Tracing Method of False News Based on Python Web Crawler Technology

Hongmei Ye[1(✉)], Yandan Lu[2], and Gang Qiu[3]

[1] Department of Chinese, Changji University, Changji 831100, China
Yehm1025@126.com
[2] School of Literature and Media, Guangxi Normal University for Nationalities, Chongzuo 532200, China
[3] Department of Computer Engineering, Changji University, Changji 831100, China

Abstract. At this stage, false news is rampant in the new media environment. Due to the wide dissemination channels of false news, the large amount of information and data, and the difficult governance of false news in the news communication industry, this paper proposes a false news Traceability Method Based on Python web crawler technology, builds a false news traceability management mechanism, and combines Python web crawler technology to build a false news traceability evaluation system to achieve the goal of false news traceability, Finally, through experiments, it is proved that the method of tracing the source of false news based on Python web crawler technology has high practicability and accuracy in practical application, and fully meets the research requirements.

Keywords: Python network · Reptile technology · Fake news · Information traceability

1 Introduction

In the new media environment, with the development of Internet technology and the upgrading of mobile terminals, great changes have taken place in people's way of obtaining information and reading behavior. Under the background of massive information, users tend to follow blindly to obtain information, making false news shops. Under the Internet technology, the social media platform is the main platform for information release and transmission. Under the social media platform, everyone can act as the publisher of news and release information in real time [1]. Under the diversified social platforms, the communication modes and purposes of large amount of information have also become diverse. Based on the communication mechanism of false news under the background of new media and combined with the relevant theories of management, this paper analyzes 110 samples of false news in the past 11 years, studies the development characteristics and corresponding governance means of false news in the past 11 years, changes the previous governance at the macro level, and explores based on tracing the source, Make false news on the basis of evidence-based governance. The proliferation

W. Fu and L. Yun (Eds.): ADHIP 2022, LNICST 469, pp. 489–502, 2023.
https://doi.org/10.1007/978-3-031-28867-8_36

of false news in the era of self media makes the responsibility unclear in the process of governance, early warning and accountability of false content become difficult problems. Putting forward the false news Traceability Method Based on Python web crawler technology is the fundamental means to solve the false news. Starting from the definition of the concept of traceability, learn from the successful experience of traceability in other fields, and explore the traceability mechanism suitable for false news [2].

According to the distribution and development trend of social public opinion, it is predicted whether changes in public opinion will be leveled after the source of false news dissemination is analyzed. At present, there is no technical breakthrough in the supervision of false news at home and abroad. A "decentralized traceability database" based on traceability is proposed. By using the "public chain" and "alliance chain" under Python technology and the python network crawler technology under government supervision, the unification from macro to micro is realized, and the double protection of the authenticity of information is realized. False news has existed for a long time, but its communication mode and characteristics are also changing with the passage of time. Therefore, under different circumstances, different management measures must be taken to effectively curb the spread of false news [3]. At present, with the development of network technology and the diversification of social networks, the propagation speed, coverage and far-reaching impact of false news in the new media environment must be traced from the source to achieve comprehensive management if we want to solve the problem at the source.

2 Fake News Information Feature Collection Based on Python Web Crawler Technology

2.1 Fake News Traceability Management Mechanism

The current Traceability Technology is mainly based on Python web crawler, and realizes systematic traceability with the help of the decentralized characteristics of Python web crawler technology. As a comprehensive method and means to solve problems, "mechanism" needs to be tested and proved by practice. Traceability has been proved to be an effective method to solve problems through practice in other fields, and even corresponding mechanisms have been formed in other fields for unified and coordinated operation [4]. At the same time, the mechanism itself contains institutional factors and requires relevant personnel to abide by them. In the above chapters, this paper has carried out a detailed analysis and summary on the release and dissemination process of false news in the new media environment. On this basis, by establishing a set of false news traceability mechanism for overall cooperative operation, this traceability mechanism involves all information receivers, communicators, platform builders, news supervision departments and other parts in the context of new media. On the premise of all parts' participation, analyze the traceability mechanism of false news to make it more systematic and theoretical, so as to guide practice more effectively [5]. At present, the tracing of false news needs to learn more from the application of sources in other fields, and find a set of experience mechanism that can guide its own development in combination with the characteristics of false news communication., As the quality level of journalists

is becoming more and more uneven, many news editors have not received professional training and obtained relevant qualifications, resulting in frequent news communication events that ignore professional ethics and violate the principle of news authenticity, and false news is prohibited repeatedly.

