure the same dimension number and vocabulary items in the topic probability distributions. The smaller the results obtained by these three calculation methods, the smaller the difference in probability distribution between topics, and the greater the similarity between the two topics.

4.2 Similar Topic Pair Selection

Based on similarity calculation results, determining similarity relationships between topics also requires setting corresponding filtering rules and thresholds to filter out topic pairs with low similarity and retain truly similar topic pairs to constitute candidate similar topic pairs.

The simplest and most direct method is to select several pairs of topics with relatively large similarity, calculate their similarity average value, and set it as the determination threshold for similar topic pairs. Obviously, this method is greatly influenced by subjectivity and has strong randomness. If the similarity between each topic and other topics is sorted in descending order, and assuming that for T_i, the topic with the maximum similarity is T_j, then it is considered that there is a topic association between T_i and T_j. However, this may lead to weak actual associations between topics because the association is purely based on maximum similarity determination and construction. Therefore, an improved association filtering method is needed, recording the semantic similarity calculation result between topic T_i and topic T_j as $S(T_i, T_j)$:

(1) Set a critical threshold ε; topic pairs with similarity less than this threshold have no similarity association;

(2) For topic T_i, if it has similarity associations greater than the critical threshold with both topic T_j and topic T_m, and if the similarity $S(T_i, T_m) < \theta \times S(T_i, T_j)$, where θ is the set association threshold, then topic T_i has

a stronger similarity association with topic T_j, and comparatively, the similarity association between topic T_i and topic T_m is too small to be considered.

After the above filtering processing, the set composed of several topic pairs with similarity associations can be regarded as a topic subset, i.e., $T_i, T_j \in TC$.

4.3 Topic Cluster Extraction Conditions

The similar topic pairs after association filtering form several topic subsets based on their similarity relationships. Topic subsets need further pruning through judgment conditions to ensure that topics within the cluster have tight similarity relationships before the final topic cluster extraction results can be determined. The following three judgment conditions can be considered to determine whether a topic subset constitutes a complete topic cluster:

(1) For topics constituting a topic cluster, there must be similarity associations between every pair;

(2) For topics constituting a topic cluster, each topic only needs to have similarity associations with at least one other topic within the cluster;

(3) The number of similarity relationships among topics constituting a topic cluster needs to be discussed based on the number of topics contained in the cluster. If a topic cluster contains more than three topics, each topic needs to have similarity associations with at least three other topics within the cluster; if a topic cluster contains three or fewer topics, each topic needs to have pairwise similarity associations with other topics in the cluster.

Among them, the first two limiting conditions may lead to topic clusters that are too small or too large due to being overly strict or lenient. Comparatively, the third judgment condition is more appropriate. Therefore, this paper selects the third judgment condition to prune the obtained topic subsets to determine topic clusters.

5 Empirical Analysis

5.1 Data Collection and Preprocessing

Posts by high-influence users on social media attract more attention from followers and ordinary users. Opinion groups composed of high-influence users trigger exponential growth in topic attention through responsive posting and repeated exposure, thereby influencing the development direction of online public opinion and even real-world events. Therefore, analyzing the topic and topic cluster evolution processes of UGC from opinion groups composed of high-influence users and measuring topic continuity can reveal the characteristics and changes of opinion groups in discussion content and style, providing references for effective

communication strategies in emergency response and network public opinion management.

This study first constructs a user influence evaluation index system by reviewing existing research on social media user influence measurement indicators [33-35], uses the analytic hierarchy process to determine index weights, and respectively identifies high-influence users participating in the discussion of the social focus event "Jiang Ge Case" on Sina Weibo and Zhihu platforms to form opinion group samples for the social media platforms. Posts from the two opinion groups over 37 months are collected as data sources for this empirical analysis. Subsequently, the post data is cleaned, obtaining 124,556 valid posts from the Sina Weibo opinion group and 2,833 valid posts from the Zhihu opinion group. After preprocessing such as word segmentation and stop word removal, the word vector representation function of Baidu AI is called and combined with TF-IDF calculation to screen feature words for each post to form text vectors. The optimal topic number K value is determined by calculating the sum of squared errors and silhouette coefficient using Python's Scikit-learn functions. K-means clustering yields 70 topics for the Sina Weibo opinion group and 35 topics for the Zhihu opinion group. Finally, the TF-IDF values of keywords in each topic cluster are calculated, and the top 10 words in descending order are used as topic words to describe topic content. Topic labels are determined by referencing Sina Weibo, Zhihu, public opinion websites such as Qingbo Index and Zhiwei Shijian, and news websites such as People's Daily Online and Sina News to facilitate subsequent analysis and expression.

