第18卷 第2期 2018年1月
1671—1815(2018)002-0279-06
科 学 技 术 与 工 程
Science Technology and Engineering
Vol. 18 No. 2 Jan. 2018

引用格式:陈海蕊. 面向电子档案大数据的可视化组织与分析[J]. 科学技术与工程, 2018, 18(2): 279—284
Chen Hairui. Visual organization and analysis of large data oriented electronic archives[J]. Science Technology and Engineering, 2018, 18(2): 279—284

自动化技术、计算机技术

面向电子档案大数据的可视化组织与分析

陈海蕊
(中原工学院软件学院,郑州 450007)

摘 要 电子档案发展至今已经积累到大数据状态,数据可视化成为解决电子档案管理问题的主要方法。根据以上描述,提出面向电子档案大数据的可视化组织与分析新方法。所提新方法将瓦片模型用于进行电子档案大数据可视化组织工作,借助瓦片图层的金字塔结构与矢量特征建立可视化组织结构。讨论了电子档案大数据编码方法与矢量瓦片可视化绘图流程,提高了可视化组织工作的准确性与效率。依次通过可视化图层捕获、非关系数据存储、数据流处理、已处理数据读取、分析结果显示以及图层更新,进行电子档案大数据可视化分析。实验结果显示,所提新方法的可视化效果优异,大数据读取效率较高。

关键词 电子档案大数据 瓦片模型 可视化 组织 分析
中图法分类号 TP311.12; **文献标志码** A

电子档案可用来存储一个人的学习、工作、健康与社会活动信息,也可用来存储物品的出厂信息、技术规格、出入库时间、库存量等,是思想道德建设与社会发展下的产物。电子档案中的数据主要是供人们审阅与查询的,科学技术的应用以及大数据时代的来临为电子档案的发展提供了新的契机,在这种新形势下,积极推进电子档案大数据的可视化建设显得异常重要。

大数据的产生一般来自于数据实时采集以及用户对数据通信效率与处理精度的高等级要求[1]。如果无法对大数据进行正确处理,电子档案将难以开展有效管理,电子档案的社会效用与服务效果也就难以发挥,大数据可视化将成为空谈。经以上分析可知,对电子档案大数据的可视化过程进行有效组织与准确分析是非常重要的。目前,一些科研工作单位已经开始着手进行这项研究,所提出的方法却大多不具备有效性[2]。其中,基于语义分析的可视化组织与分析方法最为实用,特别是在进行图书馆中进行书籍电子档案管理工作时,其可视化效果与大数据读取效率均要优秀于其他方法,但仍然未能满足用户日趋增长的可视化组织与分析需求。

在这种社会趋势下,一种面向电子档案大数据的可视化组织与分析的新方法成为各行各业的需求对象,这需要一些合理的科学技术为新方法进行理论与实践支持。时至今日,有效的研究成果仍然处于真空地位。

1 面向电子档案大数据的可视化组织与分析新方法

1.1 基于瓦片模型的电子档案大数据可视化组织方法

瓦片模型[3]是一种较为常用的多维数据组织模型,特别适合进行多样化的大数据可视化组织。瓦片模型采用空间定位与切割技术,将大数据内不同分辨率的数据点组织为若干具有相关联系性的可视化图层,图层形状类似于瓦片相互接触却又不属于同一水平区间的无缝连接结构,因此该图层又称为“瓦片图层”。

在瓦片模型中,与某一图层拥有相接点的其他图层中的长、宽、高、占用面积等特征,均可利用简单的数据表达式表示出来[4]。根据瓦片图层的特征大小对大数据进行可视化组织,建立一个金字塔结构,金字塔结构内上下图层的特征大小互为等比级数关系[5],以上便是基于瓦片模型的电子档案大数据可视化组织方法的基本原理。

