Visualization of Corpus Data by a Dual Hierarchical Data Visualization Technique

Takayuki Itoh Haruho Tachibana
Graduate School of Humanities and Sciences, Ochanomizu University
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
itot at is.ocha.ac.jp

Abstract

The paper presents a technique for visualization of corpus data consists of thousands of Japanese newspaper articles, and introduces several interesting trends discovered from the results. The technique first generates keyword-document matrices from the newspaper corpus, and respectively applies hierarchical clustering for rows and columns of the matrices. It then displays the two sets of clusters applying our own dual hierarchical data visualization technique. The visualization technique provides a mechanism to interact the two visualization components each other, so that users can freely explore the detail of the corpus data. This paper first describes the algorithm of the dual hierarchical data visualization technique, and then introduces our implementation and experiments of the visualization of the newspaper corpus data.

Keywords: *Visualization, Clustering, Newspaper Corpus, Hierarchical Data, Matrix Data.*

1 Introduction

Matrix data is a very common data in our daily life, and therefore matrix data visualization technique is an active research topic. Keyword-document matrix data is a typical matrix data used for various computer science fields, which collect importance values of selected keywords in a set of documents. The matrix data can be used for visualization of document data [15].

In many cases matrix data in our daily life is very sparse, and it is not always reasonable if we represent such sparse data using matrix-oriented visualization techniques. For example, if we have $n \times m$ matrix and values of 90% of the data items are zero, we may feel the matrix-oriented representation looks redundant. Many past visualization studies have attempted to save the display space usage by converting such sparse matrix data into node-link data such as tree or graph. It can drastically save the display space usage in many cases, because it represents only n or m data items. Matrix-oriented and node-oriented visualization techniques have their own bottlenecks, and some technical papers compare the readability of visualization results between the techniques [7].

This paper presents a new node-oriented visualization technique for such sparse matrices. The technique first generates clusters of rows and columns of the matrices. It then generates two hierarchical data from the clusters, and visualizes those using our own dual hierarchical data visualization technique. Also, our implementation provides a mechanism to interact the dual visualization components each other, so that users can explore rows and columns alternately. The mechanism is especially useful to realize detail-on-demand access of the matrix data.

We applied the dual hierarchical data visualization technique to the visualization of Japanese newspaper corpus data. Figure 1 shows the procedure of our implementation for the visualization of the corpus data, and its snapshot. Our implementation first extracts articles from the corpus, and calculates importance values of selected keywords for each article. It then constructs keyword-article matrix collecting the importance values, and visualizes the matrix data. We believe the technique can contribute so that users can freely explore and access to the interested parts of the large corpus. The paper proves the effectiveness of our technique for the newspaper corpus, by introducing several interesting trends discovered from the visualization results.

