Research on Virtual and Identity Identity, Correlation and Analysis Technology of Internet-related Economic Crimes Based on Big Data

Jing Yuan¹ and Min Tu²

¹Beijing Aerospace Ideal Technology Co, Ltd, BeiJing, 100000, China
²Jiangxi Police College, Nanchang, jiangxi, 330100, China
Email: ¹yuanjing@asideal.com.cn, ²tumin_y@163.com

Abstract. In recent years, social activities on the Internet have been frequent, and people have carried out network social activities by registering different IDs on various platforms, which has formed a new identity - Internet virtual identity. A large number of virtual identity exchanges form a social network relationship diagram, but the network virtual identity does not require real-name authentication, which increases the security maintenance and supervision of the network. Therefore, the investigation of network-related economic crimes under big data faces new challenges. In response to this situation, a combination of virtual and real identity, association and analysis techniques based on big data is proposed. Firstly, through the integration and processing of heterogeneous data, the data resources integration under the condition of decentralized data source and multi-database data source is realized. Secondly, the verification and analysis platform of Internet information is built by Solr which is an open source framework. Finally, using full-text search technology to achieve target search and virtual identity association analysis under big data.

1. Introduction
The rise of the Internet era has made online virtual socialization more convenient and more frequent. People use a large amount of virtual identity information on various platforms, including static data of users registering virtual accounts, and dynamic data such as interactive information of virtual accounts. A massive information resource pool has been formed. The massive information resource library has three characteristics: 1) The information storage has a large amount of data and the update frequency is fast. The data stored in the network can reach terabytes or even petabytes. There are frequent social activities between a large number of virtual accounts, and information is updated all the time. 2) Diversification of information structure. Information data sets are disorganized and diverse, including both structured and unstructured data, such as numbers, text, images, and even multimedia files. 3) Information distributed storage. There are many platforms for information sources, and the types of information storage on each platform are different. The data on each platform is relatively isolated, lacking in relevance, and cannot be interoperable. In the past, the public security department's verification of identity can only use the internal data of the public security for identity verification. However, since criminals often use Internet tools to communicate, how to effectively link Internet data with the internal resources of the public security department in such large and complex data, and to link and associate false and real identity has become an urgent problem to be solved.

Big data-based virtual and real-identity string concatenation, correlation, and analysis techniques rely mainly on the development of technologies such as big data storage, big data mining, big data retrieval technology, and text analysis. Scholars at home and abroad have conducted in-depth research
and discussion on this issue from different angles and different focuses, but there is no complete solution.

This paper proposes a virtual data-based identity, correlation and analysis technology based on big data, and builds a verification and analysis platform for Internet information. This platform adopts a fast and convenient search method, which is convenient for quick verification of Internet information identity, and the search results are subjected to secondary research and judgment. The results can be displayed in a visual manner, and the relationship between characters can be quickly analyzed and analyzed.

2. Demand Analysis and Research Content

In the process of public security investigation and analysis of incidents and the investigation and judgment of cases, there will be a large amount of network data. It is not enough to rely on the internal database of the public security organs for analysis. It is also necessary to use Internet data for correlation analysis. However, the current methods of investigation are cumbersome and inconvenient, and the data obtained cannot be effectively correlated with the internal data of the existing public security organs and social resources for analysis and utilization. Therefore, it is necessary to integrate data from multiple platforms and multiple sources, and use the data reasonably for retrieval and analysis, and display the results.

The core content of this paper is as follows:

1) With the advent of the Internet age, the rapid development of big data networks is also a hot topic in recent years, because its supervision is difficult, communication is convenient, and the cost of illegal crimes is low. In order to make up for the gaps in the current field, this topic establishes an effective analysis model by analyzing the characteristics of network virtual identity, which provides a basis for subsequent related research.

2) The existing technology cannot solve the problem of the association between virtual id and real identity. This topic uses heterogeneous data integration to complete the construction of the data layer, and then realizes the virtual and real identity association function through the full-text retrieval technology and the virtual and real identity association technology.

3) Based on the above key technologies, a verification and analysis platform system for Internet information with comprehensive data comprehensive analysis and analysis capabilities has been developed. This platform can intelligently search and analyze the search targets and visualize the search results. This research has realized the theoretical research of virtual identity and real identity association technology to the transformation of actual results.

3. System Architecture Design

3.1 System Architecture

The entire Internet information verification analysis platform system architecture consists of three levels: data layer, core layer, and application layer.

Data layer: The full-text index library is the data foundation of a check-through system. It is mainly built by various business databases and eight information resource libraries through information scanning. These data are not simply stored in files, but are extracted, cleaned, and converted, integrated, loading and other steps to integrate each database (table) into the index library for use by the information query system. It not only stores the basic information of the shared data, but also stores the relationship between the shared data. Therefore, the data sources for building a comprehensive index library mainly include: other business databases, eight information resource libraries, and index information.