2017年6月5日收到 2016年河南省科技厅科技攻关项目(162102210093)资助
作者简介:陈海蕊(1977—),女,汉族,河南濮阳人,博士研究生,讲师。研究方向:大数据处理与可视化。E-mail:6453@zut.edu.cn。

为电子档案大数据进行编码，编码结构可表示为

$$J = [L_N][R_C][C_C] \tag{1}$$

式(1)中，L_N 表示数据所处区域的层次[6]；R_C 表示可视化图层行编码；C_C 表示可视化图层列编码。R_C 、C_C 的编码次序均依照“北→南”“西→东”方向[7]，如图 1 所示，这样可以形成一个大数据树状检索关系，便于描述金字塔内数据特征。电子档案大数据编码的使用方法是：如果已知某一数据点的编码结构为[4][3][6]，则可以确定与该点临近数据点的编码结构为[4][2][5]、[4][2][6]、[4][2][7]、[3][3][5]等。处理电子档案大数据时，应从临近数据点开始依次向下进行可视化组织，防止漏项、串项。

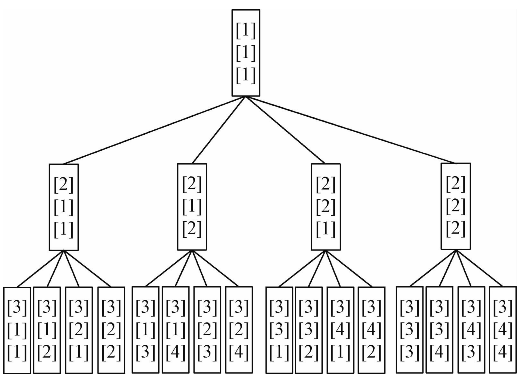

图 1　电子档案大数据编码结构

Fig. 1　Large data encoding structure of electronic archives

由图 1 可知，电子档案大数据编码结构与瓦片模型的瓦片模型的可视化图层结构是一致的，均采用金字塔结构存放大数据。越到结构底部，数据特征的描述就越详细。结构顶部是一个全局可视化场景，所存放的数据点一般是电子档案标题或数据身份加密序列[8]。同一层的编码数据拥有相同的粒度，可视化组织形式为视线平移，顶层数据拥有所有层中最大粒度值。

电子档案大数据中的数据点在经过编码后都需要进行可视化组织，也就是建立瓦片图层，具体先接入哪个数据点开展组织，所提新方法拥有一套标准流程：假设两个数据点 P_1 与 P_2 ，f_1 、f_2 分别表示这两个数据点的适应度，适应度的定义来自遗传算法[9]，其计算公式为

$$f = ml \tag{2}$$

式(2)中，f 表示适应度；m 表示数据点特征遗传力；l 表示数据点存活率。

如果 $f_2 > f_1$ ，则表示数据点 P_2 的特征更加适合遗传，与其他数据点的关联性更强[10]，应选择数据点 P_2 优先进行可视化组织。

但现实生活中存在数据点中电子档案信息不完善的情况，遇到这种情况时，以上的简单数值对比方法不成立，需要使用下式：

$$\alpha = \exp\left(\frac{f_1 - f_2}{T}\right) \tag{3}$$

$$\beta = R_d(0,1) \tag{4}$$

式中，α 表示优先规则 1；β 表示优先规则 2；T 表示电子档案平均遗漏数据量；$R_d(0,1)$ 表示随机反馈 0 ~ 1 中的任意值[11]。当 $\alpha < \beta$ 时，选择 P_1 与 P_2 中时间序列靠前的数据点优先组织；当 $\alpha > \beta$ 时，选择 P_1 与 P_2 中时间序列靠后的数据点优先组织；当 $\alpha = \beta$ 时，选择任意数据点。

根据瓦片模型的金字塔结构设计电子档案大数据可视化组织结构，如图 2 所示。每个图层都可表示一个可视化场景，电子档案大数据特征必须包含的内容是编码、元数据[12]、像素与矢量瓦片。

