Proceedings of 3rd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/42/20242501

Research on the Homogenization of Network Information Based on Chinese Social Media Weibo and Zhihu

Minxing Gu[1,a,*]

[1]*School of Computer Science, Wuhan University, Wuhan City, Hubei Province, 430072, China*
a. 2020302111157@whu.edu.cn
**corresponding author*

Abstract: This study investigates the phenomenon of information homogenization on Chinese social media platforms, specifically focusing on Weibo and Zhihu. As two of China's largest and most influential platforms, they provide contrasting environments—Weibo for mass, entertainment-driven content, and Zhihu for in-depth knowledge discussing and sharing. By analyzing data from both platforms between July 2022 and April 2024 using natural language processing and Latent Dirichlet Allocation (LDA) models, this research reveals that information homogenization has increased on Zhihu, driven by concentrated discussions around a few trending topics, and the corresponding push strategies of online platforms. In contrast, Weibo has maintained a relatively stable, high degree of homogenization throughout the period. These trends highlight the growing challenge of content diversity on Chinese social networks, posing implications for both users and platforms. The findings underscore the need for strategies that balance user engagement with the promotion of diverse content, contributing to a healthier online ecosystem and better information consumption.

Keywords: Homogenization, Social Media, Network Information.

1. Introduction

Launched on January 26, 2011, Zhihu is a mainland Chinese question-and-answer platform. The name "Zhihu" means "Do you know?" in classical Chinese, and its layout is similar to that of the American website Quora. The website claimed to have over 100 million registered users as of September 20, 2017. Among these users, 26 million were active every day, spending an average of one hour on the site, and producing 18 billion page views each month. By the end of 2018, registered users exceeded 220 million, with over 29 million questions and 125 million answers. On March 26, 2021, Zhihu became public on the New York Stock Exchange. On April 11, 2022, it started trading on the Hong Kong Stock Exchange.

Like Quora, Zhihu lets people post questions, answer them, and upvote the answers they think are useful. In addition, it offers monthly and daily bulletins along with expert panels. At first, registration in Zhihu required a lengthy application procedure. It was largely by invitation or required recommendations from several hundred current users. Zhihu set out to establish a professional Q&A platform that would eventually open up to a larger population. In March 2013, open registration was introduced, enabling users to post questions and get responses from other users. Users are able to follow updates from other users and keep tabs on the progress of their own queries, as well as conversations pertaining to their responses and remarks.

DOI: 10.54254/2753-7064/42/20242501

Zhihu's professional focus and steady growth from an exclusive, invitation-only platform to one of China's largest social media sites highlights its role as a hub for knowledge sharing and expert interaction, while, in contrast, Weibo's rapid rise to prominence underscores its emphasis on fostering broad user relationships and real-time content dissemination across diverse demographics.

Weibo, formerly called Sina Weibo, is a well-known microblogging site in China that was introduced on August 14, 2009, by Sina Corporation. As of Q1 2022, it has over 582 million monthly active users, of which 313 million are actively using the site. Notably, 50.10% of users are male and 49.90% are female. 85% of users access Weibo via a mobile device, and 70% of users are college aged. With over 100 million messages written every day, the network has had tremendous financial success as seen by growing stock prices, robust advertising income, and great quarterly profitability.

Weibo places a strong emphasis on fostering user relationships to promote information exchange and dissemination. While others can interact through comments and multimedia chat, users can post photographs and videos for instant public sharing. Initially, the platform drew in a lot of celebrities, in order to improve communication. Since then, it has grown to include a large quantity of non-governmental organizations, businesses, governmental departments, and media personalities. Today, Weibo is still the most popular platform in China's social media market, despite fierce competition.

The development and application of large models have permeated various fields, from natural language processing and image recognition to complex data analysis. The construction and optimization of large models rely on a substantial amount of high-quality corpus information, which forms the foundation for training these models. The quality of this corpus directly affects the accuracy, generalization ability, and intelligence of the models. However, on the internet, a noticeable phenomenon has emerged: the increasing information homogenization, which is especially evident on the Chinese internet. Similar topics and content are repeatedly copied and disseminated across major Chinese internet platforms, lacking innovation and diversity. This phenomenon not only affects the efficiency and quality of information acquisition for users but also poses significant challenges to the training of large models.

