Research on the Methods and Key Techniques of Web Archive Oriented Social Media Information Collection

Xinping Huang

School of Management, Jilin University, Changchun 130012, China
E-mail: huangxinping@jlu.edu.cn

Received 15 July 2021; Accepted 15 August 2021;
Publication 06 November 2021

Abstract

Social media information collection and preservation is a hot issue in the field of Web Archive. This paper makes a comparative analysis of the different social media information collection methods, deeply analyzes the key techniques of the three important parts-collection, evaluation and preservation in the information collection process, and provides the solutions for the problems in the key techniques. Through analysis, the collection method suitable for the social media information is found. In terms of the problem that social websites impose restrictions on the call frequency of API, the paper provides solutions, for example, use the multiplexing mechanism, use the naive Bayesian algorithm to solve the spam filtering problem, and use MongoDB Dbased distributed storage to store collected massive data.

Keywords: Social media, web archive, information collection, long-term preservation, technical strategy.

Journal of Web Engineering, Vol. 20_8, 2473–2490.
doi: 10.13052/jwe1540-9589.20812
© 2021 River Publishers

Introduction

Social media is the technological platform based on Web2.0 technology, which conducts information exchange and communication through publishing, comments and discussions. At present, the presentation forms of social media are becoming more and more diversified, and platforms or tools have emerged to meet the personalized needs of different user groups. Kantar media CIC, an authoritative consulting company, divides social media into social networks (Facebook, Twitter, LinkedIn, etc.), photo sharing (Instagram, Pinterest, etc.), video sharing (Youtube, Vimeo, etc.), interactive media (Wechat, Snapchat, Tiktok, etc.), blog/community construction (Weibo, Tumblr, Wikipedia, etc.) [1]. People can express views, exchange opinions, and share experience anytime and anywhere on social media, so as to participate in the information creation and utilization activities. The huge user group of social media produced massive native information, mainly including user-generated content (or UGC). The information consists of the popular events, hot news, network public opinion, dynamic trends at present, which reflects the traces of human life in the new media era and serves as a social "public memory bank" that condenses the wisdom of social groups. It has the important preservation and utilization value, and is an important object of the current long-term preservation of the network information resources (Web Archive) [2]. As the primary part in the preservation of network resources, information collection plays a key role in the whole preservation process. How to effectively collect social media information has become a new concerned topic in the research field of network resources' preservation.

The information released by Weibo, Twitter, Facebook and other social media are closely related to the context and fleeting, so it is not beneficial to the selection, collection and preservation of information, which has become a key problem restricting the collection of social media information. Social media information is different from the regular website information, most of which is user-generated content, containing a large number of texts, images, audios, videos and other heterogeneous network information [3]. The differences between the two kinds of information, in terms of data redundancy, timeliness, grabbing technology, lead to the result that the network crawler and other information collection methods used in most of the present website web preservation projects is not applicable to social media information collection, therefore, developing network information collection tools oriented towards social media is an important challenge in the long-term preservation of social media information.

Literature Review

In terms of the research on information collection and preservation of social networks, foreign research focuses on how to obtain effective data from social networks, and then analyzes the data from different perspectives, which has made technological breakthroughs in the similar recommendation of user habits and media communication influence. Advanced network technology mainly uses the Twitter application interface (API) for data collection, then stores and analyzes the data in unstructured forms. This data collection, storage, and analysis method provides great technical ideas for many experts and scholars, and some research institutions use crawler methods to obtain user information and operation information provided by Twitter website, and evaluate the user experience and influence through network topology technology. For example, Saito K et al. [4] have specially designed a distributed data collection system that can effectively collect user data in social networks and show that the system has good effectiveness and reliability through relevant tests. Aghdam S M et al. [5] believe that using simple and incomplete data sampling technology cannot fully describe the characteristics of social network. They propose the method of using the open application interfaces to collect social network data, analyze the users' use habits based on the sampling method of random steps, and finally draw the conclusion that the sampling method of random steps has better sampling effect by comparison with other sampling methods. Jones S M et al. [6], based on digital library systems, developed and designed Web Archive oriented long-term storage functional module for social media information using API interface technology provided by social media software, and provided users with the utilization service of archived social media information through the digital library retrieval system.