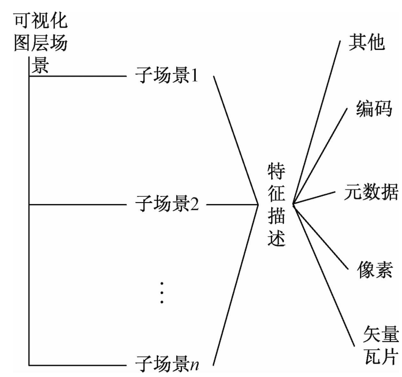

图 2　电子档案大数据可视化组织结构

Fig. 2　Electronic document large data visualization organization structure

矢量瓦片是瓦片模型中一种能够对电子档案管理服务器进行响应的参数，可以控制可视化组织效率，为可视化分析供应大量时间。矢量瓦片通过进行可视化绘图与服务器进行关联，其流程如图 3 所示，可视化绘图样本一般默认不存入所提新方法的数据库中，用户也可对默认项进行修改。

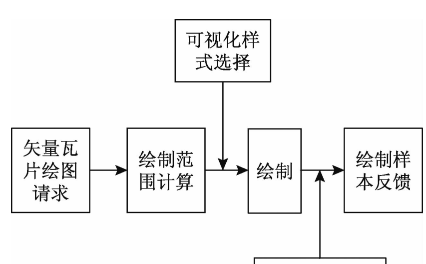

图 3　矢量瓦片可视化绘图流程

Fig. 3　Visual flow chart of vector tile

假设矢量瓦片可视化绘图任务为 $D = \{d_1, d_2, \cdots, d_n\}$,图层的行数、列数分别表示为 x、y[13],用 B 表示电子档案大数据空间范围,可视化组织区域为 V,可视化像素范围为 W,则可视化组织区域的分辨率 r 可用下式进行计算:

$$r = \frac{BW}{V} \tag{5}$$

根据分辨率计算出可视化图层中矢量瓦片可视化绘图范围,即 $[x_1、x_2][y_1, y_2]$[14],有

$$\begin{cases} x_1 = \left\lfloor \dfrac{Bx_{\min} + r\pi}{256r} \right\rfloor \\ x_2 = \left\lceil \dfrac{Bx_{\max} + r\pi}{256r} \right\rceil \\ y_1 = \left\lfloor \dfrac{By_{\min} + r\pi}{256r} \right\rfloor \\ y_2 = \left\lceil \dfrac{By_{\max} + r\pi}{256r} \right\rceil \end{cases} \tag{6}$$

式(6)中,算数符号 $\lfloor\ \rfloor$ 表示先进行向下运算,后进行向下运算,结果取整;$\lceil\ \rceil$ 表示先进行向上运算,后进行向上运算,结果取整[15]。$x_{\min}$、$x_{\max}$ 为可视化图层中最小、最大行数,$y_{\min}$、$y_{\max}$ 为可视化图层中最小、最大列数。

因此,基于瓦片模型的电子档案大数据可视化组织方法监控到的服务器时间代价可表示为

$$C = E + A + S \tag{7}$$

式(7)中,C 表示时间代价;E 表示矢量瓦片请求读取时间;A 表示矢量瓦片可视化绘图前的准备时间;S 表示矢量瓦片可视化绘图时间。

1.2　电子档案大数据可视化分析流程

图 4 是电子档案大数据可视化分析流程,依次进行可视化图层捕获、非关系数据存储、数据流处理、已处理数据读取、分析结果显示以及图层更新。

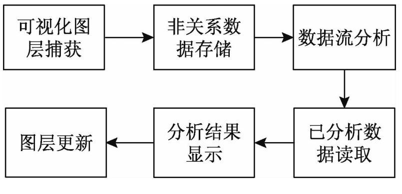

图 4　电子档案大数据可视化分析流程

Fig. 4　Visualization process of large data visualization of electronic files

可视化图层捕获的处理依据主要是电子档案大数据的类别、规模、时间以及成本因素,需要捕获到可视化图层中相对稳定、方便、安全、花费少的数据点[16]。由于电子档案大数据拥有复杂多样的特点,数据特征也是随着科技的发展而不断变更的,因此应选择捕获可视化图层中的非关系数据。