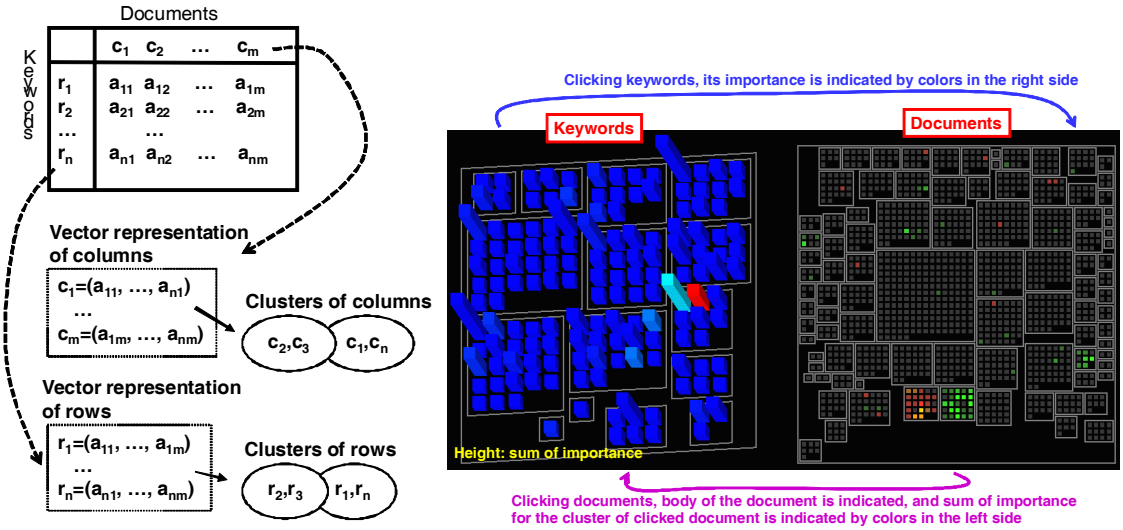

Figure 1. Overview of the newspaper corpus visualization technique applying our own dual hierarchical data visualization technique. (Left) The technique first generates clusters of columns and rows of keyword-document matrix data. (Right) The technique then visualizes the two sets of clusters applying the dual hierarchical data visualization technique.

2 Related Work

2.1 Hierarchical Data Visualization

The technique presented in this paper utilizes our hierarchical data visualization method presented in [9, 10]. This method represents a hierarchy as nested rectangles, and leaf-nodes as painted icons. The visualization technique places thousands of leaf-nodes into one display space while satisfying the following conditions:

- It does not overlap the leaf-nodes and branch-nodes in a single hierarchy of other nodes.
- It attempts to minimize the display area requirement.
- It draws all leaf-nodes by equally shaped and sized icons.
- It attempts to minimize aspect ratio and area of rectangular subspaces.

One of the features of our hierarchical data visualization technique is all-in-one visualization of lower-level data items, rather than representation and navigation of hierarchy between parents and children nodes. Many of well-known hierarchical data visualization techniques can be categorized as node-link and space-filling approaches. Space-filling visualization technique is an approach to represent all lower-level data items in one display space as well as our own technique. Quantum Treemap [4] archives to represent leaf-nodes as equally shaped and sized. Target of the Quantum Treemap is very similar to our own technique, and actually Quantum Treemap can be an alternative of our technique for the purpose of this paper. Experiments described in [9] discusses trade-offs between Quantum Treemap and our technique, where our technique had better results in aspect ratio of subregions and stability of layout among similar hierarchical data.

As well as our technique, recent Treemap also proposes hierarchy visualization from tabular data [5]. Their approach differs from ours since it clusters rows according to various attributes, since ours independently clusters columns and rows and provides user interface to interact them each other.

2.2 Matrix Data Visualization

Visualization of matrix data is an active research topic. Some early excellent works for visualization of tables, matrices, and spreadsheets, focused on exploration and navigation of the data [13]. Matrix data visualization can be often applied as an alternative of node-link data visualization for graphs or networks [3]. This approach often improves occlusion problems of graph/network visualization techniques. However, display usage of the approach is not very efficient if the input data is not dense.

Matrix zoom [1] uses both matrix- and node-

based visualization techniques, which provides overview of huge matrix and detailed view of subgraphs; however, they did not apply the node-link view to simultaneous representation of columns and rows as the presented technique focuses.

2.3 Multivariate Data Visualization

The presented technique treats columns and rows of matrix data as multivariate items. Recent parallel coordinates techniques [11] focus on cluster visualization of multivariate data, but we did not apply such techniques because it seems difficult to provide clickable user interface for arbitrary columns or rows. Principal component projection is also often used to visually discover clusters [12], but again, it is often difficult to make arbitrary data item clickable, because many data items may overlap each other on a display space. Another approach for multivariate data is using glyphs [6]. Current implementation of the technique presented in this paper simply represents attributes of data items by heights, colors, and shapes of bars, but more various representations can be realized by applying glyphs.

2.4 Document Data Visualization

Document data is one of the typical applications of information visualization. Many of presented document data visualization techniques target representation of temporal thematic change, for discovery of time-varying patterns and trends of documents. ThemeRiver [8] visualizes temporal change of frequency of specific keywords applying a visual metaphor of a river, for visual analysis and discovery of temporal thematic patterns and trends. Wong et al. [17] presented a technique to visualize frequent sequential patterns from time series dataset of documents.

