Multi-view Collaborative Federated Data Visualization Analysis Postprint

Authors: Shen Xuefeng, Ke Yongzhen, Yao Nan

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】To address the problems existing in the current knowledge discovery process for alliance data, we design a visual analysis system model for alliance data to achieve collection, mining, and visual analysis of historical information. 【Methods】We construct a visual analysis system model for alliance data, build a big data platform, and verify the usability of the model. 【Results】Experimental results show that the system can effectively perform visual analysis on massive historical data and support decision analysis. 【Limitations】The current visual analysis result views can be further enriched. 【Conclusion】This system can perform visual analysis on historical alliance data, providing scientific data support for decision-makers.

Full Text

Preamble

Visualization of Coalition Data Based on Multi-View Cooperation

Shen Xuefeng, Ke Yongzhen, Yao Nan
(School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China)

Abstract

[**Objective**] This paper proposes a visual analysis system model for coalition data to address existing challenges in knowledge discovery, enabling the collection, mining, and visual analysis of historical information. [**Methods**] We constructed a visual analysis system model for coalition data, built a big data platform, and validated the model' s usability. [**Results**] Experimental results demonstrate that the system effectively performs visual analysis on massive historical data and supports decision-making. [**Limitations**] The current visual analysis result views could be further enriched. [**Conclusion**] The system

can visually analyze historical data from coalitions and provide scientific data support for decision-makers.

Keywords: Coalition Data; Big Data; Visual Analysis; Borrowing Records
Classification Numbers: TP311; G350

Introduction

With the rapid development of information technology, multi-institution data sharing has become a common phenomenon. A major challenge facing data coalitions today is how to better organize and mine meaningful results from massive amounts of information while establishing an interactive data mining model. Visual analytics is a science and technology that assists users in analyzing and reasoning about large-scale complex datasets through interactive visual interfaces [**?**]. By employing visual analytics methods, we can address problems of information overload and ineffective communication, discover deep potential knowledge hidden in massive resources, reveal the deeper implications of results, and improve their comprehensibility and cognitive accessibility.

Library alliances represent a prime example of multi-institution data resource sharing. Exploring methods to mine library alliance data resources offers valuable insights for coalition data resource development. This paper uses the Tianjin Library Alliance big data as a case study, relying on visual analysis methods to conduct knowledge discovery and achieve decision analysis and policy formulation.

To integrate high-quality resources from Tianjin' s university libraries, the city began constructing the Tianjin University Joint Digital Library in 2002. This library alliance includes 26 libraries from 17 municipal institutions (excluding Tianjin University and Nankai University), while simultaneously enabling shared bibliographic data through gateway interlinking with the existing Unicorn systems of Tianjin University and Nankai University. The core mission of the library alliance is to establish a joint bibliographic sharing system for Chinese and foreign language books and periodicals, achieving bibliographic and technical sharing among member libraries, and completing automated management of library operations including acquisition, cataloging, circulation, periodical management, public inquiry, and interlibrary loan, thereby improving the automation management level of Tianjin' s university libraries.

After 15 years of development, the joint library has accumulated vast amounts of collection data and reader borrowing records. Applying data mining technology to quantitatively analyze reader borrowing history can reveal personalized reading needs, with mining results serving as a data reference for literature procurement decisions at each library. This can improve the quality of literature resource selection and collection utilization rates, making literature procurement more objective, scientific, and rational [**?**].

Existing research on university library book procurement has primarily focused

on analyzing data from individual libraries. Zhao Yingchun [?] employed grey relational analysis to evaluate the importance of various book categories in university libraries, comprehensively considering factors such as collection volume, borrowing volume, key discipline construction, and reader needs and evaluations to analyze the importance and correlation of each factor. However, this analysis only examined broad book categories, representing a certain limitation. Yin Jijun [?] analyzed and researched the application of neural networks for intelligent book procurement, designing an intelligent book procurement system model based on improved genetic neural networks according to the behavioral characteristics of book procurement. Li Yuan et al. [?] utilized fuzzy comprehensive evaluation to analyze borrowing data and establish a fuzzy comprehensive evaluation model for university library literature resource procurement, determining reasonable procurement amounts for various literature resource types.