存储非关系数据时,应着重考虑数据库存储器的可扩展性能以及存储效率,可选择非关系型数据库作为支持,或者设计一个云计算模型帮助进行数据引流,以达成存储器对存储实时性与特效功能的需求。

数据流分析过程要求面向电子档案大数据的可视化组织与分析新方法以中心点为核心顺次向外处理数据流,利用最小生成树处理技术形成一个高效、准确、扩展性能强的可视化分析过程。整个可视化分析过程需要迭代元素、瓦片模型深度元素[17]和深度挖掘元素的支持,来全面、深层次地开展电子档案大数据流的可视化分析。

所提新方法的数据库不具备分析结果显示链接,需要对已分析数据进行读取,这样才能实现可视化分析结果的实时在线显示,显示后再通过图层更新继续捕获可视化图层中的数据流。

2　实验

2.1　实验数据库建立

图 5、图 6 均为实验数据库概念图。其中,图 5 为本文提出的面向电子档案大数据的可视化组织与分析新方法数据库;图 6 为基于语义分析的可视化组织与分析方法数据库,这是一种主要针对元数据结构进行可视化的方法,于元数据结构分析端处进行用户身份验证,优点是读取效率高、可视化效果较好。其与本文方法在处理过程中所使用的数据类型与序列是不一致的[18],表现为本文方法的数据库采用数据接收协议提取电子档案大数据特征,经分析后将重要特征到数据库中,于数据传递接口进行身份验证。因此实验中需要建立两种数据库。

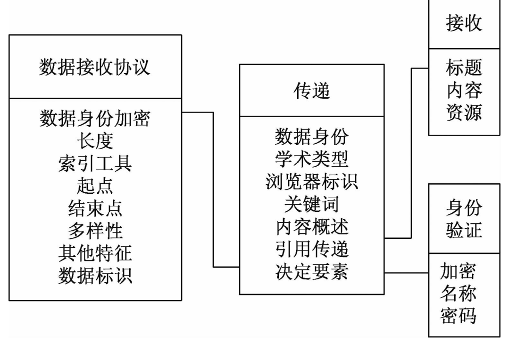

图 5　本文方法数据库概念图

Fig. 5　This method database concept map

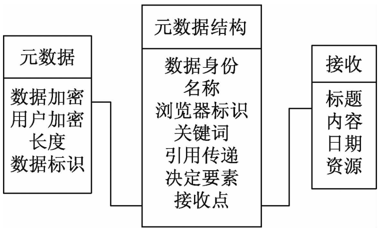

图6 基于语义分析的可视化组织与分析方法数据库概念图

Fig. 6 Visual organization and analysis method based on semantic analysis database concept map

2.2 可视化效果分析

实验所用的电子档案大数据于某省级的电子文档数据中心[19]直接下载，下载内容包括 2016 年上半年与下半年的某市图书馆书籍出入数据电子档案。上半年共涵盖 501 728 个数据点，下半年共涵盖 419 630 个数据点。

将下载好的数据依次导入本文方法与基于语义分析的可视化组织与分析方法各自的数据库内，建立上半年与下半年的电子档案大数据可视化图层，每份电子档案的各图层要素用表 1、表 2 描述。

表 1 本文方法图层要素

Table 1 Elements of the method layer

图层号	图层要素数量/个	
	上半年	下半年
1	36 175	33 250
2	21 664	20 756
3	10 246	98 442
4	96 785	83 174
5	81 771	72 563
6	46 725	40 206
7	16 573	11 357
8	12 469	6 654

