RESEARCH ARTICLE OPEN ACCESS

Correlation Method for Public Security Information in Big Data Environment

Gang Zeng*
*(Police Information Department, Liaoning Police College, Dalian, China)

ABSTRACT
With the gradual improvement of the informationization level in public security area, the concept "Information led policing" has been formed, many information systems have been built and vast amounts of business data have been accumulated down, But these systems and data are isolated and becoming the isolated information islands. This thesis proposes an architecture of information analysis system on big data platform, then discuss the question of data integration, finally proposes the correlation method for public security information: direct association and indirect association.

***Keywords* -** big data, correlation method, data integration, architecture, information analysis system

I. INTRODUCTION

With the gradual improvement of the informationization level in public security area, the concept "Information led policing" has been formed, whether social security administration, criminal investigation, anti-terrorism and stability maintenance, prohibition of gambling and opium, homeland security, or police information analysis, command decision making, the information is needed. What is the police information? Is the information the internal secret intelligence that Informer collected? Or is the "Secret" intelligence collected by special techniques? Now, the information we studied, is not the intelligence in the narrow sense, but is the information in a broad sense, the information refers to the original information obtained through public channels without analysis and evaluation, after analysis and assessment, it can be provided to leadership, project director and the front-line police to make decision.

With the popularization of information technology in public security area and the development of "Golden Shield Project", Chinese public security organs have accumulated a large amount of public data， including resident population information, transient population information, hotel guest information, internet bar guest information, criminal information, vehicle information, driver information, drug-related personnel information, fugitives information, and so on. These data have sources extensive, complex structure and the huge amount of data, only the police in the business department can use them. This information is accumulated business data by public security organs to carry out business work over the years, many isolated information islands are made for lack of scientific and rational planning, and for need of political performance, the public security system were built repeatedly. Of course, the public security information system have many problems, such as, the data were inputted repeatedly, and can’t be shared by different departments, its using rate is very low, and combat efficiency is poor. With the deepening of the "Golden Shield Project", on the basis of the previous business systems, the police information judging systems were built in every place. The unified mathematical model was established for multiple business information systems, the sharing of the public security information was realized, the isolated information islands were removed, automated or semi-automated multi-source information analysis were carried out, greatly improved the efficiency of police information analytical work, and improved the timeliness and accuracy of information judgments.

II. STATUS OF ANALYSIS ON POLICE SECURITY INFORMATION

With the development of economic and the change of the international situation, various cases have the trend of happened frequently, such as violent terrorist cases, the organized crime of underworld, vagabond to commit crimes, high-tech crime, these cases are increasing. This brings new challenges to the investigation of cases, how to expand the scope of information collected from the breadth, how to mine the association of information from the depths, how to establish a new operation mode of information-led policing, how to improve the ability of precision management, precise command, precision strike and precise prevention, these have become the direction of research and development for the countries in the world.

Now there are several issues of public security information analysis：

First, we need to expand the scope of the information-gathering. Because of the limit of information technology and the depth of people's understanding to information, the sources of traditional information analysis system is relatively narrow, the breadth of analysis is small. The traditional information analysis system usually use relational data warehouse technology to integrate data from multiple business systems, it only partially solves the problems of integration, correlation with data, data mining, and does not really solve the problems of massive heterogeneous data storage and parallel computing.

Secondly, we need to mine the variety of data in more depth. In the building and using process of traditional information system, in order to collect all kinds of data as possible, management department issued a variety of indicators to grassroots units. Grass-roots units have input a large number of duplicate data, even wrong data. This increases the burden of the information analysis system, and also reduces the accuracy of the information analysis. In the police information analytical work, we urgently need to mine a variety of data in more depth.

lastly, we need to analyse all kind of data in more breadth, the analysis breadth of traditional information systems on a variety of information is not satisfactory, and urgently need to strengthen the work in this area. In this regard the United States have had success stories, that we could learn from it. In 2006, the United States mapped the criminal records within the past 20 years and traffic accident records onto a map,

The results showed the area of a high incidence of traffic accidents are also lots of high crime area, the results showed the road of a high incidence of traffic accidents is also a high incidence of crime, after the US National Highway Traffic Safety Administration and the state judiciary united to enforce the law, traffic accidents and crime rates both declined.

III. The Architecture of Information Analysis System on Big Data Platform

The information analysis system on big data platform is a large and complex system, involved in hardware, network, operating systems, relational databases, cloud computing, big data and other aspects of knowledge and technology. In order to put them together to form an efficient, scientific system with practical application value. We need to elaborate design, and use a layered thought to simplify the system. every layer has independent function, the lower module provides service for the upper module, the upper module calls the function of the lower module, the two layers interact through interfaces, the information analysis system on big data platform are divided into the resource layer, the integration layer, the service and support layer, the interaction layer.

