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 problems in the knowledge discovery process of current alliance data, a visual analysis system model for alliance data is designed to enable collection, mining, and visual analysis of historical information. Method: A visual analysis system model for alliance data is constructed, a big data platform is built, and the usability of the model is verified. 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: The system can perform visual analysis on the alliance' s historical data, providing scientific data support for decision-makers.

Full Text

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 data 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 data analysis system model, 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 analysis. [**Limitations**] The current visual analysis result views could be further enriched. [**Conclusions**] The system can

visually analyze historical data from library alliances and provide scientific data support for decision-makers.

Keywords: Coalition Data, Big Data, Visual Analysis, Borrowing Records
Classification Number: TP311, G350

With the rapid development of information technology, multi-institution data sharing coalitions have become increasingly common. A major challenge facing such data coalitions is how to effectively organize and mine meaningful results from massive information resources while establishing an interactive data mining model. Visual analytics is a science and technology that employs interactive visual interfaces to assist users in analyzing and reasoning about large-scale complex datasets. This approach can address problems arising from information overload and ineffective communication, discover deep potential knowledge hidden within massive resources, reveal deeper connotations of results, and enhance result 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 achieve knowledge discovery, decision analysis, and policy formulation.

To integrate high-quality resources from Tianjin' s university libraries, the Tianjin Municipal Government 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 also establishing gateway-based interlinking with the existing Unicorn systems of Tianjin University and Nankai University to enable shared bibliographic data. 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 resource sharing among member libraries, and automating library operations including acquisition, cataloging, circulation, periodical management, public inquiry, and interlibrary loans to improve 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 data references for literature procurement decisions in various libraries. 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 primarily focuses on single-library data analysis. 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 scientifically

assess the importance of different book categories. 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. Li Yuan et al. utilized fuzzy comprehensive evaluation to analyze borrowing data and establish a fuzzy comprehensive evaluation model for university library literature procurement, determining reasonable procurement budgets for different literature resource types.

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

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

To observe, understand, and master 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: 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:** Performs deduplication, noise reduction, feature extraction, and related tasks to prepare data for visual analysis. 4. **Data Analysis:** Includes text vectorization representation and performs association analysis, statistical analysis, and other analytical functions on preprocessed data. 5. **Data Visualization:** Conducts visualization based on the D3 visualization component.

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

3.1 Data Preprocessing

The data preprocessing module must handle borrowing data and collection data. Borrowing data contains five fields: ID, timestamp, institution, barcode, and username. However, due to historical reasons, some library data encountered issues during consolidation into the joint library: - Data type 1: `1907863|CJ495415|2014122615|民航大学馆|张三`, where the second, third, and fourth items are out of order. - Data type 2: `1907864|M1214789|2014122615|商学院馆|C00624610|王五`, where the second item is redundant data.

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

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

Input: Call number set S
Output: Classification number

The algorithm traverses the set, taking the first character of each element (0 n length) to form set K. It then calculates the weight w_j for each element k_j in K, selects the k_j corresponding to max(w_j), and extracts from S the subset of elements whose first character begins with k_j to obtain S_1. The same process is applied to the second character of S_1 to obtain S_2. This operation continues sequentially for the i-th character of set S to obtain S_k until the number of elements in 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 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 display view. To represent quantity comparison relationships between different categories in the tree structure, we chose weighted trees for result presentation. A weighted tree displays each node' s weight alongside the node itself, using node size to represent weight magnitude. We employed 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 specific book categories, we define the book borrowing ratio ir as shown in equation (1):

$$ir = \frac{lent}{stock}$$

   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: collection quantity, borrowing quantity, and borrowing ratio. Using D3' s arc generator as the basic graphic framework, where each arc contains information including start angle, end angle, inner radius, and outer radius, we implemented this view using a polygon layout algorithm.

   Polygon Layout Algorithm:
   Input: Arc sequence arc_a, data, parameters , ,
   Output: Arc sequence arc_b, arc width w_i, triangle vertices

   Arc width calculation is shown in equation (2):

$$w_i = \text{[formula preserved as in original]}$$

   Arc point coordinate calculations are shown in equations (3) and (4):

$$x = \sin(\theta), \theta \in [1, 6]$$

$$y = \cos(\theta), \theta \in [1, 6]$$

   External point coordinate calculations are shown in equations (5) and (6):

$$x = \text{[formula preserved as in original]}$$

$$y = \text{[formula preserved as in original]}$$

When = 1, 2, 4, is = ; when = 3, 5, 6, ie = .

In this paper, the polygon layout algorithm parameters are = 100, = [value], = [value]. Additional required parameters include: outer radius, arc start angle s, and arc end angle e.

3. **View of Books and Libraries**
To understand the relationship between books borrowed by readers and libraries, we designed a third view. With 26 libraries in the university joint library system, 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, different libraries are independent, and book categories and libraries are mutually independent, G forms a bipartite graph. For bipartite graph visualization, we referenced components from Pasha' s work on bipartite graph visualization.

Books can be divided into 22 major categories by classification number. 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 Sources

The data used in this study comes from the Unicorn library automation management system of Tianjin' s university digital libraries, covering the period from system implementation to February 2015.

4.2 Visual Analysis

1. **Collection Book Analysis**
The hierarchical structure of books in the university joint library is shown in [Figure 2: see original paper]. From Figure 2: see original paper, we can clearly observe the relative relationships among book categories: industrial technology books are the most numerous, followed by literature, economics, and language/linguistics, 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 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: see original paper]. The inner ring represents collection quantity for each category, orange triangles represent borrowing quantity, and blue triangles represent borrowing ratio ir. From Figure 3: see original paper, we find that literature books have the highest borrowing ratio ir at 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: see original paper.