表 2 基于语义分析的可视化组织与分析方法图层要素

Table 2 Layer elements of visual organization and analysis method based on semantic analysis

图层号	图层要素数量/个	
	上半年	下半年
1	68 944	60 128
2	36 445	32 419
3	21 473	20 438
4	4 237	3 751
5	1 028	996
6	532	441
7	48	40
8	11	7

从表 1、表 2 中能够明显看出本文方法与基于语义分析的可视化组织与分析方法在进行电子档案大数据可视化过程中的差别之处：本文方法建立的可视化图层中要素分布比较集中，交互能力更好一些，便于进行渲染，产生的可视化效果更好。基于语义分析的可视化组织与分析方法建立的可视化图层中要素分布特别分散，不利于获取良好的可视化效果。

为了获取更精确的可视化效果对比数据，将两种方法的可视化一级分析结果（指直接显示、未进行功能填入或编辑的分析结果）展示在图 7、图 8 中，能够看出，基于语义分析的可视化组织与分析方法只对书籍电子档案的满载量、现存量、增长率等直接项目进行了展示，其他项目需要手动添加进行分析。而本文方法还展示了书籍电子档案的完整性和类型分布情况，可视化界面清晰，书籍信息一目了然，可视化效果非常优异。

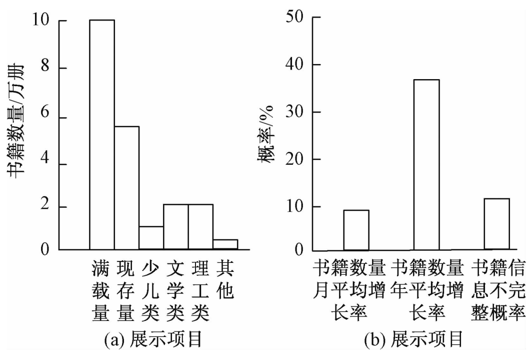

图7 本文方法可视化分析结果

Fig. 7 Visual analysis results of this method

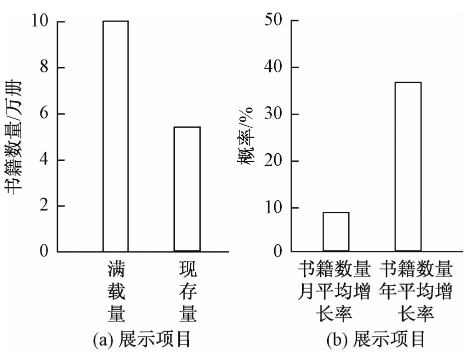

图8 基于语义分析的可视化组织与分析方法可视化分析结果

Fig. 8 Visual organization and analysis methods based on semantic analysis visualization analysis results

2.3 大数据读取效率分析

对比分析大数据读取效率的实验方法是：从每个可视化图层中各抽取一个数据包，令两种方法所抽取的数据包相同，便于进行对比分析。打开数据包进行读取，获取读取数据包的时间。采用两种不同模式读取上半年与下半年的数据包，设上半年数据包读取模式为“普通速度”，下半年数据包读取模式为“高速”。

一般来讲，高速模式的资源占用率比较大，对

电子档案管理系统的硬件设备要求较高[20],因此在本文实验中为数据点相对较少的下半年数据包选择了高速模式。表3、表4是大数据读取效率分析结果。

表3　本文方法大数据读取效率

Table 3　Efficiency of large data reading in this method

数据包所在图层号	大数据读取时间/ms	
	上半年	下半年
1	553.17	1.17
2	469.24	1.02
3	115.49	0.86
4	81.21	0.67
5	55.64	0.52
6	49.73	0.41
7	21.07	0.24
8	11.06	0.08

表4　基于语义分析的可视化组织与分析方法大数据读取效率

Table 4　Visual organization and analysis method based on semantic analysis, large data reading efficiency

