SHS Web of Conferences **7**, 02006 (2014)
DOI: 10.1051/shsconf/20140702006

Research on the Extraction Technology of the Mass Data in Citizens' Information Infringement Cases

Wu Chunsheng[1.2], Zhu Xiuyun[1.3]

[1]Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, 100088 Beijing, China
[2]Computer Network Information Center, Chinese Academy of Sciences, 100190 Beijing, China
[3]The Criminal Investigation Department of Beijing Public Security Bureau, 100054 Beijing, China

Abstract. Objective The aim of this study is to explore the analysis methods of data from citizens' personal information infringement cases. **Methods** We distinguish various types of case data according to inspection methods, and proposes three kinds of inspection methods including the methods of data conversion extraction, inspection of file size property and forensics tools. **Result** Extraction technologies can realize mass data inspection in different degrees. **Conclusion** The inspection methods is effective and need to develop software further.

Keywords. extraction technology, citizens' information, mass data

1 Foreword

At present, the puzzle is how to take the evidence of huge amounts of data from a large number of data files in the field of electronic evidence identification in our country. Especially in recent years, the use of database makes it easy to gain several millions of data information[1][2]. Through the computer network, information is collected and spread more conveniently. The actual demand of the information age results in occurrence of the person who collects and disseminates information. This will create conditions for the emergence of various kinds of illegal behavior. As a result, the cases of infringement of citizens' personal information rapidly increase in recent years. It is estimated about 40 cases in Beijing in 2010 and more than 140 in 2012. The data excludes a large number of minor events that do not meet the case standard. In detection process of such cases, the important evidence is the huge amounts of data information in files[3] and a large number of files require analysis. At present, the data is calculated only one by one by manual. This method is not efficiency and often can't finish. As a result, it becomes the priority of electronic evidence inspection to improve information extract and inspection ability from huge amounts of data in all kinds of file.

2 The realized needs and problems

In 2009, in PRC criminal law amendment (7), the crime of selling or illegally providing citizens' personal information and the crime of illegal gain citizens' personal information were added. These have provided a legal basis for all kinds of cases of infringement citizens' personal information. it specify "violating the provisions of the state, let gained citizens' personal information in performing

duties or providing services for units, sell or illegally provide to others, if the circumstances are serious, shall be sentenced to fixed-term imprisonment of not more than three years or criminal detention, and concurrently or independently be sentenced to a fine. Who illegally obtain the above information by Stealing or any other means, if the circumstances are serious, should be punished according to the above mentioned measures."

The criminal law does not provide the concrete standard for serious cases and the relevant judicial interpretations. The definition of serious case is mainly determined by the judicial authority. In face the term of serious case is referred to large number, many times and great danger [4]. In these conditions the "large quantities" is mainly on the basis of the survey report issued by judicial authentication institutions. At present Citizens information is stored mainly in the form of electronic information. It put forward the technical requirements for judicial authentication institutions of electronic evidence inspection personnel. Only relatively accurate extraction to the amount of information in the storage medium, they can provide reliable evidence for the judicial department to eventually determine the nature of the case [5].

The traditional forensics work of such cases are as follows: the necessary process for storage medium; using forensics tools to search and locate the related information file; opening one by one and examining relevant documents and data recording. But electronic information data is stored in a variety of forms, often distributed in the tree directory structure and it often have multiple forms including great amount of data in the file, as shown in figure 1.

