Identifying Digital Libraries Author Publication Pattern Using Visualization Clustering Analysis

Chunsheng Huang
School of Information Studies
University of Wisconsin-Milwaukee
huang22@uwm.edu

Abstract

Information retrieval visualization (IRV) is a powerful tool in transforming the invisible abstract data along with their semantic relationships in a data collection into a visible display and provides visualization of the internal retrieval processes for users. It assists individual to make full use of his/her own creativity and imagination to search for information from an interactive system. One of the important features of IRV is to provide an intuitive way to recognize cluster pattern in the retrieved data. The purpose of this project is to employ information visualization method to explore and perform an author clustering analysis in an online citation database.

The visualization environment for this project is the Multidimensional Scale (MDS), which is a set of related statistical techniques often used in information retrieval visualization for exploring similarities or dissimilarities in data. The technique is applied to discover relationships among information retrieval objects by visualizing them and presenting their geographic representations in a low dimensional display space. Web of Science® was chosen to serve as data source for this project for the reasons of authoritative and reliable concern. Data were collected using “digital library” as the subject field being used to search for target authors in Web of Science. In order to make the scale of the project much more manageable, the number of the most influential authors in the selected area was set to range from fifty to one hundred. Filtered by the number of publications, 70 researchers were qualified as the target authors, whose record counts in Web of Science ranging from twelve to four. Since the analysis is about the proximity of the authors, this analysis incorporates the entire publications of target authors, not limiting to the subject field of digital library. In terms of the proximity between any two authors in the visual analysis, it is primarily defined by the similarity of their publication keywords. The keywords were parsed into single words and formed a keyword-author frequency table. Then another author-author proximity matrix was constructed. The proximity matrix of the target authors was accordingly applied to perform similarity measures, including Pearson coefficient, overlap coefficient, Jaccard coefficient, and Dice coefficient.

This project is still in the exploratory phase. The stress values derived from the first phase are all lower than 0.15 with the lowest value of 0.13271 and all values of Squared Correlation Index (RSQ) are close to or larger than 0.9. The data of the MDS results were transformed into a multimeida file with vitality colors and in three-dimensional displays. The visualization result using similarity measure of the of this project clearly demonstrates the relationships between the target authors. Four clusters can be easily identified, namely red cluster (author #9, 28, 48) on the top, light green cluster (author #10, 35, 42, 64) on the left, yellow cluster (author # 52, 62) on the right, and blue cluster (author # 47, 55, 40) at the bottom.

In the second phase of the project, results of MDS visualization are to be confirmed using two different traditional clustering methods to improve the quality of the analysis. The two methods used to confirm the visual analysis result are hierarchical clustering algorithm and K-Means. The advantage of combining the visual-clustering analysis and the traditional clustering method is that it cannot only visually display the clusters in a flexible and intuitive way, but also demonstrate the clear grouping boundaries among the clusters. The two could be complemented with each other. The hierarchical clustering algorithm yields a multiple level categorical tree structure, dendrogram. It demonstrates the clusters of nearest neighbor in the data. The four clusters in MDS also appear to be the nearest neighbors in the dendrogram. The second clustering method, K-Means, identifies relatively homogeneous groups of cases based on selected characteristics. In this project, the target authors were partitioned into six categories. The previous MDS clusters also belong to the same categories. Although the groupings of the two methods are different from each other, the pattern of the MDS clusters remains the same. It can be concluded that results of the two traditional clustering analyses confirm the similarity patterns of MDS.

Several limitations of the projects are reported. Firstly, the stress value is slightly over 0.1. Secondly, the authority control of authors in Web of Science is problematic. Thirdly, the selection of the key words employed in this project includes only Keyword Plus in Web of Science. Future research could compare the similarities and differences among retrieved data using different analysis strategies, authority control, and combination of keywords to ex-

plore more possibilities of the visual representation method.

References

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3), 401-420.

Korfhage, R.R. (1997). *Information storage and retrieval*. New York: Wiley.

Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-27.

Rasmussen, E. (1992). Custering algorithms. In W. B. Frakes, & R. Baeza-Yates (Eds.), *Information retrieval: Data structures & algorithms* (pp. 419-442). Englewood Cliff, NJ: Prentice Hall.

Saracevic,.T. (2009). Information science. In: Marcia J. Bates and Mary Niles Maack (Eds.) *Encyclopedia of Library and Information Science*. New York: Taylor & Francis. pp. 2570-2586.

Torgerson, W. S. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4), 401-17.

Zhang, J., & Korfhage, R. R. (1999). DARE: Distance and angle retrieval environment: A tale of the two measures. *Journal of the American Society for Information Science*, 50(9), 779-787.

Zhang, J., & Rasmussen, E. (2001). Developing a new similarity measure from two different perspectives. *Information Processing and Management*, 37(2), 279-294.

Zhang, J., & Rasmussen, E. (2002). An experimental study on the iso-content-based angle similarity measure. *Information Processing & Management*, 38(3), 325-342.

Zhang, J. (2008). *Visualization for information retrieval*, Springer.

Zhang, J., Wolfram, D., & Wang, P. (2009). Analysis of query keywords of sports-related queries using visualization and clustering. *Journal of the American Society for Information Science and Technology*, 60(8), 1550-1571.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying web search session patterns using cluster analy sis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 60(5), 896-910.

Zhang, J., & Wolfram, D. (2009). Visual analysis of obesity-related query terms on HealthLink. *Online Information Review*, 33(1), 43-57.

Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology*, 59(12), 1933-1947.