Weibo and Zhihu, each with its own platform characteristics, display distinct differences in content production and user interaction. Weibo's social nature, fragmented information, and mass appeal make it a platform for rapid information dissemination and wide coverage. In contrast, Zhihu's atmosphere of rational and serious discussion, along with its users' professional backgrounds, provides an environment for in-depth discussions and knowledge sharing. Therefore, this study focuses on Weibo and Zhihu as representative platforms within the Chinese internet ecosystem, making the results more representative.

2. Literature Review

Due to the importance of online information sufficiency and diversity, the issue of information homogenization on the internet has attracted the attention of many scholars for research.

In social networks, users naturally gravitate towards consuming concentrated and homogeneous content. Kossinets observed that individuals with shared interests tend to grow more alike over time, indicating a strong homogeneity within social networks [1]. Mcewan introduced the concept of "self-involvement," suggesting that users seek groups that align with their viewpoints in specific discussions, which reflects their level of personal engagement with the issue. The longer users participate in these groups, the stronger their self-involvement becomes [2]. Additionally, Mcewan found that users exhibit a converging behavior once they recognize the unique characteristics of a particular social platform. Similarly, Mikal noted that explicit communication norms present in online communities play a role in shaping user behavior [3]. As users observe these public standards, they tend to adjust their own content to align with the platform's expectations, further contributing to the convergence of behavior and content within social networks.

DOI: 10.54254/2753-7064/42/20242501

On the other hand, algorithmic recommendations play a significant role in guiding users toward concentrated content, further contributing to the homogenization of online information. Hosanagar argued that while algorithms do strengthen the entertainment needs of niche groups, leading to the formation of isolated communities, overall content consumption trends show increasing similarity [4]. Over time, users' content preferences gradually converge towards a unified whole. Airold studied the network of associations between recommended music videos on user-generated content (UGC) platforms [5]. The research revealed that platforms often recommend similar videos based on the most frequently shared viewing habits among users, thereby creating interconnected clusters of related videos. This phenomenon reinforces content homogeneity. Scheufele pointed out that in modern society, information is primarily disseminated through reader comments, Facebook "likes," and article "shares"[6]. This method of communication fosters the spread of content that aligns with group preferences, pushing social networks towards a convergence of content that reflects collective biases and tastes.

The combined influence of users' natural inclinations and algorithmic recommendations has led to the emergence of "information cocoons" in social networks. Cheng Shi'an and colleagues proposed that the information cocoon is the most basic organizational unit in modern society for information aggregation and consensus-building [7]. It brings together individuals with similar interests and views into a relatively closed environment for information dissemination, which then affects their cognition and behavioral patterns. This closed loop of information reinforces users' existing beliefs and preferences, limiting exposure to diverse perspectives. Yao Wenkang pointed out that information cocoons emerge because audiences lack access to heterogeneous viewpoints, which leaves them confined to a uniform information environment [8]. Xu Xiang and his team examined Sina Weibo as a case study to assess the deepening effects of information cocoons [9]. Their findings showed that as a user's "cocoon effect" deepens, the information they consume becomes increasingly similar, reflecting convergence rather than divergence. Users with a higher degree of cocoon immersion tend to engage with content that mirrors the information of "top cocoons" and "neighboring cocoons," leading to further homogeneity. Jiang Zhongbo argued that the essence of the filter bubble phenomenon is the uniformity of information that users encounter, which narrows their cognitive scope and contributes to viewpoint polarization [10]. In this process, users are often viewed as passive participants, subjected to the limitations imposed by their information environment. This results in an echo chamber where new or differing perspectives struggle to penetrate.

The aforementioned studies and hypotheses provide preliminary background knowledge. Nevertheless, research gaps persist. Despite the abundance of material on the information homogenization on English social networks, few are about Chinese social networks, which has more than one billion active users. Moreover, most studies on the homogenization of online information are qualitative, few research are quantitative. The purposes are that after studying several mainstream Chinese social networks, researchers and marketers can gain insights into the trends and characteristics of information homogenization on the Chinese internet, as a reference for studies and decision-making. In order to meet the desired aims of this study and let the relevance of this study may be put into effect, this research will study two mainstream social networks through natural language processing tools.