With the explosive development of the self-media industry, users publish information anytime and anywhere, and attract the attention of other self-media users through the combination of pictures, texts and videos. Demand has led to the emergence of information islands. There is information asymmetry between communicators and receivers. In the new media environment, due to the competition between interests, there are also problems such as information asymmetry between communicators. In the face of a large amount of information, it is difficult for media operators to centrally share information. Even if false news occurs, error correction and clarification cannot be unified, and the rumor-refuting platforms are fighting each other. It is difficult to centralize information, and supervision and tracking become complicated and difficult [6]. Finally, some practitioners neglected their duties, copied and pasted content at will and uploaded unverified fake news. After being exposed, they evaded responsibility and shied each other's responsibilities, resulting in the ambiguity of the responsible subjects, and retrospective accountability became a necessary means. Under the massive amount of information, users are highly satisfied with the convenience and timeliness of Internet news, but they are not very recognized in terms of credibility and seriousness. On different media news clients, the same Events are reported in various forms and in huge quantities. Therefore, on this basis, false news reports are also displayed on major platforms in multiple forms and channels. Media platforms urgently need to integrate massive content resources to improve their authority [7]. Based on this, the dissemination model of Chinese online media false information is displayed, as shown in Fig. 1:

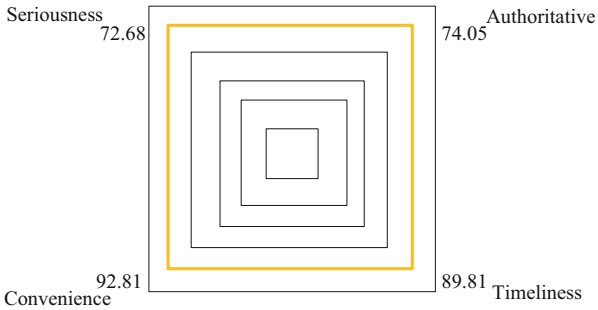

Fig. 1. The dissemination model of disinformation in Chinese online media

The definition of "tracing the source" in the online entry is: looking for the origin upstream, which is compared to tracing the source against the current, and later extended to the pursuit of the source. This term was first proposed by the European Union as a title of a relevant system for food safety. There are roughly three traditional traceability technologies. One is RFID radio frequency technology, which attaches signals to specific items and records relevant data on chips. When you need to know the detailed information of items, you can identify and read out the flow direction of relevant data products

through radio signals: the second is QR code or bar code, Record the batch, place of origin and other information of relevant products through QR code or bar code [8]. In the new media environment, the dissemination of false news presents fission and data network dissemination, and the transmission path is difficult to record and the volume is large. At present, the main means to deal with false news is to refute rumors. However, new media information takes the form of mesh geometry quantitative diffusion, which needs to be realized in combination with the decentralization technology of Python web crawler: decentralization means that the center is no longer the content of information, but the whole information dissemination process from generation, release and forwarding, that is, to control the information life cycle, Thus, a set of accounting mechanism starting from traceability is realized, and on this basis, a set of "decentralized traceability database" based on information is formed. At the same time, decentralization is not absolute "decentralization", but having multiple centers to jointly maintain and supervise the whole link. In the case of multi-party maintenance, the content cannot be easily changed and stored more safely. By providing a decentralized content market, news interviews, content publishing, text copyright and distribution of press releases can be carried out at the same time within this scope, and even a set of agreements can be established through the authenticity of content audit to give certain rewards to publishers. After tracing the source of information through decentralized technology, each news received by users can find the original publisher and the whole communication path. In addition to being forced to trace, the communication of the whole information is also determined in an organizational way. After the "decentralized traceability database" formed under the background of Python web crawler technology is determined, it still needs a set of laws and regulations and credit investigation mechanism to improve and restrict. Based on this, the technical architecture of news management based on Python web crawler is optimized, as shown in Fig. 2.