5.2 Informal Information Exchange Topic Evolution Analysis

First, based on the temporal distribution characteristics of the experimental data from Zhihu and Sina Weibo platforms, appropriate time segment division units are selected respectively. Then, according to the topic relative discussion intensity calculation formula (Formula 1), the relative discussion intensity of each topic in each time segment is calculated. Finally, based on topic relative discussion intensity, the running status and continuity of user discussion topics on Sina Weibo and Zhihu platforms within the selected experimental time span are analyzed.

For online topic discussions, topics discussed by multiple users are usually more active and have wider dissemination than topics discussed by only a few users. Therefore, among the two influence factors α and β in the topic relative discussion intensity calculation formula (Formula 1), the participant user indicator obviously has greater influence on topic relative discussion intensity than the related post indicator. Meanwhile, referencing relevant settings in existing research [29], $\alpha = 0.6$ and $\beta = 0.4$ are set. Additionally, "occasional" topic relative discussion intensity of 0 is defined as occurring once, i.e., if a topic's relative discussion intensity is 0 in only one time segment, it is still considered to have continuity.

(1) Topic Evolution Analysis on Sina Weibo. Posts on Sina Weibo are relatively frequent, so this paper considers time segment division units of both day and week. If using day as the time segment unit, according to the criteria established in this paper, none of the Sina Weibo opinion group topics have continuity characteristics, and there are many missing time segments. Therefore, this paper selects week as the time segment unit for topic evolution analysis of the Sina Weibo opinion group. The overall time interval of 37 months is divided into 161 time segments, denoted as "TS+number" hereafter. After statistical substitution of indicator parameters into the relative discussion intensity calculation formula (Formula 1), the relative discussion intensity of each topic in each time segment on Sina Weibo is obtained. Taking Topic2, Topic36, and Topic47 as examples, the topic running process is plotted as shown in Figure 1 [Figure 1: see original paper].

It can be seen from Figure 1 that Topic36, mainly focused on opinion commentary, has relatively high relative discussion intensity throughout the entire time interval, showing obvious continuity characteristics, reflecting the Sina Weibo opinion group's persistent and strong desire to express personal viewpoints. The relative discussion intensity of Topic2 and Topic47 is generally low, with topic silence appearing in several time segments where the relative discussion intensity is non-occasionally 0, showing no continuity characteristics. Among them, Topic2 revolves around holiday red envelopes, with prominent periodic topic regression; Topic47 is mainly related to Fang Zhouzi and Peng Jian's fraud security fund case. As a public opinion event dominated and promoted by high-influence users to attract other netizens' attention, its relative discussion intensity distribution shows a relatively complete life cycle process. That is, posts related to Topic47 begin with exposing relevant information (TS4), followed by continuous disclosure of relevant information and extended discussions as the event progresses; the topic discussion reaches its climax during TS30-TS39 and TS49-TS65, with relatively large relative discussion intensity and relatively continuous time segments; afterward, posts on this topic show frequent silence and regression states, indicating that related discussions have entered a decline period; finally, after TS132 and beyond, it enters a silence state until related discussions completely end and the topic terminates.

(2) Topic Evolution Analysis on Zhihu Platform. Based on the characteristics of post time intervals on the Zhihu platform, this paper considers time segment division units of both week and month. If using week as the time segment unit, due to the relatively sparse temporal distribution of Zhihu user posts, none of the Zhihu opinion group topics have continuity characteristics. Therefore, division by month is adopted, totaling 37 time segments, denoted as "TS+number". Relevant parameter values are statistically substituted into Formula 1 to calculate the relative discussion intensity of Zhihu opinion group topics in each time segment. Taking Topic1, Topic7, Topic16, and Topic17 as examples, the topic running process is plotted as shown in Figure 2 [Figure 2: see original paper].