Compared with developed countries, the domestic research on social media information collection and preservation is still in the initial stage. The related research mainly explores the strategic methods, paths and related technical problems of social media information collection and preservation from the theoretical level. For example, Xiong Zutao [7] proposed to use the Heritrix-based network crawler to collect the page data information of the blog, stored the web pages in a mirror way, then standardized the pages through the DOM tree structure, and finally marked the sensitive information of the page with the page in advance. Liu Chao et al. [8] uses integrity acquisition and CC(knowledge sharing) protocol and API (Application Programming Interface, application programming interface) to

solve key problems such as information acquisition scope, acquisition rights, acquisition methods and so on. Based on the characteristics of Web 2.0 technology, Pan Hong et al. [9] analyzed the feasibility of the block chain technology applied to social media information archiving, and put forward specific application strategies. Zengsa et al. [10] put forward the metadata framework preservation as the file based on analyzing the demands of meta data archiving in the social media.

Methods of Web Archive Oriented Social Media Information Collection

As the starting point of long-term preservation of network resources, social media information collection is the basis of building Web Archive, which is of great significance in the whole process of Web Archive. Social media information collection for Web Archive refers to the use of relevant tools to timely select and obtain social media information worthy of preservation in the given frequency and method. Social media information collection methods for Web Archive can be divided into different types according to the collection object and collection technology. According to the collection technology, this paper divides the commonly used social media information collection methods for Web Archive as follows.

Primary Collection for Long-term Preservation

This is a kind of normal collection method based on Web Archive. Since the middle of 1990s, in order to better preserve network data resources, the libraries or archive organizations in every country around the world have used the method based on the Internet crawler to collect network data and information regularly. This collection method can be divided into local collection, fixed-topic collection and fixed-point collection, etc. [11]. Among them, local collection is the integrity collection based on a specific network domain, this information collection technology mainly uses network information harvesting tools such as robot and crawler to expand from some seed URLs to the automatic collection of the whole web in a specific network domain. Compared with local collection, fixed-topic collection only collects those web pages related to the topic, saves a small number of web pages and can get faster updates, which is closer to the current real situation of the web. Fixed-point collection is the web information collection of sites

with preservation value selected from the specified web information sources according to the established rules.

Secondary Collection of Web-based Data Warehouse Storage

This secondary collection method is represented by lazy preservation, a technology jointly developed by Frank McCcnvn, the Doctor of Computer Science at the University of Harding in the United States and his team, which is mainly used in the social network platform reconstruction to provide solutions for hardware damage, illegal intrusion and missing or incorrect system files [12]. Specifically, the information archiving of Internet Archive always has the lag phase of more than half an year, and the search engine cache is also only limited to the social network information of the latest version. Therefore, the above two are organically combined for the recovery and reconstruction of social networks, and the data acquisition tool Warrick is also specially developed to twice mine the collected information content, thus obtaining more accurate and effective data.

Service-oriented Subscription Collection

Service-oriented subscription acquisition technology mainly means that professional data collection, storage and archiving services are provided by Small-scale Organizations or companies subscribed by the individuals. Web Archive is a long-term large project, requiring a lot of human, material resources and high-level application technology. So in order to save costs, many small and medium-sized enterprises, will transfer the social network information business to professional enterprises, and these professional enterprises will provide data information collection, sort and archiving services. The well-known service enterprises Hanzo Arhiv and eArhive-it companies have launched relevant information service projects [13].

Event-driven Network Information Collection

Event-driven network information collection method is to collect targeted information around a pre-set theme or event, and use the social media information capture tool of intelligent adaptive decision support to realize the collection of event driven social media information on a specific

theme. For example, when collecting social media information resources, ARCOMEM (ARchive COmmunity MEMories) project adopts the special harvesting strategy for domestic and international important historical events, and divides the archived resources into several resource collections around a historical topic according to the content, providing the public with the resource browsing mode of "special collection" [14].

