Research on College Archives Resources Information in the Background of Big Data

Zhenzi Sun

School of Finance and Economics, Dalian Institution of Science and Technology, Dalian, Liaoning 116052, China

Abstract

The data generated and accumulated by colleges and universities in the process of campus informatization already have the features of big data, and it is of great significance to extract valuable information from massive data for archival preservation and utilization. The paper outlines the latest research results of big data management from the perspective of archive information resources. Starting from the features of big data, this paper analyzes the main sources of big data sources in universities, clarifies the relationship between data and files, and constructs a life cycle management model based on big data. The article from the concept of change, organization and coordination, strategy formulation, data collection, cleaning, polymerization and other aspects, good data management and archiving of important data. How to effectively manage big data has become an important issue for the development of big data. Colleges and universities, especially the archival departments, as functional agencies in charge of historical records and information resources, should seize the opportunity and perform their functions of digital file resource management.

Keywords: Big data, file management, informatization, university.

1. INTRODUCTION

First, the information management of college archives should face theoretical challenges. These theories include the concept of generating new sources, macroscopic identification theory, document continuum theory(Boyd and Crawford, 2012). The information management of university archives must give birth to the revolution of a new round of basic theory.

Second, the information management of college archives should face the challenge of information resource management (Ajegbomogun, 2004). At present, it has come to the era of big data. The geometric growth of college archives' information volume makes the existing information management methods difficult to cope with(Carlson et al., 2011). The big data era is not characterized by information but flooding. With the rapid development of information technology such as mobile Internet and cloud computing, the whole society has entered the era of big data. Universities are no exception.

In recent years, the data generated by colleges and universities in the process of personnel training, scientific research and social services have been growing exponentially (Morrisonand Secker, 2015). These data are huge in volume and diversified in variety. They give college management departments, especially the archives which are in charge of historical records and information resources(Scaramozzino et al., 2012). The sector poses challenges, and traditional digital archive management models and methods are no longer adaptable to big data management requirements. How to extract valuable information archive storage from complicated data and make sure that it can be read and utilized effectively a few years later will become the inevitable responsibility of archives(Kansagra et al., 2016; Ryan et al., 2016). In the face of the challenge of big data, colleges and universities archives departments should actively respond to the perspective of data resources, change the working methods and ideas, master the big data-related technologies, to meet the dawn of big data.

2.THEORETICAL PREPARATION

2011 World Economic Forum released a report that big data for the new value of wealth comparable to oil. On the concept of big data, different opinions in this paper we use McKinsey's definition: Big data refers to the data collection that can not be collected, stored, managed and analyzed by traditional database software tools in a certain period of time(Appleford et al., 2014). That is, the definition of Internet data center: to meet the many types of traffic, large capacity, high value data called big data. Although there is not a definitive definition of authority for the big data community, each attempts to define it by describing and generalizing its essential features. Representative of 3V features, 4V features, that is, the scale, fast and diverse, 3V features have now been recognized by the public. IDC in 3V basis, has added a new feature, that is, value. The added value is one of the 4V's most worthy of our attention(Kambatla et al., 2014). This is because the main purpose of the big data strategy developed by each country is also to realize the value of the data. In addition to the above 4V, there are scholars proposed additional 4V features, namely, verification, variability, authenticity and proximity characteristics.

From the characteristics of big data to consider, as the ultimate habitat of information resources, archives need to manage the digital file resources have now had the characteristics of big data (Osayande, 2009). Former researchers proposed that online all kinds of file big data information is being developed and utilized new resources to explore the technology of knowledge mining related to big data for the development of digital data mining on the Internet to provide insights and reference. De Mauro from the archives resources perspective, that in the era of big data digital archives already have a certain lbig data characteristics, the first large archives data resources and the rapid growth of large, the second is a variety of data resources, diverse structures(De Mauro et al., 2016). De Mauro depth analysis of the "big data" technology core essentials, technical characteristics and development trend, put forward the strategy and method of using "big data" technology to innovate college archives intelligence information service. In summary, because big data involves a lot of professional and technical knowledge on the one hand, it is a cross of many kinds of disciplines Integration; on the other hand, big numbers The concept is novel, the relevant technology is still evolving, some experts and scholars on big data is still in the preliminary study stage, introduced the system of theoretical and empirical study of some basic concepts of big data, features and related technologies, is still lacking.