Many of other techniques focus on visualization of correlations among large number of documents, where our technique also focuses on this point. Galaxies [16] represents multi-dimensional feature values of documents as a 2D scatterplot map. InfoSky [2] plots documents in hierarchically structured spaces. These techniques well works to represent distances among documents or clusters of documents: however, the visualization results may be very sparse. Since we would like to use the document data visualization technique as graphical user interface, it is desirable that metaphors of documents are easily clickable. That is one reason why the technique presented in this paper applies our own hierarchical data visualization technique.

DualNAVI [15] is one of the techniques most analogous to ours. It displays a list of documents in the left window, and a graph of keywords in the right window, and provides a graphical user interface to interlock the left and right part of the display each other. However, it is difficult to display hundreds of documents and keywords in one display space without scroll operation.

3 Dual Hierarchical Data Visualization

The paper presents a new technique to visualize matrix data by applying dual hierarchical data visualization technique. Features of the presented technique are as follows:
Feature 1) The technique represents the hierarchical structures of rows and columns, rather than each of multivariate values of rows and columns.
Feature 2) The technique represents clustering results of rows and columns as hierarchical data, and aims to display all rows and columns in one display as clickable icons, rather than representing hierarchy relationship of parent and children nodes of the hierarchical data.
Feature 3) The technique represents the rows and columns as equally-sized small icons, because it saves the display area, and fairly represents the whole parts of the data.
Feature 4) The technique targets matrix data which have hundreds or thousands of rows and columns. We think these are appropriate data sizes for the technique, since it focuses on providing user interface so that users can easily click each of the icons of rows and columns.

3.1 Clustering of Rows and Columns

Let us describe the data items of matrix data as follows: columns c_1 to c_m, rows r_1 to r_n, and element values a_{11} to a_{nm}, as shown in Figure 1. Here the technique treats columns $C = (c_1, .., c_m)$ as n-dimensional vectors, such as $c_i = (a_{1i}, .., a_{ni})$. Similarly, the technique treats rows $R = (r_1, .., r_n)$ as m-dimensional vectors, such as $r_j = (a_{j1}, .., a_{jm})$. Value type of a_{ij} can be binary, integer, real, or any others which clustering algorithms can be applied. Example shown in Section 4 uses real values.

The technique generates clusters of columns, applying agglomerative clustering method according to Euclidian distances, where non-hierarchical clustering methods (i.e. Self organizing map, k-means method) can be also applied. Our implementation simply generates combination of columns by iteratively coupling the columns or groups of columns according to their similarity values, and then generates nested clusters by grouping the columns according user-defined

threshold values. Then, the process is also applied to rows as well as columns. Finally the technique generates two sets of nested clusters, and they can be treated as two hierarchical data.

3.2 Visualization of Clusters

The technique then visualizes the two hierarchical data. Let the visualization module for rows as the left part, and the visualization module for columns as the right part. Therefore, the left part visualizes n rows r_1 to r_n, and the right part visualizes m columns c_1 to c_m. They represent columns and rows as three dimensional bar charts, and nested clusters as nested rectangular borders. Here, colors, shapes, and heights of the bars vary according to application-oriented semantics.

3.3 Interaction between the Dual Hierarchical Data

Our technique provides two-way interaction between the left and right parts, so that users can interactively explore the data items. When users click a row in the left part, then the technique highlights columns in the right part, which are related to the clicked row in the left part. Similarly, when users click a column in the right part, then the technique highlights rows in the left part, which are related to the clicked column in the right part.

Suppose that a user clicks r_i in the left part. The technique calculates visual attributes of c_j using a_{ij}, according to user-defined conditions. The conditions can be defined according to application-oriented semantics, and an example is described in Section 4. Consequently, the technique highlights interesting columns in the right part.