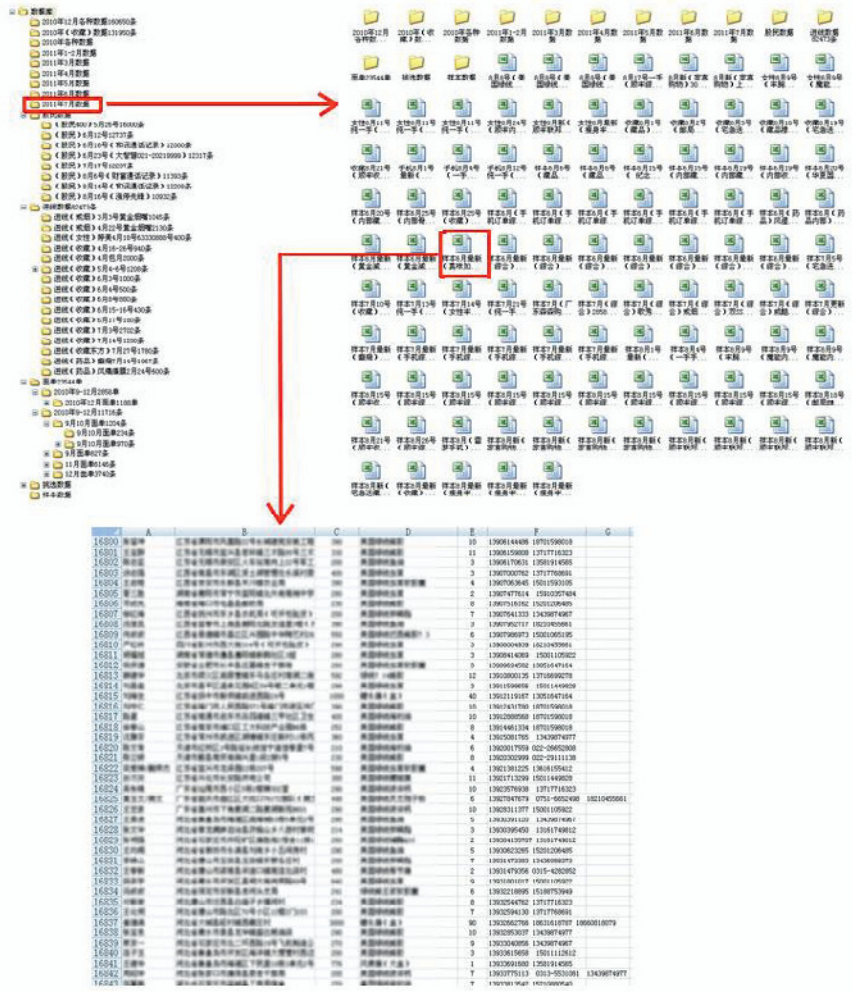

Figure1. Example of citizens' information infringement cases

This makes for involving large amount of data involved to open the confirmation and calculation one by one. There is no doubt it is a huge waste for the limited manpower. When it comes to huge amounts of data, it often appear the embarrassment that founding the evidence file but unable to statistics the number of effective and can't finish in such energy intensive work. So such cases will not be accepted or not reflected in the report in some appraisal institution. This leads to the contradiction between the real demand and the actual work.

3 Data type analysis

From a large number of case analysis, The electronic evidence involved in the infringement of information of citizens' personal cases are electronic files which store all kinds of personal information electronic documents, in addition to operating logs, application software such as attached file.

Storing files include excel, word, TXT, access, database files, database backup files, pictures and other data types. The excel and TXT document is the most common among them, it is sketched as follows.

TXT: TXT is pure text file and the format is simple and transparent, excluding structure information and encryption. It can be read with basic text editing tools. Huge amounts of data are usually stored as rows. There is a independent complete record and fixed data items in each row. There is unity separator between each data item. As shown in figure 2:

```
xiao01ong1ge2@163.com,fr8vfl6um,2295656690
r.chi@163.com,tnbok26welt80gbh,1837160521
baggiohhhhh@163.com,agd00w7n,1782758774
769704395@qq.com,7vhbyy1a0,2394030270
shuang29360@163.com,9n98t8emec1s,2172144112
364283859@qq.com,1djye36dzoe2y,1656255607
tincky2006@126.com,ufb427rgnkpyvj,1824901997
uqvdxdx@163.com,0qqq4ykp,1871915322
```

Figure2. Example of TXT file

Excel file: Microsoft excel is one of the components of Microsoft office, used for data management. Excel data are stored according to rows and columns in a table. There is a independent and complete record in each row, and the properties of each column are the same.

The various types of documents from different views can be divided as follows:

(1) The date that can or cannot be directly viewed

The data that cannot be viewed directly: all kinds of database files and backup files, which can be seen by importing the attribution database.

The data that can be viewed directly: other kinds of files, which can be seen by the method of directly opening.

(2) Structured data and unstructured data

Structured data refers to which have obvious logic division in each column, all data follow the same storage rules.

Structured data: excel, access, all kinds of database files and backup files.

Unstructured data: word, TXT, pictures and etc.

TXT or word files can be handled as a structured data many times. In addition, WPS files and ET files in the WPS word processing software can be handed as the same as word and excel.

(3) Readable data and unreadable data

The former refers to which can be directly identified by computer software through general data interface.

Readable data: excel, word, TXT and access.

Unreadable data: all kinds of database files, backup files and pictures.

The above classification provides basis for the further data extraction.

4 Data extraction methods

4.1 The method of data conversion extraction

There are often a variety of types of files in a case. If different types of data can be converted to the same type of data, it can reduce the statistical difficulty and workload. There are lots of data conversion tool software on the market at present, which can be used for almost all the mutual conversion of structured data. However the common data file has certain storage limits and read bottleneck, The more effective method is to use the data conversion tool which can batch import data files to general database, such as Oracle, SQL Server[6][7]. Then the database management system (DBMS) tools can be used for the needed data operation, to achieve the objective of data analysis and statistics. Using the DTS (data transformation services) tool, you can operate any data transfer between two data files. The data be transferred to the database can be easily analyzed and searched.

Data transformation method also has many problems, such as the need to install the corresponding database software in the inspection computer, inspection personnel must have enough database knowledge and operation skills, needs to understand different data file structure. This method is still facing a huge workload problem.