The resource layer provide the basic resources for the whole system, including hardware, network, database, to provide software, storage, calculation and data resources. The data resource is very important, it is the "ingredient" of data processing and data analysis of the system to make a meal of colorful, aromatic and appetizing "feast", we can't do it without it. Data resources, including base data and business data of public security, the public data of government agencies and community groups, the non-classified commercial data accumulated in the course of commercial business operations, the news reported by public media, the data of regional flea market site, and so on.

The integration layer is responsible for the integration of various data into the data warehouse. Information analysis requires to integrate various data sources, however, these data processing systems are different, the data structure is very complicated, the thickness and size to describe things are different. If you want to analyse the data, you must integrate them into data warehouse, you can't read them directly on the original business system. The task of the integration layer is to integrate all types of data periodically into the data warehouse, to store the data in the form of persons, things, places, organizations and other form, and to provide the base data for analysis and mining in the upper layer.

The service and support layer is the center of the information analysis system, it use the technology of big data, data mining, machine learning, provide the service and support for the upper layer, including the storage service using the Hadoop distributed file system(HDFS) to store structure and unstructured data, calculating service using MapReduce model to carry out distributed parallel computing, and data mining services such as the data sorting, association, classification, statistics.

The interaction layer is responsible for interaction between the operator and computer. According to the actual needs of public security, it provides the input and output for sorting, association, classification, statistics.

IV. Data Integration

The information analysis system involves various public security data provided by the basic department and business department, one class is identity information, including the resident population information, the temporary population information, the criminal record information, Falun Gong practitioners information, drug addicts information, Traffic illegal record information, driver information, the entry-exit personnel information, network virtual identity information, vehicle information. The other class is the track information, including the railway passenger information, civil aviation passenger information, shipping passenger information, Internet bar customer information, hotel customer information, passenger information, vehicle trajectory information. These information are important sources of police information, and a single source of information is often isolated information island, they may have the value of information only after integrated, and the process of information association is called data integration. There are two ways to integrate data: The first way is that an interface is provided in the business system for information analysis, information analysis system query and analyse the data directly in the original business system. The advantage of this way is no need to convert the original data, does not require additional storage space and computing power, but the impact to the original business system is very large, and even lead to interruption of service. The second way is to extract the data from the original business system, convert it to the format of the information analysis system, in this way, there is little effect to the original business system, but we must think carefully about the conversion of the data format, the thickness of the granularity of data description and other technical details.

The data sources of public security information analysis system are mainly structured data in the business database at present, we can convert the data from RDBMS to HDFS, HBase or Hive in the Hadoop, of course, We can develop the tools for data conversion according to special needs. This conversion can achieve specific functions, but developers need to master the underlying technology. In the future, the data may come from semi-structure or unstructured data of web pages, forums and posted bar, instant communication tools, express and logistics, e-commerce, we can use Flume or specific tools to convert it into data warehouse.

Data integration strategy are the following two aspects:

1. Batch additional way to integrate

The data structure of original business system is complex, with development of business, the amount of data is also increasing, for example, hotel customer information, passenger information, vehicle information, driver information. During integrating the data, we can use a batch integration strategy according to the conditions, the data of the hotel customer information can be integrated in the cycle of a month, two weeks or a week. The Sqoop conversion sentence is shown in the following code:

```
sqoop import --append --connect jdbc:mysql://192.168.1.100:3306/zhusudb --username root --password 123456 --query "select ID, Name, Gender,Room_ID, days, start_date, end_date ， reg_date from tab_zhusu where reg_date <= $s_date and reg_date >= $e_date" --target-dir /user/hive/warehouse/zhusudb --fields-terminated-by ",";
```

The data is converted from the database zhusudb to Hive table zhusudb, the variables $s_date and $e_date are the time when the customer check in and check out, we can append the data to the Hive table.

2. to grasp the proper granularity of data integration

In the original business system, detailed business information is recorded for the needs of business management, but these information is not always necessary for information analysis systems. For example: the integration of internet bar customer information to information analysis system, you should consider the granularity of information, internet bar management system records the customer ID, name, start-time, end-time, machine number, IP and other summary information, but the monitoring server also records details, such as the URL where the user is browsing, the game information, information of communication tools. When integrating the information, the granularity of information should be considered carefully, if we only query the time when someone is online, we should integrate the data in a coarse particle size. And if we want to know the details of a person in the internet bar, Fine-grained integration methods can be used, some details are also integrated into the information management system.