Similarly, when a user clicks c_j in the right part, the technique calculates visual attributes of rows, and highlights interesting rows in the left part.

4 Corpus Data Visualization

Let keywords be r_1 to r_n, where n is the number of keywords. Also, let documents be c_1 to c_m, where m is the number of documents. Our implementation clusters the documents and keywords as described in Section 3.1, where a_{ij} is the importance of i-th keyword in the j-th document. It then displays the clustered keywords in the left part, and the clustered documents in the right part.

As described in Section 3.2, visual attributes of nodes (i.e. shapes, colors, and heights) can be configured according to application-oriented semantics. As described in Section 3.3, interaction between the left and right parts can be also configured according to application-semantics. We modified visual attributes and interaction mechanism between the left and right parts to customize for corpus data visualization as follows:

- Heights of icons in the left part are proportional to the sum of importance, $\sum_{j=1}^{m} c_{ij}$ for i-th keyword.
- Hues of icons in the left part represent the sum of importance in a specific cluster of documents, where redness denotes that importance is high, and blueness denotes that importance is low.
- A user can choice one of the conditions to filter the articles while calculating the sum of importance of the keywords.
- When a cursor points one of the icons of keywords, it indicates the keyword.
- When a user inputs a keyword, it highlights the icon, which corresponds to the keyword, in the left part.
- When a user clicks an icon or a cluster in the left part, it indicates a list of keywords inside the cluster.
- When a user clicks an icon of keyword in the left part, the right part then represents the importance of the keyword for each document by R of RGB values.
- When a user clicks another icon of keyword in the left part, the right part then represents the importance of the keyword for each document by G of RGB values.
- When a user clicks an icon in the right part, it indicates the body of the document.
- When a user clicks an icon or a cluster in the right part, the left part then calculates the hue of icons of keywords. The hue is calculated from the sum of importance of the keywords in the documents of the clicked cluster.

5 Experiments

We implemented the dual hierarchical data visualization technique on Java 1.5, and tested on IBM ThinkPad X60 (CPU 1.8GHz, RAM 1GB) with Windows XP.

5.1 Matrix Data Generation from a Japanese Newspaper Corpus

We used a corpus of Japanese Mainichi newspaper in 1998 and 1999, where articles are stored in XML format containing date, headline, body, and additional annotations. We extracted articles which have the keyword "business information" in their annotations, where 2178 articles are extracted in 1998, and 1400 articles are extracted in 1999. We then calculated importance of words for each document, and extracted top 200 words in 1998 and 1999 respectively, according to the sum of importance values. We applied "Chasen", an open software for morpheme analysis of Japanese documents, and "termex", also an open software for importance calculation of words in Japanese documents [1]. We looked over the 200 words and manually selected 150 words in 1998 and 1999 respectively; here we preferentially selected name of companies, name of items, technical and financial terms, because we thought such words would bring trends of business information. Finally, we generated keyword-article matrices in 1998 and 1999 respectively. Figure 2 illustrates the above processes.

5.2 Visualization Results

We generated two hierarchical data from the matrix generated by the above process. Remark that the examples shown in this section applied only one threshold for clustering, and therefore the visualization results are not nested.

Figure 3 is an example of visualization of the newspaper articles in 1998. Here we clicked a cluster in the left part (indicated by a circle [A]), and looked a list of keywords in the cluster. We clicked the two keywords "USA" and "Internet" in the list, and therefore several icons in the right part were highlighted. Here, icons corresponding to articles about Internet were highlighted in red, and icons corresponding to articles about USA were highlighted in green. Intensities of the highlighted icons denoted the importance of these keywords.

From this visualization result, we observed two clusters (indicated by circles [B] and [C]) that there were yellow icons inside. We clicked the cluster indicated by the circle [B] in Figure 3(upper), and observed that several bars were then highlighted in non-blue colors in the left part, as shown in Figure 3(upper). The highlighted bars corresponded to the following words: "financial information", "services", "free", and "personal computer", and actually many of articles in the clicked cluster were about on-line financial services.