数据包所在图层号	大数据读取时间/ms	
	上半年	下半年
1	753.42	5.12
2	466.18	3.46
3	294.37	2.49
4	135.95	1.08
5	94.27	0.67
6	75.48	0.64
7	35.49	0.31
8	31.75	0.23

由表3、表4可知,在本文方法与基于语义分析的可视化组织与分析方法中,高速模式与普通速度模式的大数据读取时间相差很大。本文方法的大数据读取时间普遍偏小,从最为常用的普通速度读取模式来看,本文方法的大数据读取效率绝对高于基于语义分析的可视化组织与分析方法的大数据读取效率,而且随着时间的推移与数据包内数据量的减少,本文方法大数据读取效率的增长速度较快,可持续减少大数据所消耗的资源占用率。

3　结论

电子档案是图书馆、企业、政府、学校的重要管理资料,面对现今大量、多样、价值不统一的电子档案大数据,如何开展高效管理是科学利用档案中有价值内容的关键所在。因此,提出面向电子档案大数据的可视化组织与分析新方法,通过可视化手段解决大数据中价值信息难以提取的问题,并在实验中建立了本文方法与基于语义分析的可视化组织与分析方法的数据库概念图,将从电子文档数据中心直接下载到的图书馆书籍出入数据电子档案输入数据库,进行两种方法的对比分析,指出本文方法的优势所在。

参　考　文　献

1　程　丹,李　伟,彭　永,等.面向道路工程Web3D档案管理的三维可视化方法.计算机工程与应用,2014;50(10):156—160
Cheng Dan, Li Wei, Peng Yong, *et al*. Three-dimensional visualization for road engineering Web3D archives management. Computer Engineering and Applications, 2014;50(10):156—160

2　罗贤春,余　波,姚　明.信息价值研究的热点领域可视化分析.现代情报,2015;35(11):120—126
Luo Xianchun, Yu Bo, Yao Ming. The measure analysis of information value. Modern Information, 2015;35(11):120—126

3　杨　晶,殷伟东,陈　平,等.版式文件技术在区域电子病历平台上的应用研究.中国医院管理,2014;34(8):39—41
Yang Jing, Yin Weidong, Chen Ping, *et al*. Research of the regional electronic medical record platform construction based on technology of E-document basis. Chinese Hospital Management, 2014; 34(8): 39—41

4　张作树,徐　伟,刘金岭,等.施工企业HSE信息管理系统的设计与运用.中国安全生产科学技术,2015;11(10):191—196
Zhang Zuoshu, Xu Wei, Liu Jinling, *et al*. Design and application of HSE information management system for construction enterprises. Journal of Safety Science and Technology, 2015;11(10):191—196

5　李春英,张巍巍.全球大数据与健康管理的研究热点聚类分析.中国医院管理,2016;36(10):63—65
Li Chunying, Zhang Weiwei. Clustering analysis of research hotspots of the big data and health management. Chinese Hospital Management, 2016;36(10):63—65

6　武昭宇,张月琴,阎　华.软件测试方法的研究与应用.太原理工大学学报,2016;47(3):379—383
Wu Zhaoyu, Zhang Yueqin, Yan Hua. Study and application of software test method. Journal of Taiyuan University of Technology, 2016; 47(3):379—383

7　孙欣欣.城市突发水涝灾害大数据分析技术研究.科技通报,2016;32(4):196—201
Sun Xinxin. Study on the big data technology of the city unexpected flood disaster. Bulletin of Science and Technology, 2016; 32(4): 196—201

8　孙艳秋,刘　钢.基于大数据分析的潜在高血压病预测研究.计算机仿真,2015;32(5):386—389
Sun Yanqiu, Liu Gang. Based on the data analysis of potential hypertension prediction research. Computer Simulation, 2015; 32(5): 386—389

9　秦殿启.整合与大数据理念下的个人知识组织.情报理论与实践,2014;37(2):19—22
Qin Dianqi. Integration and personal knowledge organization under the concept of big data. Intelligence Theory and Practice, 2014; 37(2): 19—22