3. Methodology

The methodology involves scraping the trending lists and hot searches from Zhihu and Weibo from 2022 to 2024, followed by data cleaning of the collected corpus. After data cleaning, word segmentation is performed on the corpus, and the segmented text is vectorized to serve as input for the LDA model. The resulting text vectors are input into the LDA topic model, and the results are visualized using Matplotlib to plot a three-dimensional scatter plot. A k-nearest neighbors' graph is

DOI: 10.54254/2753-7064/42/20242501

then generated based on this plot. After that, the coefficient of global clustering will be calculated. Finally, statistical analysis is conducted on the monthly changes in the global clustering coefficient to draw conclusions.

Latent Dirichlet Allocation (LDA) is an algorithm that uses topic modeling to find hidden topics in a massive collection of documents. The assumption is that each document is created from a combination of different topics, and each topic is represented by a probability distribution across a set of words. By iterating optimization, LDA determines the topic distribution for each document and the word distribution for each topic by looking at the words present in the documents.

Compared to other dimensionality reduction algorithms, such as Principal Component Analysis (PCA), LDA has the advantage of focusing on the distribution of topics within documents. Each topic is composed of a group of semantically related words. This means that even if the specific vocabulary used in different documents is not identical, LDA can still identify the semantic similarity between these documents if the words belong to similar topics. In contrast, PCA tends to find the principal directions of the data (principal components) that capture the directions of maximum variance, which may not effectively distinguish different semantic topics.

The global clustering coefficient is commonly used in network science to measure the connectivity between nodes and their neighbors. It is primarily used to analyze clusters or community structures within a network. When applying the global clustering coefficient to the results of LDA (Latent Dirichlet Allocation), a higher global clustering coefficient suggests lower semantic diversity. This is because a high clustering coefficient may indicate that most documents are concentrated around a few topics, resulting in significant repetition of semantic information in the corpus. The larger the global clustering coefficient, the more concentrated the topic distribution among documents, indicating a higher degree of homogeneity in online information.

4. Results

First, we collected data from Zhihu for the period between July 2022 and April 2024. After data collection and cleaning, approximately 3,200 posts from Zhihu's trending list were gathered each month, amounting to about 100,000 characters per month. For each month's data, we processed it using the methodology described above, which yielded the global clustering coefficient for that month. We conducted sampling, statistical analysis, and calculations every three months for the data between July 2022 and April 2024, and the results are shown in table1.

Table 1: Global Clustering Coefficient of Zhihu between 2022.7-2024.4

Time	Global Clustering Coefficient
2022.7	0.6558560597703876
2022.10	0.7120610892452733
2023.1	0.7198511421546222
2023.4	0.7566940359652023
2023.7	0.7682426858238937
2023.10	0.7682426858238937
2024.1	0.772204664818436
2024.4	0.7609416497415571

The data indicates that during the statistical period, the level of information homogenization on the Zhihu platform showed a gradually increasing trend.

Similarly, we collected data from Weibo for the same period (July 2022 to April 2024). After data collection and cleaning, about 6,300 posts from Weibo's trending list were gathered each month,

DOI: 10.54254/2753-7064/42/20242501

totaling approximately 70,000 characters per month. The data for each month was processed using the same methodology, resulting in the global clustering coefficient for that month. Sampling, statistical analysis, and calculations were also conducted every three months for Weibo data during the same period, and the results are shown in table2.

Table 2: Global Clustering Coefficient of Weibo between 2022.7-2024.4

Time	Global Clustering Coefficient
2022.7	0.768745271657893
2022.10	0.8069002091938962
2023.1	0.7995007820785682
2023.4	0.7761777657745653
2023.7	0.7837032731384084
2023.10	0.8020899009508353
2024.1	0.7790842771030843
2024.4	0.7680643882705557

The data indicates that the level of information homogenization on the Weibo platform remained relatively stable over the same period.

A linear regression analysis was performed on both data sets and the results were compared in fugure1.