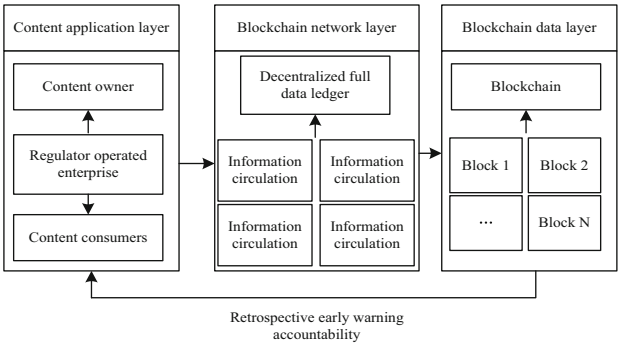

Fig. 2. News management technology architecture based on Python web crawler

In the infection graph G(V2E, P) obtained based on the IC model, due to the large scale of the real social network, it is inefficient to find the source node from all the propagation nodes. Therefore, in order to quickly and efficiently find the initial propagation node, we screen all the propagation nodes, eliminate those propagation nodes that cannot be source nodes, and narrow down the candidate node set. Due to the propagation of false

information in the network, it will definitely cause a certain influence, which means that the source node must have a certain degree of propagation. The dissemination ability is not too low, because if the dissemination ability of the source point is too low, the false information will not be able to spread well. Therefore, this paper defines the propagation capability () of a single propagation node u as the ratio of the number of neighbor nodes that accept the false information propagated by node u to become a propagator among the neighbor nodes of node M to the total number of neighbor nodes of node M, the following formula:

$$I(u) = Ea - \frac{Mn_u}{AN_u} \tag{1}$$

In the formula, n is the number of infected neighbor nodes of node M: N is the total number of neighbor nodes of node u. The propagation capacity of a single node, we can filter all propagation nodes, therefore, this paper takes the propagation capacity (a) The set A composed of all nodes greater than a certain threshold E is called the potential source node set. The value of E is based on the maximum probability that the initial propagation node can be selected into the set A as the standard. This paper believes that if the propagation capability (u) of a node in the infection graph is less than the threshold E, it is considered that the node cannot be used as a network If the dissemination source node of false information exists, the false information will not be able to spread out. Therefore, the setting of the threshold is reasonable without considering the special circumstances, and the specific expression of the node set A of the propagation source of potential false information is shown in the following formula:

$$\Lambda_u = \left\{ I(u) | \frac{n_u}{N_u} > \varepsilon \right\}, u \in V \tag{2}$$

Before the data selected in the above data mining process is formally used as the index of model construction, it is necessary to check whether the data in the table contains noise value or missing value. If so, it is necessary to clean the data, otherwise it will have a serious impact on the data mining effect, which will seriously affect the accuracy of the prediction results of the prediction model. Data cleaning refers to the last procedure to find and correct the identifiable errors in the data file, which is responsible for filling the missing parts of the data, identifying the outliers and removing the impact of noise. Due to the huge amount of data, data cleaning can be completed quickly with the help of the powerful operation ability of the computer. The data cleaning process is shown in Fig. 3.

After the above data mining and data cleaning process, the selected model construction indicators are shown in Table 1 below.

Python web crawler is an artificial network with a wide range of properties, which is composed of multiple neurons. The network is a parallel computing model designed for data analysis and summary based on the structure and operation characteristics of human brain Python web crawler. The biggest feature of the model is that it has good self-learning ability, self-organization, superior fault tolerance and nonlinear mapping ability. In the past, news ratings were affected by various factors such as broadcasting mode, content, time period and other channel programs. However, since the popularity

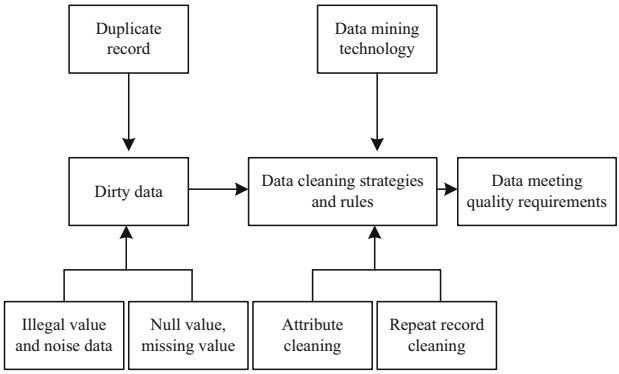

Fig. 3. Data cleaning process

Table 1. Model input indicators

Month	Network click data parameters
1	2.5% (x1)
2	2.3% (x2)
3	2.3% (x3)
4	2.7% (x4)
5	3.5% (x5)
6	2.5% (x6)
7	2.4% (x7)
8	2.4% (x8)
9	2.5% (x9)
10	2.6% (x10)