3. UNIVERSITY BIG DATA SOURCES

Under the background of big data, college students will have a lot of data such as schooling, course selection, grades, borrowing books, Internet access, forums, Weibo and teachers' basic information, courseware and video, distance education courses, etc. Meanwhile, and book information will also produce a large amount of data. In addition to producing a large amount of data in the field of personnel training, a large amount of data will be produced in university management activities, scientific research and social services. These data are huge in volume, diversified in structure, widely distributed in origin, and obviously possess big data features. Taking Nanjing University of Aeronautics and Astronautics as an example, we analyze the main sources of data sources: 1)All kinds of websites. China Southern now has 1 school homepage site, 469 secondary websites including academic institutions, scientific research institutes, party groups, administrative departments and subordinate units, as well as 3 BBS social networks, only the paper plane BBS registration number is 317358 Person, post 3630469; 2)At all levels management system. Each business unit has its own management system, such as office automation (OA), faculty, students, finance, personnel, assets, welcome and many other business management systems, every day will produce a large amount of data; 3)Scientific research data. China Southern Airlines annual research funding up to several hundred million dollars, will produce massive experimental data in scientific research.

Faced with the explosive growth of data, the traditional management mode of digital archives lags behind and faces the dilemma of not being able to adapt to big data management. How to select the valuable data from the huge amount of data for centralized storage is a problem that needs to be solved by the file forming department and the filing department. For this reason, the author's university promptly adjusts his work and thinking mode, deepens the management concept of "big data" and "big file", and all the data resources reflecting the functional activities of the school are included in the scope of filing and handing over. In our opinion, the scope of filing should include: (1) Information on management, teaching, scientific research, academic and community activities issued by our campus network; (2) Documents and data generated by OA, faculty, students, finance and personnel management systems ; (3) E-mail generated by the school during official business activities; (4) Data generated by the school on BBS, blog, Weibo, instant messenger; Quality courses such as video, text, audio information; (6) Social network published on the school's important reports. How to automatically capture and archive the above data information resources will be the focus and difficulty of current and future work ideas.

4. BUILD BIG DATA-BASED PROCESSING MODEL

4.1 Model idea

Since the data can be converted to files, then how to convert, the middle of what process, a series of issues related to big data life-cycle management. Some analysts believe that the establishment of big data life cycle should include big data organizations to assess the status quo, the development of big data strategy, data acquisition, storage, processing, retrieval, analysis and presentation and several other stages. From the perspective of preserving historical records and maintaining the authenticity of archives, the author believes that big data management needs to continuously optimize strategies, methods, processes and tools and propose the following big data lifecycle processing model, as shown in Figure 1. Big data analytics requires the "purification" of complex, semi-structured, and unstructured data that is cumbersome, integrated, and correlated and clustered, making it a value-for-money information asset built on the same principle of source consistency set, permanently stored in digital archives or school data centers.

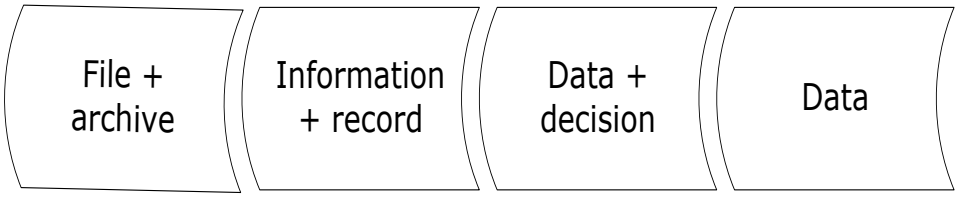

Figure 1. The relationship between data and files

4.2 Data processing

The word frequency statistics in BICOMB2 obtained 3912 non-repetitive keywords, given the frequency of each keyword and the percentage of the total frequency, combined with the total number of documents, the total number of keywords and other conditions, the threshold is set to 95 to get 20 A high-frequency keywords and word frequency. As shown in Table 1:

Table 1 Part of high frequency vocabulary frequency statistics

Keywords	Frequence	Keywords	Frequence
File manegement	357	File utilization	138
File information	236	File management mode	195
File digitization	168	Personnel files	218
Information construction	246	File business	109
Information technology	678	Countermeasures	201
Construction	114	Electronic filing	346
Informatization construction	245	File database	123
Manegement	568	Catalog database	357
Information manegement	212	Digital files	315
Digital Archives	251	File finishing	143

The co-occurrence matrix function of BICOMB2 can directly count the frequencies of high-frequency keywords co-occurring in the same document, and the word frequency threshold is determined 10. The statistics are 30× 30 matrices, and the Excel macro toolbox is combined with the cosine index:

$$\text{Cos}\,\textit{ine coefficient} = C_{ij} / \sqrt{C_i * C_j} \qquad (1)$$

Combining the cosine index formula and the Excel macro toolbox, the co-occurrence matrix is transformed into the similarity matrix. The results are shown in Table 2:

Table 2 Correlation analysis of hot words

	File manegement	File information	File digitization	Information construction
File manegement	1.00	0.19	0.09	0.32
File information	0.19	1.00	0.30	0.63
File digitization	0.09	0.30	1.00	0.03
Information construction	0.32	0.63	0.03	1.00
Information technology	0.34	0.04	0.70	0.00
Construction	0.43	0.26	0.12	0.05
Informatization construction	0.23	0.26	0.09	0.12
Manegement	0.10	0.13	0.16	0.15

	Information technology	Construction	Informatization construction	Manegement
File manegement	0.34	0.43	0.23	0.10
File information	0.04	0.26	0.26	0.13
File digitization	0.70	0.12	0.09	0.16
Information construction	0.00	0.05	0.12	0.15
Information technology	1.00	0.38	0.03	0.54
Construction	0.38	1.00	0.14	0.05
Informatization construction	0.03	0.14	1.00	0.73
Manegement	0.54	0.05	0.73	1.00

4.3 Cluster analysis

Clustering analysis is a multivariate statistical method, mainly for the classification of research samples or indicators. According to the different characteristics of a variable or a case of a batch of data, classification can be made according to the degree of density of the relationship. The result is shown in Figure 3:

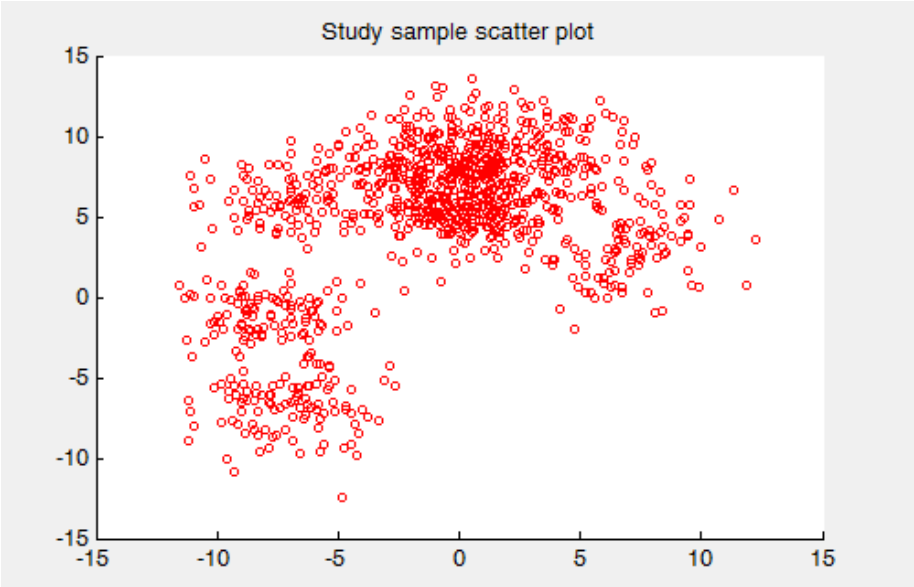

Figure 2. Sample plot graph

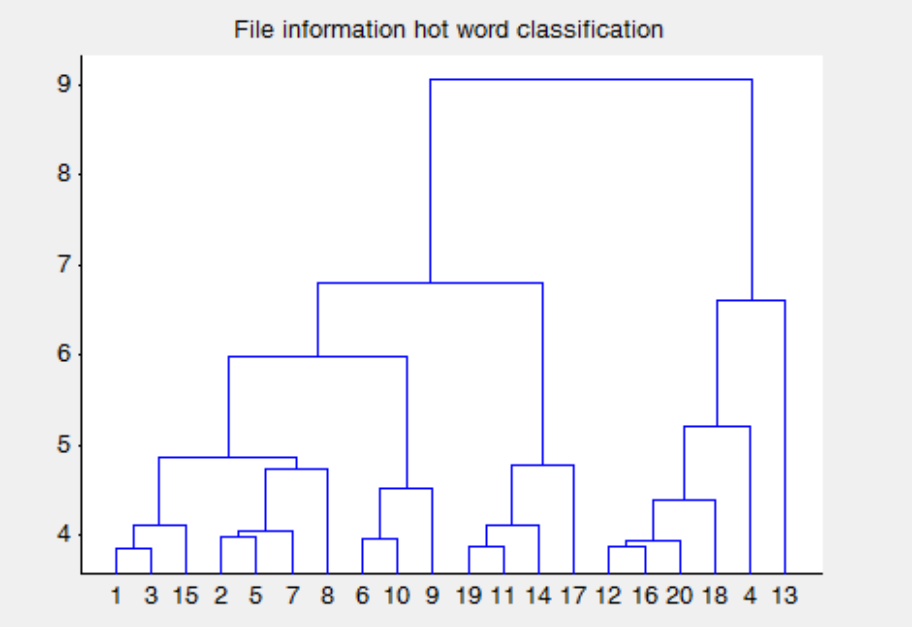

Figure 3. File information classification based on hot words

Clustering graph structure analysis. First, macroscopically observe the structure of the cluster tree. The leftmost label and number in the cluster tree represent the high-frequency keywords and their numbers. In this paper, the cohesion clustering algorithm is used to calculate the similarity between every two keywords. It is found that the first,the similarity of keywords 2, 6, 19, and 12 is the smallest among all the keyword phrases, so they are aggregated into a class first. As the distance between subsequent keywords widens, eventually all the words form a broad category. The structure of the tree diagram shows that all the keywords can be divided into five parts as a whole: Category A consists of words 1, 3 and 15, Category B consists of keywords 2, 5, 7 and 8, Category 6, 9, 10 word C, 19,11,14,17 word category D, by 12,16,20,18,4,13 word category E.

Multidimensional Scaling (MDS) is a kind of multivariate statistical method. It reflects the characteristics of many research samples and their similarities through the distribution of coordinates in low-dimensional space. Each sample is represented as a point in the spatial distribution. The distance between two points can be used to determine the degree of similarity between two points. The result is shownin Figure 4:

From the multidimensional scale analysis chart, combined with the result of clustering analysis, it is reasonable to divide all the high frequency keywords listed in the figure into three regions, and the key words in each region show the different structures of the archives informationization research.