It can be seen from Figure 2 that Topic7 has relative discussion intensity greater than 0 in all time segments after initiation, showing obvious continuity characteristics, and is more heavily weighted in opinion group posts compared to the other three topics, reflecting the characteristic of Zhihu opinion groups conducting professional interpretations from a legal perspective and continuing this preference into discussions of social focus events. Since its initiation, Topic16, except for the occasional 0 relative discussion intensity in TS25, has been discussed by the opinion group in all other time segments, also showing continuity characteristics, indicating that encouraged by Zhihu platform's encouragement for detailed content analysis, the opinion group combines sudden social focus events with personal long-term interests (such as literary works) to express their own viewpoints. Additionally, the opinion group began posting to participate in discussions related to social livelihood topics (Topic17) from TS18. Although there was brief silence in the initial stage (TS19 and TS20), the relative discussion intensity in subsequent time segments is relatively stable, showing that this opinion group has gradually developed a bias toward participating in social event discussions. Furthermore, the silence state of Topic1 in multiple time segments and the frequent alternation between silence and regression after TS18 both show that movie entertainment and other unrelated topics are not the collective preferences shared by opinion groups formed based on social focus event discussions. Such topics only represent the interests of individual members and usually do not involve metaphorical expressions associated with event information.

According to the criteria established in this paper, a total of 6 topics in Sina Weibo opinion group posts have continuity characteristics, with corresponding topic content reflecting the platform's opinion groups formed through social focus event discussions having a bias toward social event, social celebrity dynamics, and visit-related social topics, as well as participation preferences for authoritative government information release, viewpoints, and emotional exchange topics. On Zhihu, a total of 3 topics have continuity characteristics, revealing that the platform's opinion groups participate in social focus event discussions based on daily attention to overseas study and other event-related content topics, and interpret them by combining professional expertise such as law and interests such as literature when expressing opinions.

5.3 Informal Information Exchange Topic Cluster Evolution Analysis

In topic similarity calculation, since topic vectors in this study come from the same semantic space, cosine similarity is used to measure topic similarity. In association filtering, the critical threshold ε is set to 0.2-0.4 referencing existing research; the association threshold θ is typically set between 0.5-0.7 in related research [36]. This experiment calculates the number of candidate similar topic pairs extracted under three value schemes of $\theta = 0.7$, $\theta = 0.65$, and $\theta = 0.6$ in both Sina Weibo and Zhihu. Based on the number and scale of extractable candidate topic clusters, $\theta = 0.65$ is determined for Sina Weibo and $\theta = 0.7$

for Zhihu, obtaining 48 similar topic pairs in Sina Weibo opinion group posts and 41 similar topic pairs in Zhihu. The topic similarity relationship network is plotted using the visualization tool Gephi, as shown in Figure 3 [Figure 3: see original paper].

It can be seen from Figure 3 that in the Sina Weibo opinion group topic similarity relationship, there is an obvious similarity relationship between Topic3 and Topic42, and both have strong similarity with Topic66, while the similarity relationships among other topics are weak and complexly intertwined. In the Zhihu opinion group topic similarity relationship, there is obvious semantic similarity between Topic2 and Topic22, and between Topic3 and Topic26, Topic10 and Topic25, Topic11 and Topic25, each naturally forming several relatively obvious topic subsets.

(1) Topic Cluster Evolution Analysis on Sina Weibo. According to the topic cluster extraction conditions discussed in Section 4.3, this paper extracts a total of 14 topic clusters from the 70 posting topics of the Sina Weibo opinion group, among which only 1 topic cluster is composed of more than 3 similar topics, 5 topic clusters are respectively composed of 3 similar topics, and 8 topic clusters are respectively composed of 2 similar topics. The different topics constituting topic clusters reflect the opinion group's analysis, interpretation, and opinion expression on the same issue from different perspectives, revealing the implicit associations established between different topics in the subjective cognition contained in UGC, which helps to understand the entry points of opinion group posting discussions more richly and deeply. For example, Topic Cluster 1 contains 9 topics covering social, life, government affairs, current affairs, emergencies, and other topics, showing that the Sina Weibo opinion group pays attention to multiple aspects closely related to people's daily life and interests; the 3 topics constituting Topic Cluster 2 mainly involve government affairs and current affairs content, reflecting that the Sina Weibo opinion group extends discussions on domestic urban construction and management, international exchanges and cooperation, etc., from their attention to government work dynamics. Taking Topic Cluster 2 as an example, the changes in relative discussion intensity of topics in each time segment are plotted as shown in Figure 4 [Figure 4: see original paper] to analyze the relationships among topics within the topic cluster running process.