of various electronic devices, especially smart phones, news ratings are more affected by network click data. The Internet click through rate is affected by various other factors, which will show a certain disorder and nonlinearity. Therefore, it is impossible to accurately predict the fluctuation of news ratings by using linear methods such as regression prediction.

2.2 Fake News Information Management of Python Web Crawler Technology

In the Internet era, the amount of information is large, and the data collection of fake news has become more difficult. Currently, information is mainly disseminated through social media platforms, and the information content of the operating platform is collected and counted to establish a complete resource database. It is more urgent to use the decentralization characteristics of Python web crawler technology to achieve information path restoration [9]. The resource sharing is completed through the "centralized traceability

database". In the database, the background of the database can be trusted. Data such as the number of transmissions are tracked. At the same time, the system contains a large-scale corpus and data statistics of social platform users. Users in the database have an evaluation level, which is divided according to the user's usual communication behavior. The decentralized traceability database uses the corpus and user behavior database. The published content is collected, and at the same time, it will automatically analyze whether there are sensitive words or topics, and automatically generate review results based on the publisher's credit evaluation. After the first stage of verification, the decentralized traceability database will send the content and review evaluation report. To the operator platform, the platform will conduct a second review, the content is completed and passed the review, which confirms the authenticity of the information and then disseminates it to the users. In this way, the content and the disseminator start to check at the source of dissemination, ensuring the authenticity of the source of the content. Specifically as shown in Fig. 4:

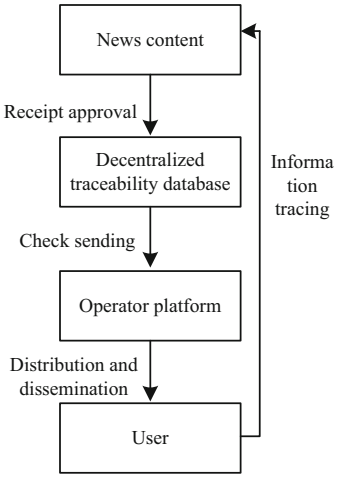

Fig. 4. Decentralized source measurement database information traceability process

In view of the large volume and numerous sources of false news, it is difficult to overcome the problem of tracing the source by relying on the strength of one party alone. It needs the joint cooperation and collaborative governance of macro and micro levels. However, from the perspective of national macro laws and policies, the binding force of law may not be able to take into account and reach all aspects of false news communication. With the development of Internet technology, false news under social media is the product of irrational competition in the communication market, while China's market economy lags behind. At the same time, the media industry belongs to the field of ideology and the forms of media are diverse, Government regulation sometimes fails to grasp the attributes of such industries: as a profit-making organization, the operation platform is committed to meeting the needs of users, often ignoring social benefits. False news is the product of the media's excessive pursuit of individual interests. Regulating users' illegal behavior on social media requires the formation of industry alliances among

industries and the formulation and observance of common industry rules in addition to the management of joint supervision platform, With the wide application of Python web crawler technology, the news industry also needs to keep pace with the times, improve the sense of responsibility, combine the third-party alliance based on "alliance chain" technology built by the operation platform with the "public chain" platform based on users, and jointly complete the source tracing of content, communication path monitoring and audience influence control under the technical means of "decentralized traceability database", as shown in Fig. 5.