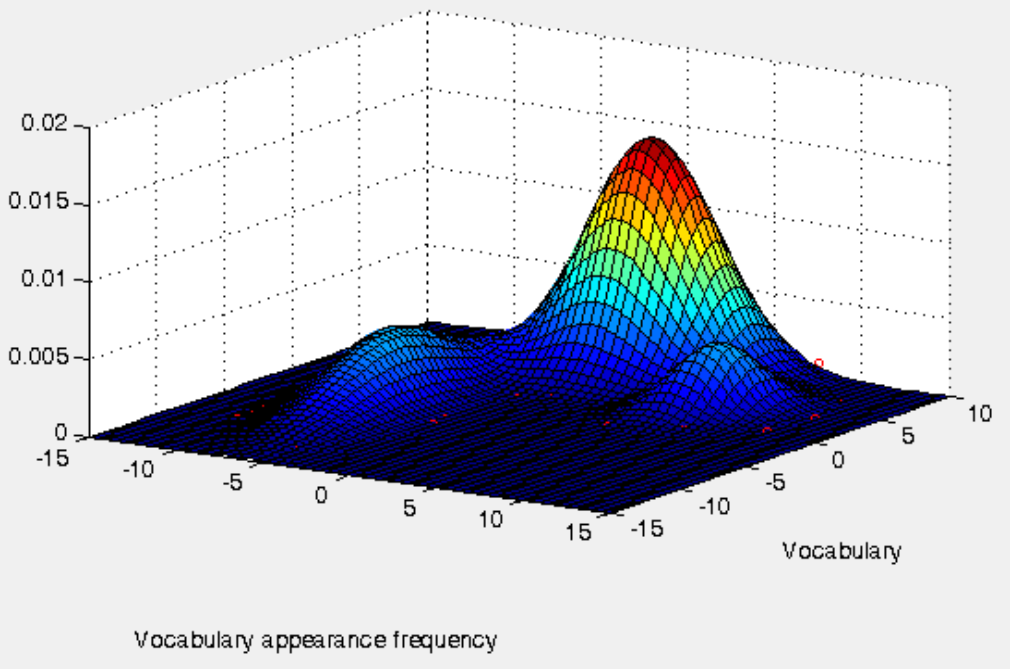

Figure 4. The scale model based on MDS

5. CONCLUSIONS

We have entered the era of big data irreversibly. It is incumbent on filing departments to take charge of historical data and information resources to collect data, protect data, discover value and maintain data validity. File information technology, information technology integrates computer file, database, internet, data mining, some of the information further comprising a push, grid like. This paper summarizes the impact of cloud computing on archives informationization, points out that we can rely on the party and government networks at all levels and the Internet to build a national cloud computing platform. Archives departments use this platform to provide digital archives management system, which provides new information management for archives direction. In this paper, big data management colleges and universities, for example, follows the data life cycle theory, trying to large data sources universities, flow, ownership, archiving and storage of exploration and research.

REFERENCES

AjegbomogunF.O. (2004). Users' assessment of library security: A Nigerian university case study, Library Management, 25(8/9), 386-390.

Appleford S., BottumJ.R., Thatcher J. B. (2014). Understanding the social web: towards defining an interdisciplinary research agenda for information systems. AcmSigmis Database, 45(1), 29-37.

Boyd D., Crawford K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

Carlson J., Fosmire M., Miller C.C., Nelson M. S. (2011). Determining data information literacy needs: A study of students and research faculty. portal: Libraries and the Academy, 11(2), 629-657.

De Mauro A., Greco M., Grimaldi M. (2016). A formal definition of Big Data based on its essential features, Library Review, 65(3), 122-135.

Kambatla K., Kollias G., Kumar V., Grama A. (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing, 74(7), 2561-2573.

KansagraA.P., John-Paul J.Y., Chatterjee A. R., Lenchik L., Chow D. S., Prater A. B., Smith S.E. (2016). Big data and the future of radiology informatics. Academic radiology, 23(1), 30-42.

MorrisonC., Secker J. (2015). Copyright literacy in the UK: a survey of librarians and other cultural heritage sector professionals, Research Journal of Botany, 8 (1), 1-14.

Osayande O. (2009). Security issues in academic libraries: The way out, Journal of Library and Information Science, 6(1), 10-17.

Ryan E.G., DrovandiC.C., McgreeJ.M.,MengersenK.L., Holmes C., Richardson S. (2016). Big data and design of experiments, Computational and Statistical Methods for Analysing Big Data with Applications, 6, 111-129.

ScaramozzinoJ.M., RamírezM.L., McGaugheyK.J. (2012). A study of faculty data curation behaviors and attitudes at a teaching-centered university. College & Research Libraries, 73(4), 349.