We also clicked the other cluster indicated by the circle [C] in Figure 3(lower), and observed that several bars were then highlighted in non-blue colors in the left part, as shown in Figure 3(lower). The highlighted bars corresponded to the following words: "company", "business", "investigation", and "personal computer", and actually many of articles in the clicked cluster were about business innovations.

As a result of above operations, we discovered that there were two meaningful clusters of articles that were related to USA and Internet.

Figure 4 is an example of visualization of the newspaper articles in 1999. Here we focused on several tall icons in a cluster in the left part (indicated by a circle [A]), and clicked the two tall icons of keywords "design" and "Toyota" (name of Japanese automobile company), and then several icons in the right part were highlighted. Here, icons corresponding to articles about design were highlighted in red, and icons corresponding to articles about Toyota were highlighted as green. Intensities of the highlighted icons denoted the importance of these keywords.

From this visualization result, we observed two clusters (indicated by two circles [B] and [C]) that there were yellow icons inside. We clicked the cluster indicated by the circle [B] in Figure 4(upper), and observed that only two bars were then highlighted in non-blue colors in the left part, as shown in Figure 4(upper). The highlighted bars corresponded to "design" and "Toyota", and other keywords of names of companies or items were not highlighted. We found that Toyota's design brought many hot topics in 1999, rather than designs by other companies or designs of other items.

We clicked the other cluster indicated by the circle [C] in Figure 4(lower), and observed that more than two bars were then highlighted in non-blue colors in the left part, as shown in Figure 4(lower). The highlighted bars corresponded to the following words: "Asahi Beer", "Matsushita Electric Industrial", and "development", in addition to "design" and "Toyota". Articles in the clicked cluster were about the joint brand by these companies. As a result of above operations, we discovered a cluster of such unexpected articles.

5.3 Feedback and Discussion

We asked 10 examinees to play with the presented technique and give us comments. All the examinees were female university student studied computer graphics, but none of them were experts of document visualization. This section introduces

[1] Chasen is distributed at http://chasen.naist.jp/hiki/ChaSen/. termex is distributed at http://gensen.dl.itc.u-tokyo.ac.jp/.

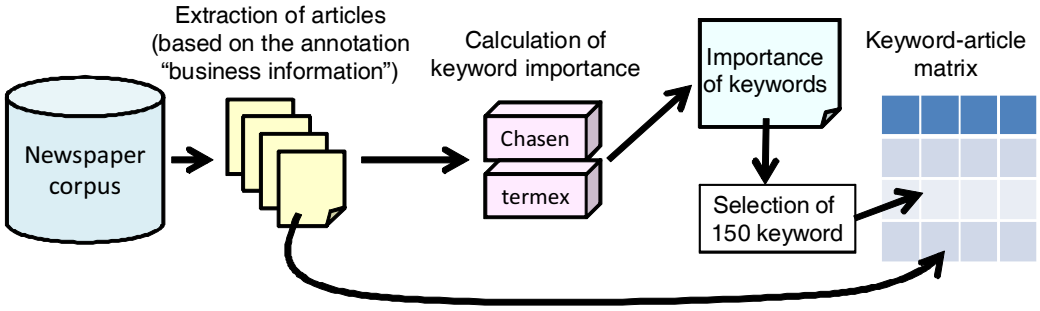

Figure 2. Processing flow for keyword-document matrix construction from Japanese newspaper corpus.

the feedback of examinees and discuss about the future improvements of the technique.

Most of the examinees said that they could immediately understand the meaning of visualization results and master the operation. Some of them said that height-based representation of keyword importance, in the left side, was useful to focus on hot keywords. Some of them also said that color representation of document importance, in the right side, was intuitive since hue denoted the corresponding keyword and intensity denoted the importance. Some of them said that it was very interesting to represent importance of keywords in a document cluster when they clicked in the right side, because they could imagine the contents of the cluster of documents from the highlighted keywords.