API Based Network Information Collection

At present, mainstream social media software such as Flickr and Twitter support users to embed the acquired data into various mobile apps, website apps, social network platforms, online communities and other specific application services by using open application program interfaces (APIs). Based on this principle, the API interfaces provided by various social media software can be used to realize the corresponding social media information collection [15]. Compared with the above social information media collection methods, this method does not need to develop special crawler tools, but uses the rest API Application interface technology provided by the corresponding social media to collect social media information.

Key Problems and Solutions in the Social Media Information Collection Process

Basic Procedures of Social Media Information Collection

According to domestic and foreign network information collection and long-term storage projects, the social media information collection process is divided into the following steps: the selection of collection service, the permission of copyright owners, the collection of data information, the building of metadata, the audit of collection quality, and information archiving [16]. Of course, information data collection projects appropriately adjust the above procedures as needed, for example, for building metadata, before or after data collection, or even some collection projects will omit data collection procedures; for example, network collection projects restricted by some laws and regulations cannot be restricted by the copyright owner. In the actual information and data collection process, the above collection procedures need to be referred to, but the specific analysis of specific problems should be needed, so that the network collection process can be carried out in a regular and orderly manner.

Key Issues and Solutions in the Social Media Information Collection Process

Information selection

The following content will take Weibo in social network as an example. Generally speaking, Weibo only provides functions such as Following function, blog forwarding and blog comments, which shows that the information correlation of Weibo will be higher than that of web content. Among them, the correlation of Weibo is mainly reflected in the two aspects: on the one hand, to use the following function of Weibo to establish connections; on the other hand, to establish connections through the functions of the forwarding and comments of Weibo. It needs to be noted that the relevance of Weibo will have rich semantic information, if the association is organized and saved, it will not only be conducive to meet people's personalized needs for Weibo and save the important index information, but also ensure the reliability, integrity of Weibo and the information security. In April 2010, Twitter entered into an agreement with the Library of Congress. As required in the agreement, Twitter should unconditionally share all available Twitter information with the Library of Congress so as to seek functional maturation and information expansion services from the Library of Congress. There is no denying that to better meet the needs of users, it is very important to ensure the integrity of information collection.

Of course, the disclosure of social network information is the basis of information collection and preservation, but the users' private information should also be respected and protected in the collection process. It is because of its sharing and openness that Weibo is recognized worldwide, so users only have to register one Weibo account to view the information of other users at will. In order to better reflect the respect for users' information, users can use functions such as the cancellation of following, setting password access and setting up the account only visible to their friends, so as to make some microblog information more private.

Information collection

Social networking is a virtual network. In order to better obtain experimental data, this study requires obtaining users' information at random rather than the extraction within a specific user group. The information obtained by the data collection system mainly includes users' personal information, users' relationship information and microblog information. The collection system designed in this study adopts the application interface provided by the Weibo

platform to obtain the data of the social network, and the breadth-first retrieval strategy is used in the collection process, including the seed node, the data collection and the database. The architecture diagram of the data collection system is shown in Figure 1 below.

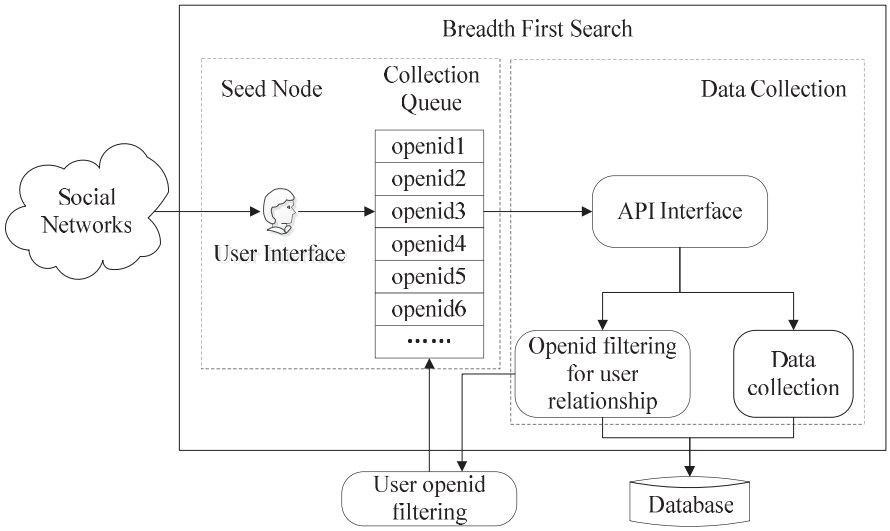

Figure 1 The architecture diagram of data collection system.