In the big data context, some scholars [?]-[?] have explored book procurement models and applied data mining techniques to support university library book purchasing plans. Although Chi Chunjia et al. [?] discussed the feasibility of applying data mining in formulating book procurement plans, they did not provide specific examples. While Feng Na [?] provided examples, the data was based on questionnaires, resulting in strong subjectivity.

Domestic research on book procurement across multiple libraries remains relatively scarce, and even fewer studies employ visual analysis methods. Book classification information constitutes hierarchical data, and the visualization of hierarchical data has always been an important research topic in information visualization [?]. Related work primarily falls into two categories: node-link diagrams using explicit representation and space-filling methods using implicit representation [?]. Node-link diagrams represent parent-child relationships between nodes as connecting lines, which clearly display hierarchical relationships [?]. Space-filling methods use blocks with certain areas or volumes to represent individual nodes in the data, with treemaps and their variants [?]-[?] being representative examples. Compared with node-link diagrams, space-filling methods generally allocate most space to leaf node presentation, making it difficult to identify hierarchical or adjacent relationships among non-leaf nodes [?]. In practical application domains, data composition is becoming increasingly complex, with most data possessing not just single data characteristics but multiple characteristics simultaneously [?]. For such complex data, Chen et al. [?] proposed employing two or more visualization methods to address the inability of existing visualization and visual analysis methods designed for single data characteristics to meet analysis needs.

To observe, understand, and grasp results from different perspectives, we employ multiple views to achieve effective organization and expression of massive resources from different dimensions, designing and implementing a coalition data visual analysis system.

3. Coalition Data Visual Analysis System Model

Our coalition data visual analysis system model is based on the Hadoop platform, using HDFS as the massive data storage platform. The entire model includes five components: Hadoop infrastructure, data collection, data preprocessing, data analysis, and data visualization, as shown in Figure 1 [Figure 1: see original paper].

The modules are described as follows: (1) **Hadoop Infrastructure**: Provides operation interfaces for Hadoop distributed data (index library, Hive data warehouse, analysis library) and the MapReduce parallel computing framework; (2) **Data Collection**: Collects corresponding data according to specific requirements; (3) **Data Preprocessing**: Completes data deduplication, noise reduction, feature extraction, and related work to prepare data for visual analysis; (4) **Data Analysis**: Performs vectorization representation of text, association analysis, statistical analysis, and other analytical functions on preprocessed data; (5) **Data Visualization**: Conducts visualization based on D3 [**?**] visualization components.

This paper uses library alliance data as an example to validate the model' s feasibility, focusing primarily on the data preprocessing and visual analysis modules.

3.1 Data Preprocessing

The data preprocessing module needs to process borrowing data and collection data. Borrowing data contains five fields: ID, time, institution, barcode, and username. However, due to historical reasons, some libraries' data had the following issues during the merger into the joint library: `1907863|CJ495415|2014122615| 民航大学馆 | 张三`, where the second, third, and fourth items are out of order; `1907864|M1214789|2014122615| 商学院馆 |C00624610| 王五`, where the second item is redundant data.

To address these issues, each data record needs to be formatted into a unified format, with field positions swapped or removed for non-compliant data.

Each collection data record contains barcode and call number information. Since a book's position on the shelf may change, one barcode may correspond to different call numbers. For example, barcode `ZY8027501` corresponds to call numbers `D125/4`, `D125/1`, `D125/C`, `D125/A.L.X`, `D08/ELX(LS)`, `D751.664`. Extracting classification numbers from the call number set requires a branch-and-bound algorithm.

Input: Call number set S
Output: Classification number S_s

The algorithm traverses the set, taking the first character of elements ($0 \leq n \leq |S|$) to form set K. It then calculates the weight w_j of each element in K, selects the element k_j corresponding to $\max(w_j)$, and extracts from S the element set S_1 whose first character begins with k_j. For the second character of S_1, the

same step is repeated to select set S_2. This process continues sequentially for the i-th character of elements in set S to obtain set S_k, ending when the number of elements in set S_k becomes unique.

3.2 Data Visualization

For collection and borrowing data with hierarchical and multi-dimensional characteristics, we designed three views to display the hierarchical structure information and multi-dimensional attribute information of books.