There are many data conversion tool which can realize the database connection, on the market, but the key point is data format conversion and data bridge joint of different data source. There is not an architecture and utility in electronic forensics field to extract and statistically analysis a mass of multiple source date, such as: automatically import multiple data files of the same structure at the same time, and the structured conversion of the text data in word and TXT file.

4.2 The method of using file size property inspection

The size of the file storage space is actually in proportion to the amount of data storage, so you can judge roughly the number of data from this attribute of file size. The relationship between file size and storage space based on test, are shown in table 1:

Table 1. File size attribute list

	Whether useful	Blank size (K)	Store size per 10 thousand characters (K)
txt	yes	0	about 19.5
Word2003	yes	23.5	about 26
Word2007	yes	0	about 11.5
Excel2003	yes	13.5	about 36.4
Excel2007	yes	9.6	about 15.2
picture	no		
database	no		

Notice: the word documents belong to the accumulative storage mode.

If word was saved several time, it would occupy much redundant space. If there are other fonts and format in paper in addition to the five words font, it will occupy more storage space. There is same situation in excel documents.

In most cases the way often used to generate data file is automatically batch conversion, the default format and font is seldomly changed. Then you can use the aforementioned mode to calculate date size. Although access file can be opened directly to intuitive viewing, there are many indexes data in it. Index data can occupy extra space, thus it isn't suitable for this method.

Through the following formula we can simply calculate the number of data. the average number of words in each row can be obtained by rough estimate. S is file size, S0 is blank file size, S10000 is Store size per 10 thousand characters, N is average word in each row and R is number of rows.

$$\frac{S - S_0}{S_{10000} \times N} = R$$

This method is simple, fast and easy to use. It is relatively effective for rough statistics of a large amount of data, but not for more accurate statistics. Due to the complexity format and no unified word number of each line, it is difficult to obtain the exact figures by this method. It is recommended for statistical order of magnitude of information.

4.3 The method of forensics tools inspection

Automated extraction method is the practical demand in actual inspection work. It is the most direct and fast way among three methods. This method is mainly using professional forensics tools software; realize the automatic identification of a large amount of data involved in inspected material; distinguish the relevant data of different files types; do the data statistics of sensitive information; collect the statistics number of each part.

This method is applicable to simple structured data. The data held the most part in the cases of infringement of citizens' personal information. Through the human-computer interaction special, forensics tools data read data and retrieve data according to its structural features. Usually such tools should have basic functions are as follows:

(a) Have a convenient human-computer interaction window, easy to operate.

(b) Data file read: provide users to choose a variety of data files in local folders, mainly include: structured data storage file, such as excel, access, TXT. It should effectively open and read the contents of the file. If the file contains multiple tables, it should list the name of table for user to choose. It can show the content preview of selected tables.

(c) Data statistics: according to the user need to select a field in a table for statistics, including total and total after duplicate rechecking. According to the user need to quickly select several tables in a file for automatically statistics. According to the user need to quickly select lots of tables in several files of a folder for automatically statistics.

(d) Data classification and summary: according to the date regular, such as the identification card number and phone number, it can automatically gather ID number, phone number and other information from a table, a file or multiple files, output qualified record number. It help users set the keywords, search a table, a file, multiple files using keywords and output the number of records accordant with keywords.

Using a special tool will greatly shorten the inspection period of such cases, improve the inspection accuracy. The biggest problem of this method currently is that there is not satisfactory software tools on the market, which need development.

5 Conclusion

To sum up, the one of key is the amount of data involved for the cases of infringement of citizens' personal information. There is efficiency and accuracy problem in current inspection methods. A set

of effective inspection methods and the applicable software tools need be designed and developed to provide strong support for such cases. The three kinds of inspection methods of paper still need to practice during the working process. We should increase the development speed of specialized tools at the same time, make it suitable for the requirements of practical application and provide technical support for such cases inspection work.

Acknowledgement

This paper is supported by the Opening Project of Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education. (GN: 2012KFKH03)

References

1. Xu Rongsheng, Wu Haiyan, Liu Baoxu. Computer Forensics Introduction[J]. Computer Engineering and Application, 2001(21):114.
2. Jiang Ping. Electronic Evidence[M]. Beijing: Tsinghua University Press, 2007.
3. Huang Jun, Wang Binjun. Probative Value Analysis of Document Evidence[J]. Netinfo Security, 2010(03):34-36.
4. Li Ziping, Zhou Jianda. The Brief of Plot Serious of Illegal Gain Citizens' Personal Information Crime [J]. Law Review, 2012(5):146-152.
5. Zhang Yunquan. The Electronic Evidence Forensics and Limits[D]. Beijing: China University of Political Science and Law, 2006(09):45.
6. Dong Mingrui, Shen Limin Zhao Guangjian. User-oriented Data Integration Model Research[J]. Micro Computer Information, 2010(7):26-28.
7. Su Zhuo. Integration of Heterogeneous Information Resources in the XML Model of Manufacturing Automation[J]. 2011(02):203-206.