As shown in the figure above, the seed nodes can be interpreted as Weibo users, whose collection team is listed as a fan list. It can be authorized using OAuth2.0, paid a visit after authentication. According to the collection data, the seed node should be randomly chosen, so a specific user is not set as the starting point of the collection information, but the list of celebrity users is used as the starting point of the seed node. It should be noted that celebrity user openid is the only identification of users and all information of these celebrities is added during the data collection process.

Data collection technology is to use the open application interface to collect users' personal information, users' relationship data and blog data, among which the users' relationship data includes the users' attention to other users' list and fan list. In the specific operation process of collecting users' relationship data, the system will refilter the collected celebrity users and fan list, and repeatedly perform the iterative collection, and always adhere to the unity and integrity of data collection; when collecting blog data, the system mainly uses celebrity users or fan list to filter, collect and delete processing.

Considering the technical advantages of the network crawler, the system adopts a multi-threaded network crawler mode. The workflow of the system and the multi-threaded call mechanism are introduced below. The work flow chart is shown in Figure 2 below:

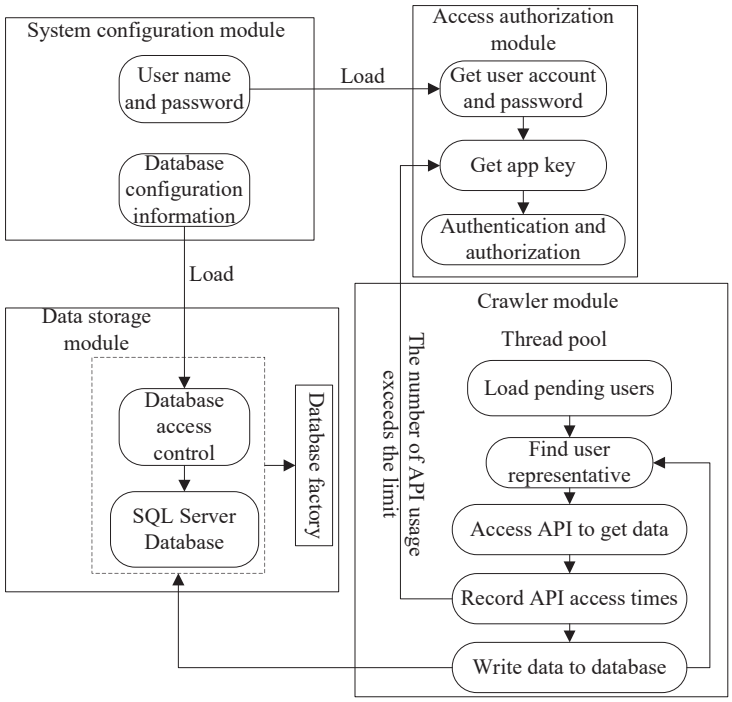

Figure 2 The flow chart of the system work.

As shown in the figure above, the whole system can be divided into four modules. The first module is the system configuration module. When the user wants to have access to the system, he or she should enter the user name and password and then load the corresponding database and data table; the second module is to obtain the authorization module. When the user enters the account password, the system will randomly select App Key to match, and the system uses Weibo OAuth2.0 technology to enable the data scraper technology to proceed smoothly. The third module is the data storage module. After the system loads the corresponding configuration data, the SQL Server database is used to conduct data reading and storage. It is worth noting that this part is the core of the whole data processing, and the system directly operates the database using the SQL language. The fourth module is the

crawler module, the system uses a multi-threaded way to climb the Weibo information, and when the application interface usage exceeds the limit, it will use other stored App Key to reauthorize the crawl.