10　Beghini L L, Pereira A, Espinha R, *et al*. An object-oriented framework for finite element analysis based on a compact topological data structure. Advances in Engineering Software, 2014;68(2):40—48

11 张 凯. 大数据网络入侵过程的痕迹数据监测方法研究. 科学技术与工程,2016;16(14):254—258
Zhang Kai. Big data network intrusion traces of process data monitoring method research. Science Technology and Engineering,2016;16(14):254—258
12 Trumpy E,Bertani R,Manzella A,*et al.* The web-oriented framework of the world geothermal production database:a business intelligence platform for wide data distribution and analysis. Renewable Energy,2015;74(5):379—389
13 宋 雨,焦 谱,李 刚. 大数据预处理中属性约简的特性保持分析. 计算机测量与控制,2015;23(12):191—194
Song Yu,Jiao Pu,Li Gang. Characteristics analysis of attribute reduction in big data preprocessing. Computer Measurement & Control,2015; 23(12):191—194
14 陈世敏. 大数据分析与高速数据更新. 计算机研究与发展,2015;52(2):333—342
Chen Shimin. Big data analysis and data velocity. Journal of Computer Research and Development,2015;52(2):333—342
15 周小娟. 一种轻量级大数据分析系统的实现. 电子设计工程,2016;24(8):40—43
Zhou Xiaojuan. Implementation of lightweight big data analysis system. Electronic Design Engineering,2016;24(8):40—43
16 Singh P,Negi P,Laezza F,*et al.* Multiscale analysis of neurite orientation and spatial organization in neuronal images. Neuroinformatics,2016;14(4):1—13
17 吴纯青,任沛阁,王小峰. 基于语义的网络大数据组织与搜索. 计算机学报,2015;38(1):1—17
Wu Chunqing,Ren Peige,Wang Xiaofeng. Survey on semantic-based organization and search technologies for network big data. Chinese Journal of Computers,2015;38(1):1—17
18 Wanner F,Jentner W,Schreck T,*et al.* Integrated visual analysis of patterns in time series and text data-Workflow and application to financial data analysis. Information Visualization,2016; 5(9):283—442
19 王学娟,杨 阳,闫桂莲. 浅谈大数据时代下承德高校电子档案管理存在的问题和对策. 承德医学院学报,2016;33(6):537—539
Wang Xuejuan,Yang Yang,Yan Guilian. Problems and countermeasures of Chengde university electronic archives management in the era of big data. Journal of Chengde Medical College,2016;33(6):537—539
20 徐 飞,徐绪堪,吴广印. 大数据环境下科技文献用户阅读行为知识组织模型研究. 情报学报,2015;34(9):912—921
Xu Fei,Xu Xukan,Wu Guangyin. Research on knowledge organization model of scientific literature based on users' reading behavior in the background of big data. Journal of the China Society for Scientific and Technical Information,2015;34(9):912—921

Visual Organization and Analysis of Large Data Oriented Electronic Archives

CHEN Hai-rui
(School of Software, Zhongyuan University of Technology,Zhengzhou 450007, China)

[**Abstract**] The development of electronic archives has accumulated to the status of large data, and data visualization has become the main method to solve the problem of electronic records management. According to the above description, a new method of visual organization and analysis for large data files is proposed. The proposed method will be used for tile model visualization of big data organization of electronic archives, with the pyramid structure and characteristics of the vector tile layer to establish electronic archives data visualization organization structure, discusses the electronic archives data encoding method and vector tile visualization drawing process, improve the efficiency and accuracy of visual organization. In order to do the visualization analysis of large data of electronic files, it includes visual layer capture, non relational data storage, data stream processing, processed data reading, analysis result display and layer updating. Experimental results show that the proposed method has excellent visualization effect and high data reading efficiency.

[**Key words**] electronic archives big data tile model visualization organization analysis