Figure 3: see original paper reveals 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 top three categories by borrowing ratio ir are light industry/handicrafts (28.2%), automation/computer science (24.3%), and architecture (21.4%), with atomic energy technology having the lowest ratio (1.6%). The generally low borrowing ratios reflect low utilization rates of collection books, consistent with the objective reality of reduced demand for paper books in the internet environment.

3. **Analysis of Books and Different Libraries**
[Figure 4: see original paper] displays the relationship between different book categories and libraries. The university joint library system comprises 26 libraries, and this view enables analysis of borrowing patterns for the same book 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: see original paper]. From [Figure 3: see original paper], we identified three industrial technology subcategories with borrowing ratios ir exceeding 20%: automation/computer technology, architecture, and light industry/handicrafts. Figure 5: see original paper shows that automation/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%). Figure 5: see original paper indicates that 58% of architecture books are borrowed by Urban Construction College Library. Figure 5: see original paper shows that light industry/handicrafts readers are primarily concentrated at Tianjin Polytechnic University Library (45%) and University of Science and Technology Library (18%), although Fine Arts College Library readers account for only 8% in this category, Fine Arts College Library shows particularly high demand–Figure 5: see original paper reveals that Fine Arts College Library users borrow 32% of light industry/handicrafts books. This may relate to school majors or reader preferences, requiring further investigation. This data can serve as a basis for allocating procurement budgets among different schools and for further allocation within schools across book categories.

Further analysis of the three categories with relatively low borrowing rates is shown in [Figure 6: see original paper]. From Figure 3: see original paper, we identified three categories with borrowing ratios around 2%: atomic energy technology (1.6%), mining engineering (2.0%), and metal-

lurgical industry (2.3%). Readers of these three categories are primarily concentrated at Polytechnic University Library and Industrial University Library, but Figure 6: see original paper shows that Medical University Library readers account for 6% of borrowing in these categories. While the reasons require further study, this data reflects reader demand for these categories, enabling procurement staff to make purchases based on reader needs.

The view can also analyze different demands of different library readers for various book categories, as shown in [Figure 7: see original paper] and [Figure 8: see original paper]. We find that readers across all libraries generally show high demand for automation/computer science books within the industrial technology category, particularly at medical libraries, which should be considered during procurement to address these readers' needs.

Through multi-dimensional visual analysis of the alliance library' s collection data and reader borrowing records, we identified the problem of low book borrowing rates in alliance libraries. Multi-view visual analysis methods not only more clearly display data hierarchical structures but also facilitate natural interactions such as hierarchical drilling down and rolling up. Experimental results can assist libraries in book procurement activities. Although this paper 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] 任磊, 杜一, 马帅, 等. 大数据可视分析综述 [J]. 软件学报, 2014, 25(9): 1909-1936. (Ren Lei, Du Yi, Ma Shuai, et al. Visual Analytics Towards Big Data [J]. Journal of Software, 2014, 25(9): 1909-1936.)

[2] 贺德方, 曾建勋. 基于语义的馆藏资源深度聚合研究 [J]. 中国图书馆学报, 2012, 38(200): 79-87. (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] 赵迎春. 灰色关联分析在高校图书馆图书采购中的应用 [J]. 农业图书情报学刊, 2016, 28(9): 114-118. (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] 尹纪军. 基于改进遗传神经网络的图书采购系统研究 [D]. 镇江: 江苏大学, 2007. (Yin Jijun. Research on Book Purchasing System Based on Improved Genetic Neural Network [D]. Zhen Jiang: Jiangsu University, 2007.)

[5] 李媛, 胡蓉. 模糊综合评判法在高校图书馆文献采购中的应用 [J]. 农业图书情报学刊, 2014, 26(5): 72-75. (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] 迟春佳, 毛志勇. 基于数据挖掘的高校图书馆图书采购计划辅助决策研究 [J]. 现代情报, 2009, 29(7): 108-110. (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] 冯娜. 浅议基于数据挖掘的高校图书馆购书计划 [J]. 农业图书情报学刊, 2016, 28(4): 112-114. (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] 赵海森, 吕琳, 薄志涛. 面向层次化数据的变分圆形树图 [J]. 软件学报, 2016, 27(5): 1103-1113. (Zhao Haisen, Lǚ 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): [page numbers].

[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): [page numbers].

[14] 陈谊, 甄远刚, 胡海云, 等. 一种层次结构中多维属性的可视化方法 [J]. 软件学报, 2016, 27(5): 1091-1102. (Chen Yi, Zhen Yuangang, Hu Haiyun, et al. Visualization Technique for Multi-Attrbute 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 Contribution Statement:
Yao Nan: Provided original data and performed basic data analysis.
Shen Xuefeng, Ke Yongzhen: Proposed research ideas, designed research plan,

finalized the paper.
Shen Xuefeng: Conducted experiments, collected, cleaned and analyzed data, drafted the paper.

Conflict of Interest Statement:
All authors declare no conflict of interest.

Supporting Data:
Supporting data is self-archived by the authors, E-mail: 812876188@qq.com.
[1] Shen Xuefeng. library_book, library_lent. Original collection data and original borrowing records.
[2] Shen Xuefeng. book_denoised. Collection data after deduplication.
[3] Shen Xuefeng. lent_Statistical data. Borrowing records statistical data.

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

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