Due to the large number of time segments in Sina Weibo and the frequent changes in discussion intensity of topics within this topic cluster, the scatter lines in Figure 4 are complexly intertwined, so various running states are not labeled in the figure. It can be seen that Topic15 and Topic31 show continuation states in most time segments, and based on changes in relative discussion intensity values, the two topics alternately occupy the discussion focus of the opinion group within this topic cluster; Topic23 gradually declines after a period of frequent posting (TS1 to TS32), with topic running frequently switching between silence and regression states. After analyzing the running processes and dominant topics of all Sina Weibo topic clusters, it can be found that the

vast majority of topic clusters have running states similar to Topic Cluster 2 shown in Figure 4, where multiple active topics alternately dominate the discussion within the topic cluster, revealing the rich and diverse characteristics of Sina Weibo topic cluster evolution. Additionally, a few topic clusters containing only 2 topics show running states where one topic occupies the vast majority of discussions, i.e., Topic Clusters 11, 12, and 14, whose running processes are similar to those shown in Figure 1 in Section 5.2.

(2) Topic Cluster Evolution Analysis on Zhihu Platform. According to the same topic cluster extraction conditions, a total of 11 topic clusters are extracted from the 35 posting topics of the Zhihu opinion group, among which 4 topic clusters are respectively composed of more than 3 similar topics, and the remaining 7 topic clusters are respectively composed of 2 similar topics. The different topics constituting topic clusters also reflect the Zhihu opinion group's interpretation and expression characteristics of analyzing the same issue from different perspectives. However, compared with the relatively rich content coverage of Sina Weibo opinion group posting topic clusters showing horizontal association characteristics, the topic clusters extracted from Zhihu opinion group posts are relatively focused in content coverage, showing vertical depth characteristics. Taking topic clusters containing more than 3 topics as examples, Topic Cluster 1 mainly covers film and entertainment content, Topic Cluster 2 focuses on legal professional content, Topic Cluster 3 focuses on competitive sports content, and Topic Cluster 4 emphasizes social life content. Some of these contents are related to the analytical perspectives chosen by opinion groups when participating in social focus event discussions, while other contents reveal that opinion groups have relatively broad and sustained interests. Taking Topic Cluster 1 as an example, the changes in relative discussion intensity of each topic are plotted as shown in Figure 5 [Figure 5: see original paper].

It can be seen from Figure 5 that in the early stage (TS1 to TS17), the discussion mainly focused on Topic16, occasionally touching on other topics, with relatively single overall content; in the later stage (after TS17), posts on different topics increased, making the discussion content in this topic cluster richer. This characteristic of topic cluster content gradually expanding and enriching over time is relatively common in Zhihu opinion group posting data. In the specific running process, Topic16 not only has continuity characteristics but also occupies the hot discussion position in the described topic cluster in most time segments. In the running processes of other topics, there exist either long or short silence periods or frequent switching between silence and regression. Although Topic24 and Topic29 show fluctuations around a relatively high level of relative discussion intensity after TS17, and occupy the hot position in Topic Cluster 1 in individual time segments, Topic16 remains the core topic in Topic Cluster 1 in most cases.

6 Conclusion and Outlook

This paper introduces conversation analysis theory to reveal the evolution characteristics and patterns of informal information exchange at micro and meso levels by analyzing the running processes of topics and topic clusters, and proposes exploration of topic continuity as a measurement standard for evolution analysis. In the empirical analysis, taking opinion groups composed of high-influence users in social focus events as examples, the running processes of topics and topic clusters on Sina Weibo and Zhihu are analyzed to reveal the bias characteristics and changing trends of opinion groups in topic discussion content and style in informal information exchange, aiming to provide references for developing effective communication strategies in network public opinion management.

The analysis finds that topic continuity reflects the obvious bias of opinion groups on the topic content and indicates the main entry points of opinion groups' viewpoints in social focus event discussions. Meanwhile, the obvious differences between Sina Weibo and Zhihu opinion groups in topics with continuity characteristics reveal the differences in UGC content between the two social media platforms regarding event-related and daily states of high-influence users in social focus event discussions, indicating their different roles in informal information exchange in the online environment. The relationship network formed by candidate similar topic pairs shows the characteristics of complex content interweaving and blurred boundaries in Sina Weibo, and clear content similarity differences and clear boundaries in Zhihu, which originate from multiple factors such as UGC posting characteristics and topic identification methods on the two platforms. Under these circumstances, the method of using similarity relationship quantity determination in topic cluster extraction conditions helps accurate determination of complete topic clusters. Meanwhile, topic cluster running process analysis demonstrates that Sina Weibo opinion groups diverge and explore different topics within a certain range, while Zhihu opinion groups always focus on core topics.