(1) Collection Book Display View

For representing hierarchical relationships among collection books, we selected node-link diagrams as the view for collection book display. In tree structures, to represent quantity comparison relationships between different categories, we chose weighted trees as the presentation method. A weighted tree displays the weight of each node alongside the node itself, using node size to represent weight magnitude. Here, we selected the open-source weighted tree component Vizuly [**?**] as the display view for the hierarchical structure of collection books.

(2) Comparison View of Borrowed Books and Collection Books

To analyze the utilization rate of certain book categories, we define the book borrowing ratio i_r as shown in Equation (1):

$$i_r = \frac{\text{lent}}{\text{stock}}$$

where i represents a book category, stock represents the collection quantity of that category, and lent represents the borrowing quantity of that category. To represent quantity differences among different book categories, we designed a second view where each book category is expressed by an arc segment and two triangles conveying three dimensions of information: collection quantity, borrowing quantity, and borrowing ratio. We used D3' s arc generator as the basic graphic framework, with each arc containing information such as start angle, end angle, inner radius, and outer radius, completing the view using a polygon layout algorithm. The polygon layout algorithm is as follows:

Input: Arc sequence $\text{arc}_a = \langle a_0, a_1, \ldots, a_n \rangle$, data $D = \langle d_0, d_1, \ldots, d_n \rangle$, parameters λ, α, β
Output: Arc sequence $\text{arc}_b = \langle b_0, b_1, \ldots, b_n \rangle$, arc width w_i, triangle vertices V

The arc width calculation is shown in Equation (2):

$$w_i = \lambda \times \frac{d_i}{\sum_{j=0}^{n} d_j} \times (\beta - \alpha)$$

The horizontal and vertical coordinates of points on the arc are calculated using Equations (3) and (4):

$$x_i = r \times \sin(\theta_i), \quad \theta_i \in [1, 6]$$

$$y_i = r \times \cos(\theta_i), \quad \theta_i \in [1, 6]$$

The horizontal and vertical coordinate calculation formulas for points outside the arc are shown in Equations (5) and (6):

$$x_i = (r + w_i) \times \sin((\theta_s + \theta_e)/2), \quad \theta_i \in [1, 6]$$

$$y_i = (r + w_i) \times \cos((\theta_s + \theta_e)/2), \quad \theta_i \in [1, 6]$$

When $i = 1, 2, 4$, $\theta_i = \theta_s$; when $i = 3, 5, 6$, $\theta_i = \theta_e$.

Using the arc generator to produce arc sequence arc_a, for each arc a_j, we calculate its start angle $\theta_{j,s}$ and end angle $\theta_{j,e}$. Based on parameters λ, α, β and the start angle $\theta_{j,s}$ and end angle $\theta_{j,e}$, we derive the arc width $w_{j,i}$ and triangle vertices V. According to the size termination of the arc sequence, we output arc_b and $w_{j,i}$. In this paper, the polygon layout algorithm parameters are $\lambda = 100$ and $\alpha = 0$. Additionally, the algorithm requires other parameters including: outer radius r, arc start angle θ_s, and arc end angle θ_e.

(3) View of Books and Libraries

To understand the relationship between books borrowed by readers and libraries, we designed a third view. Since the university joint library comprises 26 libraries, to analyze how readers at each library borrow different categories of books and how different categories of books are borrowed at different libraries, we abstracted the relationship between book categories and libraries as a graph $G(V, E)$, where vertices V represent book categories and libraries, and edges E represent relationships between book categories and libraries. Since book categories are independent of each other, different libraries are independent of each other, and book categories and libraries are also independent, G is a bipartite graph. For bipartite graph visualization, we referenced components related to bipartite graph visualization by Pasha [**?**].

Books can be divided into 22 major categories according to classification numbers. This paper selects industrial technology books for case analysis, focusing on analyzing the borrowing ratio of this category and its borrowing patterns across different libraries.

4.1 Data Source

The data used in this paper comes from the Unicorn library automation management system of Tianjin University Digital Library, spanning from the system' s launch date to February 2015.

4.2 Visual Analysis

(1) Analysis of Collection Books

The hierarchical structure of books in the university joint library is shown in Figure 2 [Figure 2: see original paper]. From Figure 2(a), we can clearly observe the relative relationships among book categories in the collection: industrial technology books are the most numerous, followed by literature, economics, and language categories, while aerospace and astronomy/earth science books are relatively scarce. This phenomenon relates to the majors offered at the 17 institutions—for example, only Civil Aviation University of China offers aerospace programs, and none of the 17 institutions offer astronomy/earth science programs.