(1) Design of the Collection Module of Users' Relationships

Social networks are mainly established through the mutual attention of users. Both the analysis of users' habits and information communication analysis need to grasp the interconnection between users, so they need to collect all the users' relationship data. The collection of users' relationship data mainly includes the user's attention list and the followed users' attention list. The flow chart of the collection of users' relationship data is shown in Figure 3 below.

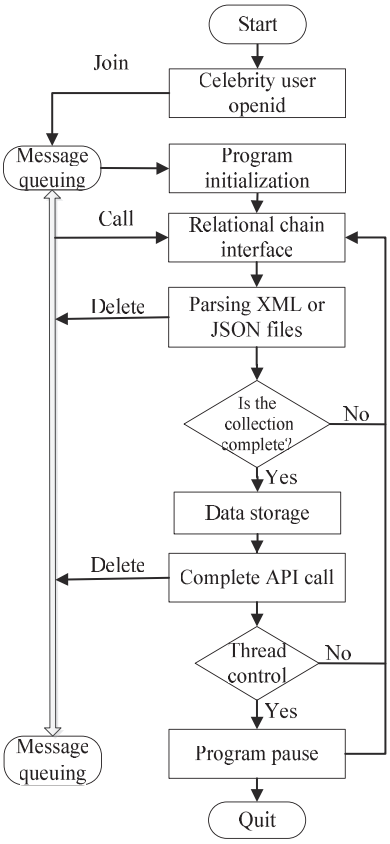

Figure 3 The flow chart of user relationship data collection process.

(2) Design of Blog Post Content Collection Module

After collecting users' relationship data, the system can use the openid of celebrity users to collect the latest blog information. The flow chart of blog content data collection is shown in Figure 4 below.

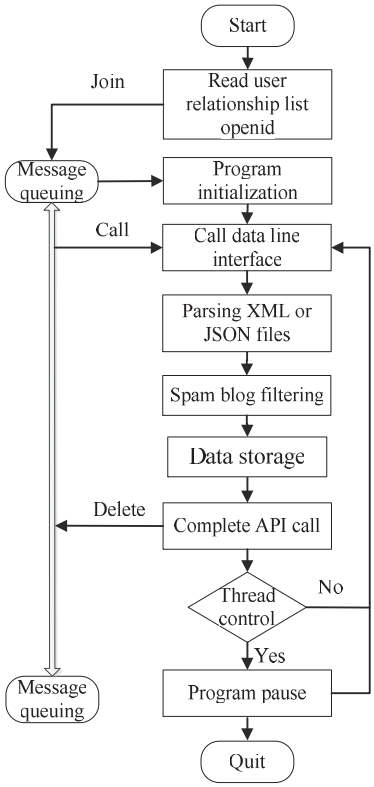

Figure 4 The flow chart of blog content data collection.

First, the required Weibo celebrity user openid, is retrieved in the database, then put the openid into the message processing queue, and use the timeline interface program to locate the latest blog message information of a specific user, then analyze the JSON or XML file, and finally collect the required blog message information. This study uses the hash method to delete duplicated data of users' relationships, so when truly collecting users' blog information, the hash method is used to remove the duplicated celebrity user -openid in the message queue at first, to avoid duplicated data collection based on the ID of the user and avoid excessive waste of resources.

Metadata management

For subsequent information management, storage and utilization, it is necessary to edit and rank the collected data information based on metadata, including creating metadata, creating metadata specifications, setting storage rules, application scopes of metadata, storage location, etc.

Filtering of useless information

Social network is an open and equal network platform, as long as the users register in the network, they can subscribe and publish information on the network due to the openness and arbitrariness of social network. So the restrictions on users' published blog text are not very big, of course, the blog content effect based on machine learning that the platform used is not very good. Therefore, on various network platforms, when preparing to collect blog information, the blog information need to be filtered and selected at first, for example, advertisements and uncomfortable remarks should be eliminated, leaving effective blog information to facilitate the deep mining of blog information.