On the other hand, some of examinees mentioned problems of the presented technique. An examinee claimed that category-based keyword structuring may be better than current co-occurrence-based keyword clustering. Another examinee claimed that it might be sometimes confusing, because the left and right sides looked somewhat similar, but semantics of operation were quite different. Some others suggested interesting future works, which are listed in the next section.

6 Conclusion

This paper presented a technique for visualizing corpus data, which applies our own dual hierarchical data visualization technique. We extracted thousands of articles from a Japanese newspaper corpus, and constructed keyword-article matrices based on the importance values of the keywords in the articles. The paper presented several visualization results of the matrix data, and discussed visual discoveries of trends from the articles.

A journal paper on this work [14] describes more detail of algorithms and experiments.

Reflecting feedback of our examinees, we point out the following as potential future works, including:

- visualization of correlation of more than two keywords,
- application of more sophisticated clustering algorithm such as co-clustering,
- more usability tests and objective/subjective evaluations,
- visualization of non-newspaper corpus such as database of technical papers or patents, and
- application of the dual hierarchical data visualization to non-corpus data.

Acknowledgements

We appreciate Prof. Ichiro Kobayashi, Ochanomizu University, and Prof. Tsuneaki Kato, The University of Tokyo, for their suggestions on document processing. Japanese newspaper corpus has been provided by A Workshop on Multimodal Summarization for Trend Information (MuST). This work has been partially supported by Japan Society of the Promotion of Science under Grant-in-Aid fir Scientific Research (C) No. 18500074.

References

[1] Abello J., van Ham F., Matrix Zoom: A Visual Interface to Semi-external Graphs, *IEEE Information Visualization 2004*, 183-190, 2004.

[2] Andrews K., et al., The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities, *Information Visualization*, 1(3), 166-181, 2002.

[3] Becker R. A., et al., Visualizing Network Data, *IEEE Transactions on Visualization and Computer Graphics*, 1(1), 16-28, 1995.

[4] Bederson B., Schneiderman B., Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies, *ACM Transactions on Graphics*, 21(4), 833-854, 2002.

[5] Chintalapani G., et al., Extending the Utility of Treemaps with Flexible Hierarchy, *8th International Conference on Information Visualization*, 335-344, 2004.

[6] Forsell C., et al., Simple 3D Glyphs for Spatial Multivariate Data, *IEEE Information Visualization 2005*, 119-124, 2005.

[7] Ghoniem M., Fekete J., Castagiloia P., A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations, *IEEE Information Visualization 2004*, pp. 17-24, 2004.

[8] Havre S., et al., ThemeRiver: Visualizing Thematic Changes in Large Document Collections, *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 9-20, 2002.

[9] Itoh T., Yamaguchi Y., Ikehata Y., Kajinaga Y., Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, *IEEE Transactions on Visualization and Computer Graphics*, 10(3), 302-313, 2004.

[10] Itoh T., Takakura H., Sawada A., Koyamada K., Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, *IEEE Computer Graphics and Applications*, 26(2), 40-47, 2006.

[11] Johansson J., et al., Revealing Structure within Clustered Parallel Coordinates Displays, *IEEE Information Visualization 2005*, 125-132, 2005.

[12] Marks J., et al., Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation, *ACM SIGGRAPH '97*, 389-400, 1997.

[13] Rao R., Card S. K.., The Table Lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information, *Computing Systems (CHI'94)*, 318-322, 1994.

[14] Tachibana H., Itoh T., Sakyo and Ukyo: A Technique for Visualization of Large-scale Table Data, *The Journal of the Society for Art and Science*, 7(2), 22-33, 2008. (in Japanese)

[15] Takano A., et al., Associative Information Access Using DualNAVI, *Kyoto International Conference on Digital Libraries (ICDL'00)*, 285-289, 2000.

[16] Wise J. A., et al., Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents, *Reading in Information Visualization: Using Vision to Think*, 442-450, 1999.

[17] Wong P. C., et al., Visualizing Sequential Patterns for Text Mining, *IEEE Information Visualization 2000*, 105-111, 2000.