(2) Comparative Analysis of Borrowed Books and Collection Books

The comparative analysis of borrowed books and collection books is shown in Figure 3 [Figure 3: see original paper]. The inner ring represents the collection quantity for each category, orange triangles represent the borrowing quantity, and blue triangles represent the borrowing ratio i_r. From Figure 3(a), we can see that among the 22 major categories, literature has a borrowing ratio i_r of 47.2%, indicating strong reader demand for this category. To further display borrowing patterns of subcategories within the 22 major categories, we selected the industrial technology category for deeper analysis, with results shown in Figure 3(b).

From Figure 3(b), we can see that automation/computer science books have the largest collection quantity, while weapons industry books have the smallest. Automation/computer science books also have the highest borrowing quantity, while atomic energy technology books have the lowest. The three categories with the highest borrowing ratio i_r are light industry/handicraft (28.2%), automation/computer science (24.3%), and architectural science (21.4%). The lowest borrowing ratio is atomic energy technology (1.6%). The generally low borrowing ratios reflect low utilization rates of collection books, which aligns with the objective reality of reduced demand for paper books in the internet environment.

(3) Analysis of Books and Different Libraries

Figure 4 [Figure 4: see original paper] shows the relationship between different book categories and libraries. The university joint library comprises 26 libraries. This view allows analysis of borrowing patterns for the same category across different libraries and different categories within the same library.

Analysis of borrowing patterns for the same book category across different libraries is shown in Figure 5 [Figure 5: see original paper]. From Figure 3, we know that three subcategories of industrial technology have borrowing ratios i_r exceeding 20%: automation/computer technology, architectural science, and light industry/handicraft. From Figure 5(a), we can see that automa-

tion/computer technology books account for 49% of industrial technology books borrowed by readers, with the top three libraries being Tianjin Polytechnic University Library (26%), Tianjin Vocational Normal University Library (14%), and Urban Construction College Library (12%). From Figure 5(b), we can see that 58% of architectural science books are borrowed from Urban Construction College Library. From Figure 5(c), we can see that light industry/handicraft readers are concentrated at Tianjin Polytechnic University Library (45%) and University of Science and Technology Library (18%). Although readers from the Academy of Fine Arts Library account for only 8% of borrowing in this category, the Academy of Fine Arts has particularly high demand for these books. From Figure 5(d), we can see that readers from the Academy of Fine Arts Library account for 32% of light industry/handicraft books borrowed. This situation may relate to the majors offered at each institution or reader preferences, requiring further research. This data can serve not only as a basis for allocating procurement funds among different schools but also as a reference for allocating procurement funds across categories within each institution.

Further analysis of the three categories with lower borrowing rates is shown in Figure 6 [Figure 6: see original paper]. From Figure 3(b), we know there are three categories with borrowing ratios around 2%: atomic energy technology (1.6%), mining engineering (2.0%), and metallurgical industry (2.3%). Readers of these three categories are concentrated at Polytechnic University Library and Industrial University Library, but from Figure 6(a), we can see that readers from Medical University Library also account for 6% of borrowing in these categories. Although the reasons for this phenomenon require further research, this data reflects reader demand for these categories, enabling procurement personnel to make purchases according to reader needs.

The above analysis examines relationships between different book categories and different libraries. This view can also analyze different demands of readers at different libraries for different categories, as shown in Figure 7 [Figure 7: see original paper] and Figure 8 [Figure 8: see original paper]. We can see that readers across all libraries generally have high demand for automation/computer science books within the industrial technology category, especially at medical libraries. Procurement should focus on these reader needs.

Through multi-dimensional visual analysis of collection data and reader borrowing records from the alliance library, we discovered the problem of low book borrowing rates in the alliance library. Multi-view visual analysis methods can not only more clearly display the hierarchical structure of data but also facilitate natural interactions such as drilling down and rolling up. Experimental results can assist libraries in book procurement activities. Although this paper only uses alliance library data for case analysis, the proposed visual analysis system model remains effective for other coalition data, enabling effective visual analysis and discovering potential knowledge hidden behind the data.