Data classification is an important procedure in data mining technology. Text classification is a step-by-step learning process, which can be classified based on the text characteristics and classification model of information, and the text characteristics and classification model are available based on the provided datasets. Simple Bayesian algorithm is also a typical way of machine learning. This algorithm has the advantages of high efficiency, high accuracy in classification and simple implementation. It is a classification model with mature technology and superior algorithm, which is very suitable for screening and filtering garbage information in text [17]. The filtering method adopted in this research is implemented using the naive Bayesian algorithm, which divides blogs into two categories, normal blog and spam blog, respectively. The specific blog message information filtering process is as follows:

(1) Chinese segmentation

The significant difference between Chinese texts and English texts lies in the natural separator. If you want to get the characteristics based on Chinese texts, you need to process Chinese texts at first. The common word segmentation methods are based on semantics, dictionaries and frequency statistics, among which the dictionary-based segmentation method is the most simple.

(2) Remove the stop words

Generally, there are more modifiers in Chinese than in English, such as function words which mainly play a modification role, and do not have too much impact on the understanding of Chinese texts. So in the Chinese text segmentation processing, the deactivated vocabulary list will be referred to, if the stop word in the text is in the deactivated vocabulary list, the word is deleted. If not appearing in the table, the word is retained. In the research of establishing deactivated vocabulary, the natural language processing laboratory of Harbin University has achieved remarkable results, where the deactivated vocabulary has been built to facilitate reference for developers.

(3) Chinese classification

Chinese classification requires training in Chinese data. This study uses sogou-based Chinese database and useless database to train a special dataset. The dataset of the normal blog text is trained by the normal Chinese database, while the g useless data set is trained by the useless database, and finally the naive Bayesian algorithm for filtering and classification is used, so as to filter out the useless blog information. The useless blog filtering framework diagram is shown in Figure 5 below.

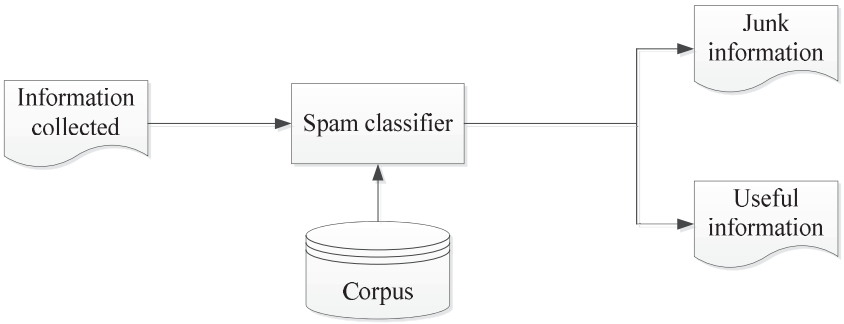

Figure 5 The useless blog filtering framework.

(4) Naive Bayesian algorithm filtering

The basic processing method of simple Bayesian algorithm is: build Chinese text as a kind of text collection $D(d_1, d_2, \ldots, d_n)$, then assume that there are categories $C(C_1, C_2, \ldots, C_n)$, and then bind the text with the most closely related category, that is, the posterior probability of each text information $(p_1, p_2, \ldots, p_n)$ according to the calculation of a certain category

$C_1, C_2, \ldots, C_n$, when a posterior probability $(p_1, p_2, \ldots, p_n)$ is greatest, the text d_i belongs to that specific category. In order to facilitate the calculation, the Bayesian algorithm defaults to the influence of each text attribute on a specific category, and its specific calculation formula is shown in Equation (1).

$$P(C_1|d_i) = \frac{P(C_j)P(d_i|C_j)}{P(d_i)} \tag{1}$$

It should be noted that the text probability $P(d_i)$ is a constant, which is not difficult to see according to the above formula: when the maximum value $P(C_j)P(d_i|C_j)$ is taken. The prior probability formula is $P(C_j) = n_j N$, among which $P(C_j)$ is the probability of text C_j in the training library, the number n_j is the number of text based on the classification of C_j. Specifically, for a text $d(e_1, e_2, \ldots, e_n)$, it exists $P(C_1|d_i) = P(e_1, e_2, \ldots, e_n|C_j)$, but due to the large joint probability computation $P(e_1, e_2, \ldots, e_n|C_j)$, the independent conditions assumed by the naive Bayesian formula are reduced to:

$$P(d_i|C_j) = P(e_1|C_j)P(e_2|C_j)\cdots P(e_n|C_j) = \prod_{i=1}^{n} P(e_i|C_j) \tag{2}$$