Core layer: The core layer is the core part of the verification and analysis platform of the entire Internet information, and is the connection between the data layer and the application layer. Its main function is to receive the external instructions transmitted by the application layer, and through the
full-text search technology, search and analyze the data of the data layer, and display the results in a visual form.

Application layer: The application layer includes an information access portal and an external system interface. The information access portal is a unified web interface for users to log in to the information query system and use various functions provided by the system. The external interface includes: an interface with the CA authentication system, an interface with the data exchange platform, an interface with the request service platform, and an interface with the mobile police.

3.2 System Operating Environment
The system operating environment refers to the relevant software and hardware environment designed and used to ensure the normal operation and normal service of the full-text retrieval system. It includes the following aspects:

Network layer: various types of network equipment on the information network of the public security department.

Host and storage layer: Servers and storage devices with different performances, including database server, application server, HTTP server, etc.

Operating system layer: including mainstream operating systems such as Windows, Linux, and Unix.

System software layer: including database system software and middleware products.Application software layer: full-text search system and other application systems.

System operation management system: covers all aspects of system operation, including computer room management system, personnel management, auditing, etc.

System security management system: Covers all aspects of technology composition, including operational security, application security, data security, host security, network security, and physical environment security.

Browser: The user uses the operation entry of the full-text search system.

HTTP server: After deploying related software, the HTTP server can provide task distribution for the application server distributed, so that the application server cluster where the full-text retrieval system is located is load-balanced. To improve system security and stability, it is recommended that multiple HTTP server groups be clustered and equipped with intelligent switches to distribute access requests from front-end users to different members of the HTTP server cluster.

Application Server: Run the full-text search system and middleware software on the application server. It provides the system with the core functions of information query application and full-text search library construction. It needs to have large memory space, powerful computing power, high concurrent processing capability, and high reliability and maintainability. In order to improve the reliability and maintainability of the full-text search system, it is recommended to deploy in a cluster mode.

Database server: The database server will store and manage data. It needs high centralized processing capability and high concurrent processing capability, and has high reliability and maintainability. In order to ensure data security, optimized storage and easy management of the full-text retrieval system, a cluster cluster system is used to build a database server cluster system.

3.3 System Flow
System workflow:
1. Configure the data source parameters and website parameters to be retrieved, so that the system can connect and have the right to query the database and access the website information.
2. Using the above parameters to scan and extract the data source, construct a full-text index library file, and store it in the storage of the server. In this process, it is necessary to perform the rule according to the full-text search.
3. Each system module queries and retrieves the full-text index library through the corresponding query retrieval strategy.
4. Key Technologies

This project needs to establish full-text indexing for each business database, resource library, and social resource library. Provide users with a cross-platform, cross-database one-stop search platform to fully mine the clue information hidden in each system. There are two main types of data used in the current big data virtual and real identity string, association and analysis techniques: structured data and unstructured data. The structured data is data of a fixed format or length, and the unstructured data is data of no fixed format or variable length. The retrieval of unstructured data is also called full-text search. Therefore, the research focus lies in heterogeneous data integration, text clustering methods, full-text retrieval technology and virtual identity association technology.

4.1 Heterogeneous Data Integration

Currently, databases are storing massive amounts of data, both structured and unstructured, but these databases are distributed across different systems. There is an increasing demand and situation in which various business systems take data from these databases, which is extremely inconvenient for data management and analysis. This requires a unified data management and access platform to facilitate unified maintenance and management, and to provide "one-stop" data access services. However, due to the large amount of data, the resources of a single machine are limited. The organization of data is not the physical integration of data, but the logical integration based on business needs and data conditions.

We add a data integration layer between the source data layer and the upper application layer to isolate the source data from the upper application. The main function of this data layer is to clean and integrate the source data and store it in a new database. It not only ensures the integrity of the source data, but also speeds up the data retrieval by the upper application.

Processing for unstructured data: A structured database as a data source requires an open database interface for the metadata management system to extract data structure information from the source database and store it in the metasystem. The service generation module can query each business system metadata stored in the metadata system. It automatically generates code blocks for extracting data by simple operations such as checking and combining fields. At the same time, the partial code block is packaged into a webservice service, stored in the service running module, and the service is registered to the enterprise service bus to provide data services for the outside.

For NoSQL databases, because there is no uniform data structure, it is impossible to automatically generate code blocks and publish them as services in the above manner. However, it is still possible to generate a webservice service by customizing the service interface, and then integrating it through the ESB and publishing it to the data integration platform to uniformly provide external services. In this case, custom development of webservice can only be done for each interface.

4.2 Text Clustering Method

Text clustering is mainly based on the well-known clustering hypothesis: similar documents have large similarity, but different types of documents have less similarity. As an unsupervised machine learning method, clustering requires neither a training process nor manual labeling of documents in advance. It can be divided into partitioning method, hierarchical method, density-based method, grid-based method, and model-based method.