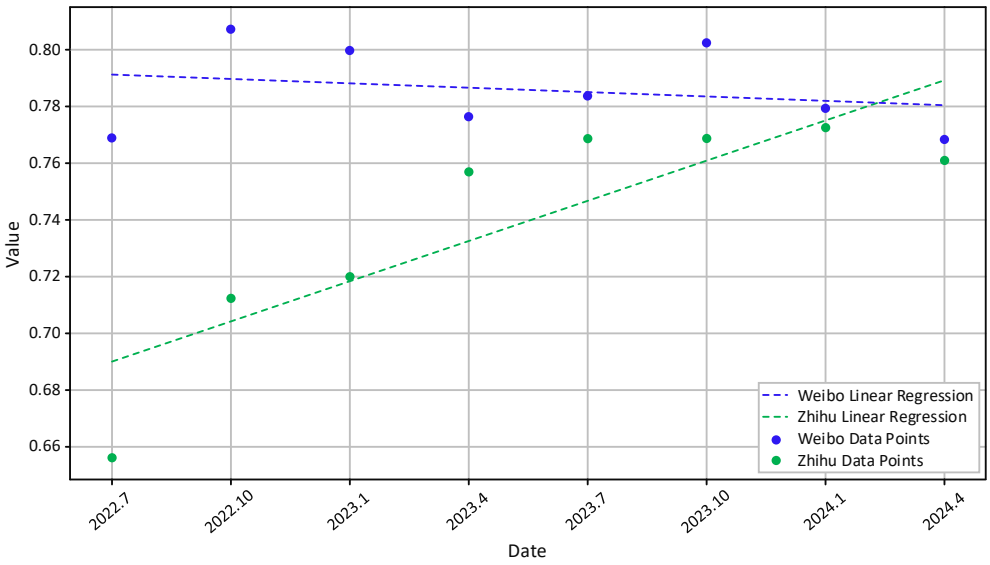

Figure 1: Linear Regression of Weibo and Zhihu Clustering

The data shows that from 2022 to 2024, Zhihu's clustering coefficient exhibited a rapid upward trend (Figure 1). In comparison, Weibo's clustering coefficient remained consistently high throughout the period. This suggests that information homogenization on Weibo might have been a long-standing issue, while on Zhihu, it appears to be a more recent development in the past few years.

5. Discussion

First, between July 2022 and April 2024, there has been a noticeable increase in the level of information homogenization on the Zhihu platform. As a Chinese internet platform primarily focused on long-form content and ideas, Zhihu has objectively become a significant source of thought

DOI: 10.54254/2753-7064/42/20242501

leadership in the Chinese internet space. On the other hand, it is also more susceptible to the influence of the broader Chinese online discourse environment. Originally, as a knowledge-sharing platform, Zhihu's users were more focused on professional insights, in-depth discussions, and perspectives from various fields, leading to a relatively low degree of content homogenization. However, starting in 2022, the emergence of several issues that polarized public opinion on the Chinese internet, such as the Russia-Ukraine war and COVID-19 prevention policies, led to highly intense discussions on various topics among Zhihu users. A large portion of these discussions became concentrated on a few key issues. On the other hand, by promoting these topics, the Zhihu platform was able to attract more traffic and increase user engagement. For these reasons, Zhihu's trending searches became focused on a few specific subjects, contributing to the increasing homogenization of information on the platform as reflected in its trending topics.

Second, this trend on Zhihu also mirrors the broader tightening of information control by authorities. As an important hub for idea generation and dissemination in the Chinese internet space, the platform is more significantly affected by this regulation. Starting from the second half of 2022, due to the impact of COVID-19 control policies, long-standing economic issues began to surface, leading to a period of economic downturn in China that eventually resulted in deflation. Economic hardships intensified social tensions, driving the need for greater control over public discourse. This heightened need for regulation was implemented across Chinese internet platforms. Specifically, for propaganda purposes, certain topics appeared frequently in trending sections, dominating the lists for days at a time, while other topics that did not align with the desired messaging were quickly removed. This selective amplification and suppression likely further contributed to the increasing homogenization of content on Zhihu.

In contrast to Zhihu, the degree of information homogenization on the Weibo platform remained relatively stable from July 2022 to April 2024, maintaining an already high level. The difference between these two platforms can likely be attributed to Weibo's nature as a more mass-oriented and entertainment-focused platform. On Weibo, a large proportion of trending topics are dominated by entertainment-related news. Additionally, Weibo's massive user base, which is much broader than Zhihu's, has led to a tendency to cater to popular interests and entertainment trends. As a result, user discussions tend to concentrate on mainstream topics and entertainment events, contributing to a higher degree of content similarity. Furthermore, Weibo's content dissemination mechanisms amplify this effect. The platform’s trending topics and recommendation systems often prioritize content that already has high engagement, quickly amplifying popular information while limiting the exposure of more diverse viewpoints. Due to the reasons mentioned above, the high degree of information homogenization on the Weibo platform is not limited to the period from July 2022 to April 2024; it is, in fact, a long-term phenomenon.