Therefore, this study only needs to compute the posterior probability $P(C_j|d_i)$ of the category corresponding to each text blog message, which is deleted if the posterior probability based on junk blog information reaches the greatest.

Information store

According to the actual application requirements, the main objects collected by the system are blog information data and users' relationship data, and the designers can collect different types of data according to different needs. According to the characteristics of large social network information volume and fast upgrading speed, the system uses MongoDB-type and text file methods to store data respectively. Finally, according to the specific experimental requirements, the database is divided into three parts, namely the blog library, the attention user library and the fan user library.

(1) MongoDB data storage design

MongoDB is a NoSQL database oriented toward document storage service, owning high degree of freedom and consisting of multiple documents that can be nested between them. According to the collected data characteristics

and the storage characteristics of MongoDB, the whole database is divided into three parts for storage, respectively blog collection, users' attention list collection and users' fan list collection. The specific diagram of the relationship between each set is shown in Figure 6, in which the embedded tag is represented as an embedded document.

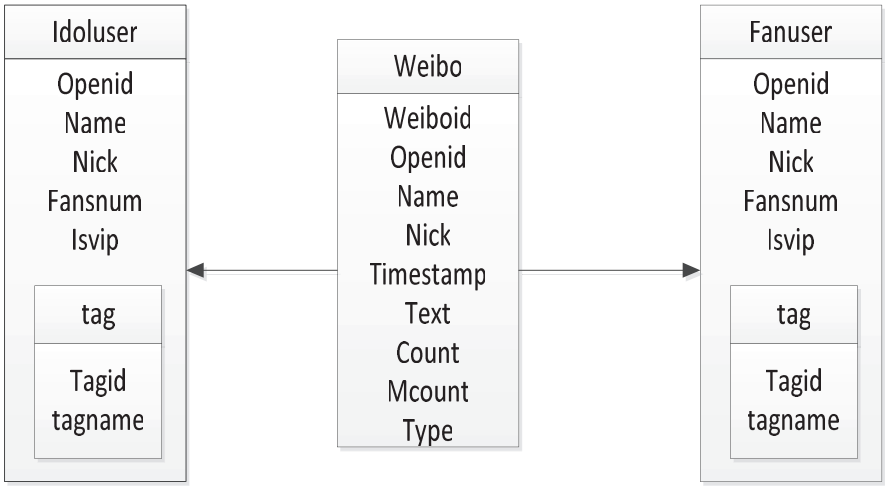

Figure 6 The relationship between data sets.

(2) Text File Storage

This study can also store data in a text file mode, firstly using the application interface to collect data and return it in a JSON or XML file format. The whole system adopts the data flow of data processing, so that the data processing is independent of the hardware facilities. Text file mode not only provides convenient conditions for later data processing, but also is more intuitive, but when the amount of data increases, the processing efficiency will deteriorate. MongoDB storage can deal with mass data, when ensuring the quality of storage, it also has a strong query function.

Conclusion

The huge use group of social media has produced a large amount of native information, among which user-generated content mainly condenses the ideological wisdom of social groups, and has the characteristics of originality,

timeliness, wide influence and transiency. As the "public memory bank" of the society, this information has important scientific research value and is an important object for the long-term preservation of the current network information resources. As the primary link of network resource preservation, information collection plays a key role in the whole preservation process. This paper studies the social media information collection methods and corresponding key technologies. Aiming at the problems faced by the key links such as acquisition, screening and preservation of social media information collection, such as limited API call frequency, useless information filtering and massive heterogeneous data storage, technical strategies such as multi app key reuse mechanism, naive Bayesian machine learning classification method and distributed storage based on JSON data storage Mongodb are proposed respectively. The research results have certain reference value for solving the technical problems of the information collection and preservation faced by the current long-term preservation practice of social media information. The research of this paper is limited to the theoretical exploration of technical strategies. In the future, it is necessary to experiment the social media information collection method for Web Archive, design and implement the corresponding prototype system to verify whether the proposed methods are effective and whether the corresponding technical strategies are practical.