References

[1] Ren Lei, Du Yi, Ma Shuai, et al. Visual Analytics Towards Big Data [J]. Journal of Software, 2014, 25(9): 1909-1936.

[2] He Defang, Zeng Jianxun. Study on In-depth Integration of Library Collections Based on Semantics[J]. Journal of Library Science in China. 2012, 38(200): 79-87.

[3] Zhao Yingchun. Application of Grey Relation Analysis Method in the College Libraies' Books Acquisition[J]. Journal of Library and Information Sciences in Agriculture. 2016, 28(9): 114-118.

[4] Yin Jijun. Research on Book Purchasing System Based on Improved Genetic Neural Network [D]. Zhen Jiang: Jiangsu University, 2007.

[5] Li Yuan, Hu Rong. The Application of Fuzzy Comprehensive Evaluation Method in the Document Purchasing of University Library[J]. Journal of Library and Information Sciences in Agriculture. 2014, 26(5): 72-75.

[6] Chi Chunjia, Mao Zhiyong. Research on Assistant Decision-making in Formulating University Library Book Purchasing Plan Based on Data Mining[J]. Journal of Modern Information, 2009, 29(7): 108-110.

[7] Feng Na. A Brief Discussion of University Library' s Book Procurement Plan Based on Data Mining[J]. Journal of Library and Information Sciences in Agriculture, 2016, 28(4): 112-114.

[8] Zhao Haisen, Lǔ Lin, Bo Zhitao. Variational Circular Treemaps for Hierarchical Data[J]. Journal of Software, 2016, 27(5): 1103-1113.

[9] Schulz H J. Treevis.net: A Tree Visualization Reference[J]. IEEE Computer Graphics and Applications, 2011. 31(6): 16-23.

[10] Schulz H J, Schumann H. Visualizing Graphs—A Generalized View[C]//Proceedings of the Conference on Information Visualization (IV 2006). Washington, USA: IEEE Computer Society, 2006, 166-173.

[11] Tak S, Cockburn A. Enhanced Spatial Stability with Hilbert and Moore Treemaps[J]. IEEE Transactions on Visualization and Computer, 2013. 19(1): 141-148.

[12] Lam H C, Dinov I D. Hyperbolic Wheel: A Novel Hyperbolic Space Graph Viewer for Hierarchical Information Content[J]. ISRN Computer Graphics, 2012(6): 487-493.

[13] Ham F V, Wijk J V. Beamtrees: Compact Visualization of Large Hierarchies[J]. Information Visualization. 2003. 2(1): 31-39.

[14] Chen Yi, Zhen Yuangang, Hu Haiyun, et al. Visualization Technique for Multi-Attribute in Hierarchical Structure[J]. Journal of Software, 2016, 27(5): 1091-1102.

[15] Chen Y, Zhang X Y, Feng Y C, et al. Sunburst with Ordered Nodes Based on Hierarchical Clustering: A Visual Analyzing Method for Associated Hierarchical Pesticide Residue Data[J]. Journal of Visualization, 2015. 18(2): 237-254.

[16] Bring Data to Life with SVG, Canvas and HTML[EB/OL]. [2016-11-04]. https://github.com/d3/d3.

[17] Vizuly. Weighted Tree [EB/OL]. [2016-11-04]. http://vizuly.io/product/weighted-tree/?demo=d3js.

[18] NPasha. Bipartite Graph [EB/OL]. [2016-11-04]. http://bl.ocks.org/NPasha.

Author Contributions

Yao Nan: Provided original data and conducted basic data analysis.
Shen Xuefeng, Ke Yongzhen: Proposed research ideas, designed research plans, and finalized the manuscript.
Shen Xuefeng: Conducted experiments, collected, cleaned, and analyzed data, and drafted the manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is stored by the authors, E-mail: 812876188@qq.com.
[1] Shen Xuefeng. library_{book}, library_{lent}. Original collection data and original book borrowing records.
[2] Shen Xuefeng. book_{denoised}. Collection data after deduplication.
[3] Shen Xuefeng. lent_{Statistical} data. Borrowing record statistical data.

Received: 2016-11-14
Revised: 2017-02-23

Note: Figure translations are in progress. See original paper for figures.