This study still has certain limitations and shortcomings, which can be improved in future research from the following two aspects: First, explore topic continuity from the perspective of intermittent continuation and compare it with the coherent continuation in this study to further enrich research on topic evolution measurement standards; second, compare and analyze the similarities and differences between event-related and conventional state UGC content from the perspective of topic and topic cluster evolution running, combined with evaluation objects and emotional tendencies, to enrich the analysis levels of cognitive consistency and dissonance research in emergency stimulus contexts. Additionally, expansion can be made from the perspective of mining the correlation of users' opinion expression changes in cross-platform UGC to deepen understanding of users' opinion expression in informal information exchange in the online environment.

References

[1] Wang Zhenhuang, Chen Siming, Yuan Xiaoru. Visual analysis research oriented to Weibo topics [J]. Journal of Software, 2018, 29(4): 1115-1130.

[2] Bex F J, Lundgren L, Cripps K J. Scientific Twitter: the flow of paleontological communication across a topic network [J/OL]. PLoS One, 2019, 14(7). [2021-05-06]. http://doi.org/10.1371/journal.pone.0219688.

[3] Bokaei M H, Sameti H, Liu Y. Unsupervised approach to extract summary keywords in meeting domain [C]// Dugelay J L, Stock D. Proceedings of the 23rd European signal processing conference. Piscataway: IEEE, 2015: 1406-1410.

[4] Chen Y H, Lu J L, Meng F T. Finding keywords in blogs: efficient keyword extraction in blog mining via user behaviors [J]. Expert systems with applications, 2014, 41(2): 663-670.

[5] Gu Ying, Li He, Li Yeye, et al. Research on enterprise competitive intelligence demand mining based on online reviews [J]. Modern Information, 2021, 41(1): 24-31.

[6] An Lu, Li Qian. Detection of secondary and derivative events in emergencies based on hotspot topic identification [J]. Information and Documentation Services, 2020, 41(6): 26-35.

[7] Zhang Y H, Mao W J, Zeng D, et al. Topic evolution modeling in social media short texts based on recurrent semantic dependent CRP [C]// Benjamin V, Li W F. Proceedings of 2017 IEEE international conference on intelligence and security informatics. Piscataway: IEEE. 2017: 119-124.

[8] Liao Haihan, Wang Yuefen, Guan Peng. Topic mining and viewpoint identification of different communicators in Weibo public opinion dissemination cycle [J]. Library and Information Service, 2018, 62(19): 77-86.

[9] Liang Xiaohe, Tian Ruya, Wu Lei, et al. Review of Weibo topic discovery research methods [J]. Library and Information Service, 2017, 61(14): 141-148.

[10] Liang Xiaohe, Tian Ruya, Wu Lei, et al. Weibo similarity based on hypernetwork and its application in Weibo public opinion topic discovery [J]. Library and Information Service, 2020, 64(11): 77-86.

[11] Sasaki K, Yishikawa T, Furuhata T. Online topic model for Twitter considering dynamics of user interests and topic trends [C]// Martony M. Proceedings of 2014 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2014: 1977-1985.

[12] Liu Y P, Peng H, Li J X, et al. Event detection and evolution in multilingual social streams [J/OL]. Frontiers of computer science, 2018, 12(5). [2021-05-06]. http://doi.org/10.1007/s11704-019-8201-6.

[13] Dehghani N, Asadpour M. SGSG: semantic graph-based storyline generation in Twitter [J]. Journal of information science, 2019, 45(3): 304-321.

[14] Goyal P, Kaushik P, Gupta P, et al. Multilevel event detection, storyline generation, and summarization for Tweet streams [J]. IEEE transactions on computational social systems, 2020, 7(1): 8-23.

[15] Huang J J, Peng M, Wang H, et al. A probabilistic method for emerging topic tracking in microblog stream [J]. World Wide Web, 2017, 20(2): 325-350.

[16] Cai H Y, Huang Z, Srivastava D, et al. Indexing evolving events from Tweet streams [J]. IEEE transactions on knowledge and data engineering, 2015, 27(11): 3001-3015.