Acknowledgments

This research is funded by the National Social Science Foundation of China, grant number 18CTQ040.

References

[1] Kantar Media CIC. 2020 China Social Media Research Report [EB/OL].[2021-06-07]. http://www.ciccorporate.com/index.phpoption=com_content&view=article&id=1079&catid=archives-2020&Itemid=194&lang=zh.

[2] Zhang Yan. Research on the Transformation of Archival Memory Reproduction in New Media Age [D]. Shanghai: Shanghai University, 2020.

[3] Huang Xinping. Comparison and Its Reference of Social Media Information Long-term Preservation Projects Founded by FP7 in the European Union [J]. Research on Library Science, 2019(17):2–9.

[4] Saito K, Kimura M, Ohara K, et al. Behavioral Analyses of Information Diffusion Models by Observed Data of Social Network[C]. 3rd International Workshop on Social Computing, Behavioral Modeling and Prediction, Bethesda, MD, MAR 30–31, 2010.
[5] Aghdam S M, Khansari M, Rabiee H R, et al. WCCP: A congestion control protocol for wireless multimedia communication in sensor networks[J]. Ad Hoc Networks, 2014, 13:516–534.
[6] Jones S M, Klein M, Weigle M C, et al. MementoEmbed and Raintale for Web Archive Storytelling [EB/OL]. [2021-06-11]. https://arxiv.org/pdf/2008.00137.pdf.
[7] Xiong Zutao. Analysis of Micro-blog Public Opinion based on Text Information Extraction from Webpage [D]. Xi'an: Xi'an University of Science and Technology, 2013.
[8] Liu Chao, Zheng Jiancheng. Discussion on the Key Issues of Chinese Micro-blog Information Collection from the Perspective of Long-term Preservation [J]. Library and Information Service, 2015, 59(3):134–139.
[9] Panhong, Wang Zipeng. Application of blockchain technology to social media information archiving [J]. China Archives, 2018(06):74–77.
[10] Zeng Sa, Huang Xinrong. Construction of social media file archive metadata scheme in China [J]. Research on Library Science, 2020(20):58–66.
[11] Huang Xinping, Wang Ping. Recent Home and Abroad Studies on Progress of Web Archive Technology Research and Application [J]. Research on Library Science, 2016(18):30–35+19.
[12] Liu Lan, Wu Zhenxin. Research on Web Archive Information collection process and key issues[J]. Information Studies: Theory & Application, 2009(8): 113–117.
[13] Library of Congress. Update on the twitter archive at the Library of Congress [EB/OL]. [2021-06-17]. http://www.loc.gov/today/pr/2013/files/twitter_report-2013jan.pdf.
[14] Thomas Risse, Elena Demidova, Stefan Dietze, etc. The ARCOMEM Architecture for Social and Semantic Driven Web Archiving [J]. Future Internet, 2014, 6(3):688–716.
[15] Intelligent Archiving of the Social Web [EB/OL].[2021-06-17]. http://pierre.senellart.com/talks/diadem-20111003.pdf.
[16] Huang Xinrong, Gao Chenxiang. Review of social media archiving technology from the perspective of process [J]. Research on Library Science, 2019(02):2–11.

[17] Zhang J, Feng S. Machine Learning Modeling: A New Way to do Quantitative Research in Social Sciences in the Era of AI [J]. Journal of Web Engineering, 2021, 20(2):280–301.

Biography

Xinping Huang received the Ph.D. degree in information science from the Jilin University, Changchun, China, in 2017. He is currently an associate professor with the Department of information management, School of management, Jilin University, China. His current research interests include Information Management System and Web Archive.