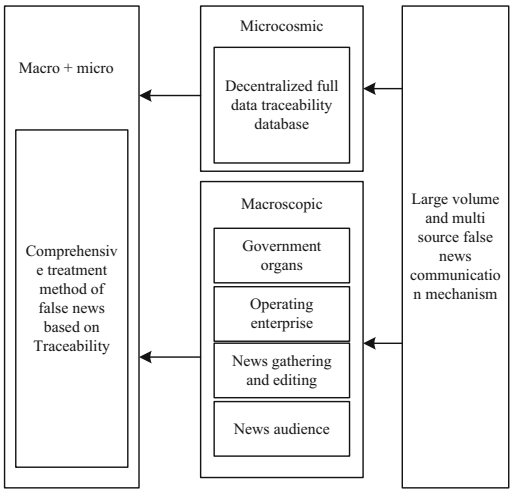

Fig. 5. Integrated management platform for virtual news

Cultivation of news audience awareness of traceability: For news recipients, it is more difficult to identify the true and false information in the context of new media, but it still requires a certain degree of rational judgment. In the future when information is traceable, news audiences It is very important to cooperate with them, not to blindly spread or believe rumors, and resolutely resist news extortion. Macroscopically, abide by national laws and regulations, correctly use various search engines, portal websites and other intelligent terminal platforms, and at the same time be good at using their own supervision power, and have a certain binding effect on news editors. Government agencies, enterprise platforms, news editors, and audience interaction, in order to realize the promotion and development of traceability technology.

2.3 Realization of Fake News Information Traceability Based on Python Web Crawler Technology

The data patterns of different data sources are different in semantics and syntax. When crawlers crawl the data of online social networks or load data from existing data sets, they will inevitably face the problems caused by heterogeneity. Considering the structure of data from a more abstract level can effectively reduce the required workload. Firstly,

the source schema or source attribute name will be mapped to the classes and attributes in the semantic model. An automatic mapping algorithm is usually used to generate candidate mapping rules, and these mappings will be further checked. Then, data conversion is carried out based on these rules. Most of them are one to many and many to many relationships. The identification, mapping and transformation of these relationships require some additional manual operations [10]. However, after the mapping rules are established, the ETL process is automatically executed through the program. Some common patterns used to establish relationships between individuals are summarized in Table 2:

Table 2. Common models of object properties

Pattern	Type	Example
Mode A	Embedded	Microblog posts are returned in JSON format and directly embedded into posts. The relationship between posts and user accounts is embedded
Mode B	External & amp; explicit	In some data sets, posts are stored in a directory named after the platform, so the connection between the online platform and posts is external and displayed
Mode C	External & amp; implicit	An organization has established accounts on different social networking platforms, which are relational, external and implicit

Python web crawler technology based on input and output parameters mainly includes three core components: preprocessor, service filter and service matcher. The overall framework of Python web crawler technology is shown in Fig. 5. Among them, service filter and service matcher involve using the relationship in ontology for calculation. When there are many candidate services, the calculation takes a long time. Therefore, in the preprocessor, the search space is reduced according to the ontology referenced in the semantic description of service and request. After that, the service filter is responsible for further filtering the service, checking the consistency between the request and the input and output of the service, excluding completely inconsistent candidate services, and further reducing the search space. Finally, the service matcher outputs the sorted candidate service list according to the semantic similarity. The overall framework of Python web crawler technology is shown in Fig. 6:

When the information is propagated from the source node to other nodes, the longer the propagation time interval on the propagation path, the later the node that the path reaches will receive the information. The propagation time interval is the forwarding event interval 7684 on each edge on the propagation path. Therefore, if the distance of nodes on the network is properly defined, nodes that are further away from the source node will be activated later. Replacing the source node with another node would violate this fact. Using the above characteristics, the traceability estimation function can be designed. To take advantage of this property, the required traceability function needs to

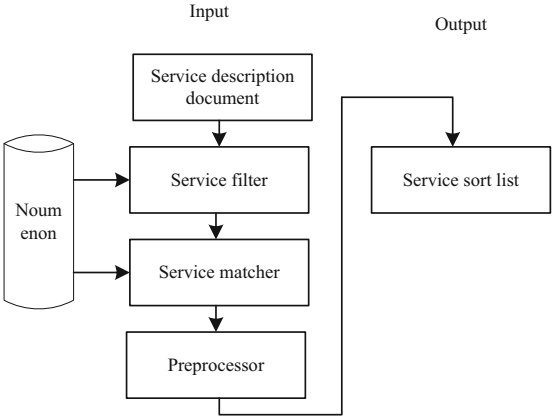

Fig. 6. The overall framework of Python web crawler technology

be able to reflect how close a node is to an early-activating node and far from a late-activating node. The overall framework of the information traceability method is shown in Fig. 7.