When converting the data with Sqoop, you can use --columns, --where options to limit the data to be converted, then control the granularity of information. As shown in the following example.

```
sqoop import --append --connect jdbc:mysql://192.168.1.200:3306/netinfor --username root --P  --table information --columns "Id, name, machine_id, start_date, end_date, IP" --where "start_date >=  $s_date  and end_date>= $e_date" --hive-import --hive-table netinfor --fields-terminated-by ",";
```

V. THE METHOD OF INFORMATION ANALYSIS

After the integration of information in the business systems, we can carry out the analysis of data mining. The most common method of data mining is association analysis. The isolated information in data warehouse can be associated by their common characteristics, then find some valuable clues, to associate the similar cases, to get the suspect’s travel path, to narrow down the field of inquiry, to provide the direction of investigation, until cracked the case and arrested the suspect. Type associations generally include direct association, indirect association, text association.

1. Direct Association

Direct association can associate the isolated information by the common attributes, such as ID number, phone number, bank account number, address, etc. for example, during investigating the case, we may get some information of suspect, we can query the suspect's details in the resident population database, such as name, ID number, birthday, home address, then query the information in other databases by ID number, in order to achieve a direct association.

1) to Query Criminal Record

After querying the suspect's basic information, query multiple data tables with his basic information, the condition of connections is the ID number. Information tables involved include the criminal table, fugitives table, drug addicts table, traffic offense table, and so on. to find out whether they have a criminal record, to provide support for improving query results and expanding the scope of information. The correlation method in Hive data warehouse is very simple, it can be realized by sub-queries and join queries.

2) to Query Travel Path

If the suspects commit the series crime all over the country, we can query their information by the temporary staff table, hotel customers table, internet bar customers table, bath accommodation customers table. After the customers information of the rail , aviation, shipping, road have been integrated, we can get suspects travel path by ID number, if the suspects have a vehicle, we can query the plate number in the vehicle table and driver table, then query the travel path of the vehicle by video monitoring system in every city.

3) to Associate the Phone Number

In modern society, the telephone is the main communication tools between people. When the telecommunications company registered customer information, they will record the customer's name, home address, phone number, and so on; In many places the information have been reserved, including cell phone numbers, landline numbers, fax numbers, etc. we can query the customer's name, home address, ID number, detailed call list. These information can form a network of relationships, we can find other relationship people, through relative network, we can find other related suspects.

4) Correlation Analysis of Belongings

Suspect's belongings can expand the scope of information, we can query the suspect's belongings in lost items database, figure out the relationship between the belongings and a case. For example, the suspect is driving a car, we can query the car by the car's large frame number and the engine number in the lost vehicle database, if the car has the relationship with a case, we can consider that the suspect has a certain relationship with this case.

2. Indirect Association

On the surface, there is no direct relation between an information and another information, but they may have indirect relation through some hidden clues, if the hidden clues can be found, the scope of information can be expanded, thereby the misrelated information can be associated and provide valuable information for the investigation of the cases, the method of indirect association mainly includes riding the same travel tools in travel, living in the same hotel in the same trip. In the same family in household registration system, the related persons in the entry and exit management system, these clues can provide great help for indirect association between misrelated information.

For example, During querying the same travel tools in data warehouse, the condition of query is that the starting date, place of departure, destination are identical, and the seats are adjacent, the booking numbers are connected. During querying living in the same hotel, the condition of query is living in the same room at the same date, or living in adjacent rooms and registration at the same time, the settlement time are same and living rooms are adjacent.

According to the above analysis, after integration of case data, a data warehouse is formed. In order to find the satisfying information from the data warehouse, the most commonly query methods are subqueries and join queries. We can find more cases with same characteristic from different databases by join queries, the result of subqueries are the conditions of other query. Through subqueries and join queries, we can associate a case with other cases, and also mine the cases.

VI. CONCLUSION

In this thesis we discuss the question of the isolated information islands, and present an architecture of information analysis system on big data platform, then discuss the question of data integration, finally propose the correlation method for public security information: direct association and indirect association.

Acknowledgements

This work was partly supported by general scientific research project of education department of Liaoning Province of China (No. L2013490).

References

[1] C. Dobre, F. Xhafa. Intelligent services for Big Data science. Future Generation Computer Systems 37 (2014) 267–281

[2] Nanning Zheng, Jun Zhang, Chenghong Wang. Special issue on big data research in China. Knowledge and Information Systems. 2014,41(2):247-249.

[3] Gencaga, D.;Malakar, N. K.;Lary, D. J. Survey on the Estimation of Mutual Information Methods as a Measure of Dependency Versus Correlation Analysis. AIP Conference Proceedings. 2014, 1636 (1):80-87.

[4] Yi Tong Liu, Li Li. A Method of Building Correlation Relationships to Thesauri Based on Improved Mutual Information. Applied Mechanics and Materials. 2014,431:7-11.

[5] Hock Chuan Chan, Varsha Guness, Hee-Woong Kim. A method for identifying journals in a discipline: An application to information systems. Information and Management. 2015,52(2): 239-246.

[6] Zhu, T., Xiong, P., Li, G., Zhou, W. Correlated Differential Privacy: Hiding Information in Non-IID Data Set. Information Forensics and Security. 2014,10(2):229-242.