[17] Abulaish M, Fazil M. Modeling topic evolution in Twitter: an embedding-based approach [J/OL]. IEEE access, 2018, 6. [2021-05-06]. http://doi.org/10.1109/ACCESS.2018.2878494.

[18] Pruss D, Fujinuma Y, Daughton A R, et al. Zika discourse in the Americas: a multilingual topic analysis of Twitter [J/OL]. PloS one, 2019, 14(5). [2021-05-06]. http://doi.org/10.1371/journal.pone.0216922.

[19] Wang Zhenhuang, Chen Siming, Yuan Xiaoru. Visual analysis research oriented to Weibo topics [J]. Journal of Software, 2018, 29(4): 1115-1130.

[20] Liu Yashu, Zhang Haitao, Xu Hailing, et al. Research on network public opinion emergency evolution topic graph based on multi-dimensional feature fusion [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(8): 798-808.

[21] Sacks H, Schegloff E A, Jefferson G. Simplest systematics for the organization of turn-taking for conversation [J]. Language, 1974, 50(4): 696-735.

[22] Wu Yaxin, Yu Guodong. Rectifying the name of conversation analysis [J]. Journal of Shanxi University (Philosophy and Social Science Edition), 2017, 40(1): 85-90.

[23] Zhao Yan, Zhang Qiwei, Xu Rui, et al. Translanguaging and identity construction as learning means for bilinguals [J]. Modern Foreign Languages, 2021(2): 258-270.

[24] Stommel W, Van Goor H, Stommel M. Other-attentiveness in video consultation openings: a conversation analysis of video-mediated versus face-to-face consultations [J]. Journal of computer-mediated communication, 2019, 24(6): 275-292.

[25] Avison D, Banks P. Cross-cultural (mis)communication in IS offshoring: understanding through conversation analysis [J]. Journal of information technology, 2008, 23(4): 249-268.

[26] Wu Yaxin, Liu Shu. A study on the sequential organization of the subtlety of request behavior [J]. Modern Foreign Languages, 2020, 43(1): 32-43.

[27] Peng Xin, Zhang Wei. Conversation analysis of storytelling in daily conversation [J]. Journal of Shanxi University (Philosophy and Social Science Edition), 2019, 42(4): 137-144.

[28] Lu Heng, Zhang Xiangxian, Zhang Liman, et al. Research on user interaction behavior characteristics in virtual academic communities from the perspective of conversation analysis [J]. Library and Information Service, 2020, 64(13): 80-90.

[29] Bazhichao, Li Gang, Mao Jin, et al. Analysis of network structure, behavior, and evolution of internal information exchange in WeChat groups: from the perspective of conversation analysis [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(10): 1009-1021.

[30] Gu Y, Li X Y, Huang K X, et al. Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder [C]// ACM. Proceedings of the 26th ACM multimedia conference. New York: ACM, 2018: 537-545.

[31] Housley W, Albert S, Stokoe E. Natural action processing: conversation analysis and big interactional data [C]// ACM. Proceedings of the halfway to the future symposium. New York: ACM, 2019: 1-4.

[32] Kono S, Aihara K. Prototype of decision support based on estimating group status using conversation analysis [C]// Yamamoto S. Proceedings of the 18th international conference on human-computer interaction. Berlin: Springer, 2016: 40-49.

[33] Zhang Xing, Wei Shufen, Wang Li, et al. Empirical research on influencing factors of Weibo opinion leaders in crisis events [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 66-75.

[34] Cui L, Pi D C. Identification of micro-blog opinion leaders based on user features and outbreak nodes [J]. International journal of emerging technologies in learning, 2017, 12(1): 141-154.

[35] An Lu, Hu Junyang, Li Gang. Research on profiling high-influence users on social media in emergency contexts [J]. Information and Documentation Services, 2020, 41(6): 6-16.

[36] Guo Xiaoli, Zhou Zilan, Liu Yaowei, et al. News topic evolution analysis based on DTS-ILDA model and association filtering [J]. Journal of Applied Sciences, 2017, 35(5): 634-646.

Author Contributions: Wang Xiao: Proposed research ideas and framework, collected and analyzed data, wrote the paper. Ma Chao: Processed and analyzed data. Zhai Shanshan: Provided paper revision suggestions and participated in paper revision.

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