[18] Yang J., et al., InterRing: An Interactive Tool for Visually Navigating and Manipulating Hierarchical Structures, *Information Visualization*, 2(1), 16-30, 2003.

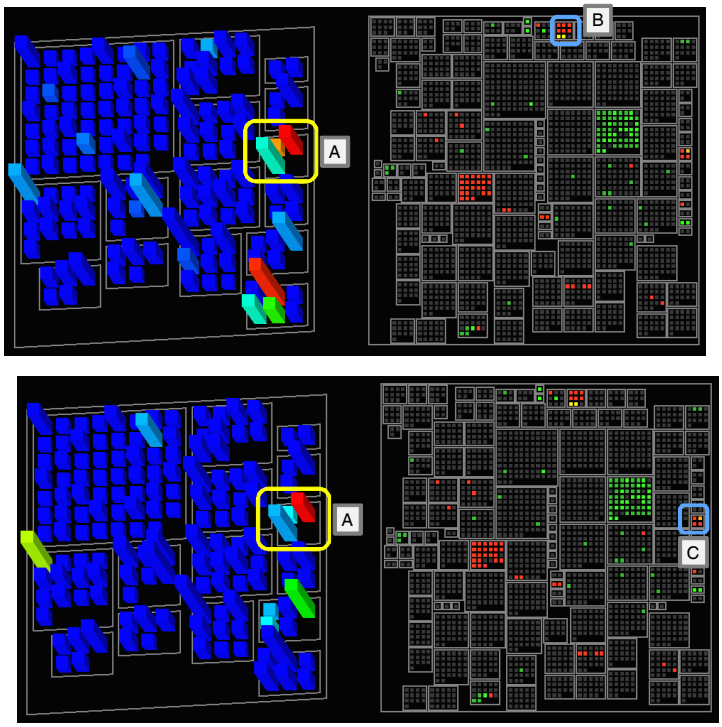

Figure 3. Example (1). The circle [A] surrounds icons corresponding to two keywords "USA" and "Internet", and the other circles [B] and [C] surround two clusters of different contents of articles related to both two keywords.

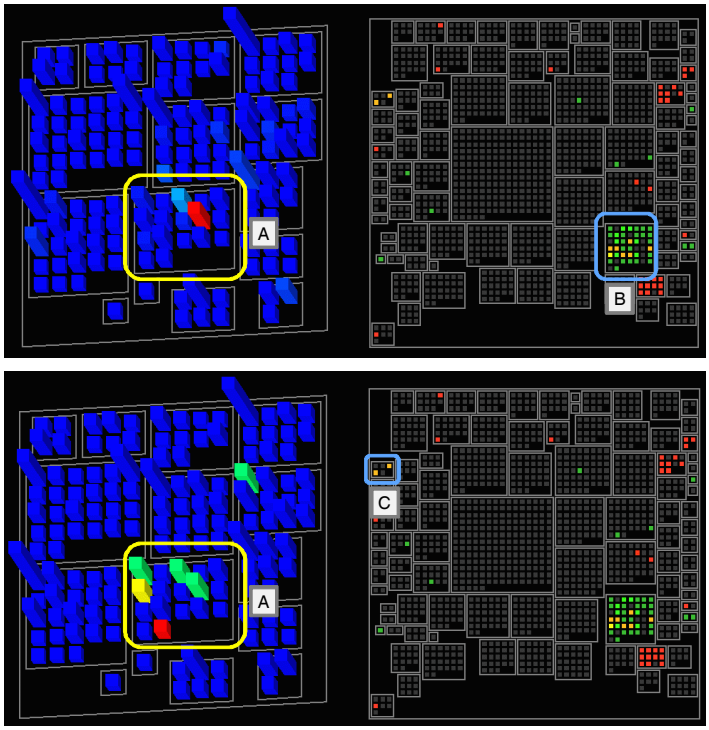

Figure 4. Example (2). The circle [A] surrounds icons corresponding to two keywords "design" and "Toyota", and the other circles [B] and [C] surround two clusters of different contents of articles related to both two keywords.