The homogenization of content on Weibo has been a long-standing phenomenon, whereas the homogenization of content on Zhihu has only emerged in recent years. This reflects a broader trend of increasing information homogenization within the Chinese internet, contributing to the gradual deterioration of the online ecosystem. From a business perspective, for Chinese internet platforms like Zhihu, which rely heavily on content, the time cost of obtaining useful information for learning purposes is steadily rising. Additionally, the difficulty of collecting valuable data from these platforms for research, large language model training, or business decision-making is also increasing. This presents greater challenges for companies and analysis teams that rely on Chinese internet data.

On the other hand, platforms like Weibo, which are mass social media networks, have long been in a state of high information homogenization, flooded with various forms of "junk information." From a business standpoint, however, this situation benefits Weibo by maintaining large traffic volumes and a vast user base, enhancing its value for advertising and commercial promotion. From an ethical standpoint, these platforms should strive to balance commercial interests with social

DOI: 10.54254/2753-7064/42/20242501

responsibility by fostering content diversity, rather than solely prioritizing profit through trending topic promotions, which could further degrade the overall online ecosystem.

6. Conclusion

The research conducted on information homogenization in Chinese social media platforms, particularly Weibo and Zhihu, highlights significant trends in the concentration of online content. Over the period from July 2022 to April 2024, it was observed that while Weibo maintained a consistently high level of homogenization, Zhihu experienced a sharp increase in content similarity. This rise in homogenization on Zhihu can be attributed to the growing influence of a few polarized topics that dominated discussions and attracted a substantial amount of user attention. These findings suggest that the platform's role as a hub for knowledge exchange is being overshadowed by the repetition of popular themes, driven by user interest and external factors like the regulation of information.

In contrast, Weibo's already high level of homogenization stems from its design as a mass entertainment platform. The platform thrives on user engagement driven by trending entertainment news and social events, which in turn results in the amplification of widely shared content and limits the diversity of information available. The research indicates that while this homogenization is not new for Weibo, it remains a persistent issue that affects the richness of information consumption.

The growing homogenization across these platforms raises concerns about the overall health of the Chinese digital ecosystem. Users are increasingly trapped in "information cocoons," where they are exposed to limited viewpoints and repetitive content. This has significant implications for the training of machine learning models that rely on diverse and high-quality datasets. As social media continues to shape public discourse, platforms like Weibo and Zhihu need to balance commercial interests with the need to foster content diversity, ensuring that users can access a broader range of perspectives and reducing the adverse effects of content repetition.

References

[1] *Kossinets, G & Watts, D. Origins of homophily in an evolving social network. Journal of Sociology, 2009,115(2), 405–450.*

[2] *McEwan, B. Carpentier, J & Hopke, E. Mediated skewed diffusion of issues information: A theory. Social Media + Society,2018,4(3), 1–4.*

[3] *Mikal, P. Rice, E & Kent, G. Common voice: Analysis of behavior modification and content convergence in a popular online community. Computers in Human Behavior, 2014(35), 506–515.*

[4] *Hosanagar, K. Fleder & D. Lee, D. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. Management Science, 2014,60(4), 805–823.*

[5] *Airoldi, M., Beraldo, D & Gandini, A. Follow the algorithm: An exploratory investigation of music on YouTube. Poetics, 2016,57, 1–13.*

[6] *Scheufele, A & Nisbet, C. Commentary: Online news and the demise of political disagreement. Annals of the International Communication Association, 2013,36(1), 45–53.*

[7] *Cheng Shian, Shen Enshao. Explanation and Reconstruction of Organizational Communication Theory in the Digital Era: From the Perspective of Technological Progress and the Evolution of Communication Laws. Journalism University, 2009(2):119-124.*

[8] *Yao Wenkang. The "Information Cocoon" Effect and Reflections on Aggregated News Apps: A Case Study of "Today's Headlines". Media Forum, 2020(3):151-153.*

[9] *Xu Xiang, Ao Ziqi, Shi Jingyuan, et al. Convergence Through Different Paths: Homogenization of Information Coccoons in User-Generated Content on Social Media—An Empirical Analysis Based on Sina Weibo. Journal of Xi'an Jiaotong University (Social Science Edition), 2022, 42(03):133-140.*

[10] *Jiang Zhongbo, Xue Danyang. Analysis of the "Echo Chamber" and "Filter Bubble" in the Era of Social Media. Journalism and Communication Review, 2024, 77(03):101-114.*