This project plans to use the improved K-MEANS algorithm for text clustering. The principle of K-MEANS is to give a set of data points and the required number of clusters k (k is specified by the user). The k-means algorithm repeatedly divides the data into k clusters according to a certain distance function. However, for the K-means algorithm to converge to the local minimum, rather than the global minimum, it needs to be improved. Therefore, SSE (Sum of Squared Error) is used to measure the clustering effect, that is, the sum of squared errors, which is the sum of the squares of the error distances of all the data points in all clusters to the center of the cluster. The idea of the algorithm is to minimize the SSE value, which means that the data points are closest to their cluster center point, and the clustering effect is best. Therefore, the entire sample set D is first treated as a cluster, and then the cluster is divided into two clusters, that is, the K-means algorithm of K=2 is calculated. Then select one of the clusters to continue the partitioning, and which cluster is selected depends on whether it is
possible to minimize the value of SSE. The above division process is repeated until the number of divided clusters reaches the K value. The specific process is as follows:

1) input a data set ( ) with a sample number of m, and determine the K value;
2) Treat the entire data set as a cluster and calculate the mean of the data points in each cluster as the cluster center;
3) Perform a K-means algorithm for K=2 on a cluster;
4) comparing the SSE values of the respective clusters in the divided data set, and repeating the step 3 in which the cluster having the smallest SSE value is repeated;
5) After multiple iterations of steps 3 and 4, until the number of cluster centers reaches a preset K value.

4.3 Full-text Search Technology
At present, the search for structured data is relatively mature and fast, and database data or metadata can be searched by using SQL statements or by using Windows search.

There are two ways to search for unstructured data: sequential search and full-text search. In the case of a small amount of data, sequential search is convenient and simple, but in the case of big data, the speed of this method is very slow, and the use of full-text search can achieve faster search.

Full-text search draws on the search idea of structured data, reorganizes unstructured data and establishes a certain structure, and then searches the data. Therefore, the most critical technology for full-text retrieval is index creation and index search. Index creation: The process of extracting information from all structured and unstructured data to be searched and then creating an index. Search index: Receive the user's query request, then search for the created index, and finally return the result.

The technology is implemented based on the full-text search engine Solr. Solr is a standalone enterprise search application server that provides an API interface similar to Web-service. The user can submit a certain format XML file to the search engine server through the http request to generate an index. It is also possible to make a lookup request through the Http Get operation and get the returned result in XML/Json format.

4.4 Virtual Identity and Real Identity Association
To identify and correlate virtual identities, an identity model must be constructed to calculate the degree of association between different identity keywords in the word vector matrix. Then, find the weight coefficient of the keyword. The correlation calculation is performed after the linear representation, and finally, it is confirmed whether it is the same virtual identity according to the result of the association calculation.

Due to the diversity of data sources, in addition to the association of different virtual ids, the association between virtual identity and real identity can also be made. And related extensions of other information, such as ID number, telephone number, license plate number, address class, etc. The system supports the association, divergence and mining of the searched information according to the business process flow, and displays it in a visual image manner (displayed according to the analysis model).

5. Conclusion
This paper combines heterogeneous data integration, full-text search technology and traditional identity-related technology to develop an Internet information verification analysis platform. The communication number data collection and analysis system of the Internet information verification analysis platform uniformly cleans and integrates the data. At the same time, the text is clustered, and then the index is built to improve storage efficiency, and the data is correlated and analyzed. Then, through the information verification module, in response to the user's request, the data of the Internet communication number data collection and analysis system is analysis and judgement, and displayed to the user through visual interaction.
6. References

[1] DeJong, G. An Overview of the FRUMP System [J]. Strategies for Natural Language Parsing. 1987. 149-176.

[2] Tait, J. Automatic Summarizing of English Texts [D]. Ph.D. Dissertation, University of Cambridge. 1983.

[3] Boguraev, B., and Kennedy, C. Salience-Based Content Characterization of Text Documents [C]. In Proc. of the ACL’97 Workshop on Intelligent Scalable Text Summarization. 1997.

[4] Dave, K., Lawrence, S., and Pennock, D., Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Review [C]. WWW’03. 2003

[5] Wu Peng; Construction of virtual identity in network art [D]; Central Academy of Fine Arts; 2017

[6] Deng Wei, Han Weihong, Liu Dong, Xiong Ying, Research on eID Virtual Identity Data Storage [A], Proceedings of the 28th National Computer Security Academic Exchange Conference [C]. 2013

[7] Deng Wei, Research and implementation of key technologies for storage management of massive virtual identity data [D], National University of Defense Technology, 2013

[8] Zheng Jie, Research on the construction of virtual identity knowledge map for network threat discovery [D], Jinan University, 2015

[9] Lan Shaowu; Extraction and Analysis of Network Virtual Identity Relationship [D], Beijing University of Posts and Telecommunications, 2017

[10] Peter Harringto, Machine Learning in Practice [D]: People's Posts and Telecommunications Press, 2013

[11] Zhou Zhihua, Machine Learning [D]: Tsinghua University Press, 2016

[12] https://blog.csdn.net/google19890102/article/details/26149927