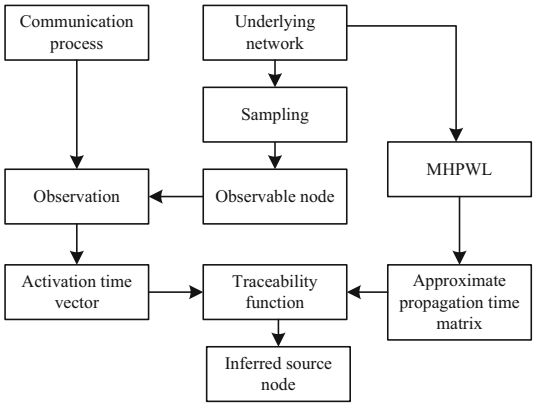

Fig. 7. The overall framework of the information traceability method

Information traceability mainly includes three key parts: (1) using sampling method to select observable nodes and observe them to obtain the activation time vector of information transmission process; (2) The approximate expression of information propagation time between nodes is obtained according to the underlying network; (3) According to the approximate propagation time and activation time vector from each node to observable node, the possibility of each node may be the source node is estimated by using the information traceability function, and the inferred source node s is obtained.

3 Analysis of Experimental Results

In order to better evaluate the method proposed in this paper, this paper sets the edge propagation probability to obey the uniform distribution on (0, 1). The comparison standard is based on the above algorithm idea, using the average error distance and The error rate is used to prove the accuracy of the experiment, that is, the shortest distance between nodes (Hops) to represent the error distance. If the shortest distance is 0, the node evaluated by the experiment coincides with the actual node. And the error rate represents the accuracy of identifying the source point in all infection maps, that is, the average error distance. In order to evaluate the TSRA algorithm proposed in this paper, it is compared with the existing algorithms for the single-point traceability problem.

3.1 Experimental Result

In this paper, 200 experimental simulations are performed on three real-world datasets, and 200 node information C = (V, E) is obtained, and the error distance (Hops) and error rate of the positioning source point are compared. The experimental results are recorded in Fig. 8 in:

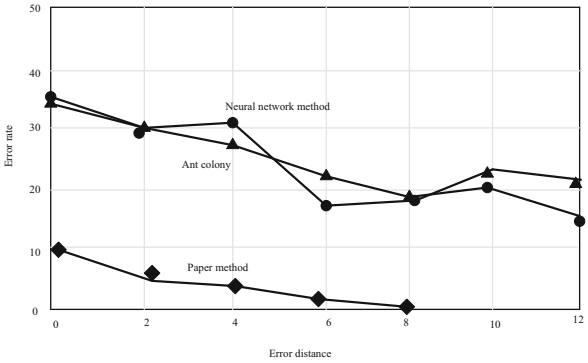

Fig. 8. Comparison of error distances in network information traceability

The traceability time factor is particularly important, and early important participants refer to users with a high comprehensive influence index r who participated in the topic earlier. Define the release time of G as T = Earliest(T: CG}, T = atest(IcCG], and the time T when v ∈ G participates in the cascade, then M = T–1. In order to distinguish the early period when M is small and r is large Important participants, because the scale and duration of different thematic events are different, the method of grading and quantification is adopted for comparison. The Mt level is the total time T-T is divided into 10 segments, and then the high-influence nodes with r > 2 are classified according to the participation level. The time of association is projected into different Δt levels. For example, 14% of the nodes of false information 1 are involved, Fig. 9 shows the comprehensive influence distribution of different news information.

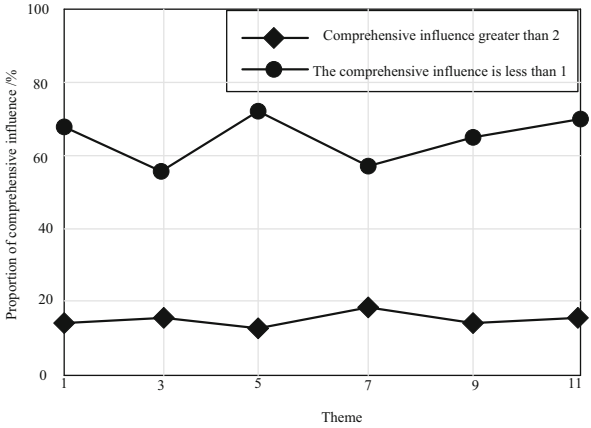

Fig. 9. Distribution of comprehensive influence of different news information

The news information nodes are projected into different Δt levels according to the time of participating in the cascade. For example, false information 1 has 14% of the nodes involved, and the figure shows its projected distribution results. Early participants may include some initiators, so the two should be combined. The traceability results of false information are shown in Table 3:

Table 3. The influence of fake news information and the importance of traceability

Serial number	1	2	3	4	5	6	7	8	9	10
Initiator	60	61	50	67	32	47	58	28	35	58
Key participants	79	49	115	78	46	39	77	68	68	29
Event source	105	119	103	103	89	63	125	98	62	78
A	√	√	√	√	√	√	√	√	√	√
B	×	×	×	×	√	×	×	√	√	√
C	√	×	√	×	√	√	√	×	√	√
D	√	√	√	√	√	×	√	√	√	√
E	×	√	×	√	√	√	√	√	√	√

Where "√" means reaching the index" ×" Indicates that the indicator is not met. It can be seen that the event source is basically locked within 129 D, and the excavation effect is achieved. Compare information traceability methods from an overall perspective. The table shows the success rate of the combination of different traceability methods (Table 4).

This success rate is calculated separately for all propagation processes and observations on each network. As shown in the table, the overall traceability effect of the traceability estimation method in this paper is significantly better than the traditional

Table 4. Success rates of different traceability methods

Information content (GB)	500	1000	1500	2000
Paper method	0.8809	0.9596	0.8785	0.8972
Neural network	0.3026	0.6006	0.5199	0.6581
Ant colony	0.4513	0.3088	0.3285	0.3785

two methods, and the traceability accuracy and traceability time are significantly higher, which can obtain a higher success rate and fully satisfy the research.

3.2 Experimental Analysis

From the experimental results, it can be concluded that the false news Traceability Method Based on Python web crawler technology designed in this paper can basically keep the success rate of traceability above 87% when tracing the source of massive news data, up to 96%, which can be highly practical and accurate, and can be used as a technical means for the news industry to control false news.

4 Conclusion

Based on Python web crawler technology, this paper designs a false news traceability method, aiming at the propagation of false news in massive new media news information, and implements false news traceability through Python web crawler technology. Through experiments, it is verified that the designed false news traceability method can be completed in the actual application process, has high practicality and accuracy, and can effectively maintain the authenticity of network news in the new media environment, Provide technical support for the governance of false news in the news communication industry.

References

1. Fang, Q., Cheng, Y.: Design and implementation of distributed crawler based on Docker container. Electron. Design Eng. **28**(08), 61–65 (2020)
2. Yuan, J., Wang, X.: Art blockchain certificate traceability model based on three chains. Appl. Res. Comput. **38**(10), 2915–2918+2925 (2021)
3. Wang, J.: Tracing and en route filtering analysis of false data based on Python crawler technology. Henan Sci. Technol. **40**(22), 27–30 (2021)
4. Jing, L., Siyu, F., Yafu, Z.: Model predictive control of the fuel cell cathode system based on state quantity estimation. Comput. Simul. **37**(07), 119–122 (2020)
5. Zhu, Q.: Design of public opinion analysis and early warning system based on Web crawler. Electron. Des. Eng. **28**(22), 56–60 (2020)
6. Wang, Y., Zhu, S., Hou, S., Wei, Z.: Application of Python based crawler technology in big data environment. Inf. Commun. **08**, 189–190 (2020)

7. Chun, L.: Design of big data acquisition system based on web crawler technology. Modern Electron. Tech. **44**(16), 115–119 (2021)
8. Yu, H., Zhang, S., Liu, Z., et al.: Propagation source tracing algorithm based on priori estimation. Pattern Recogn. Artif. Intell. **33**(01), 86–92 (2020)
9. Chen, C., Zhou, L.: Tracing and filtering of false data based on Python crawler technology. Comput. Simul. **38**(03), 346–350 (2021)
10. Wang, L., Wang, S.: Dilemma traceability and model innovation: research on personal information cooperative governance based on the blockchain. Chinese Public Administration **12**, 56–61 (2020)