Visualization of Corpus Data by a Dual Hierarchical Data Visualization Technique

Haruho Tachibana*
Graduate School of Humanitics and Sciences, Ochanomizu University

Takayuki Itoh†
Graduate School of Humanitics and Sciences, Ochanomizu University

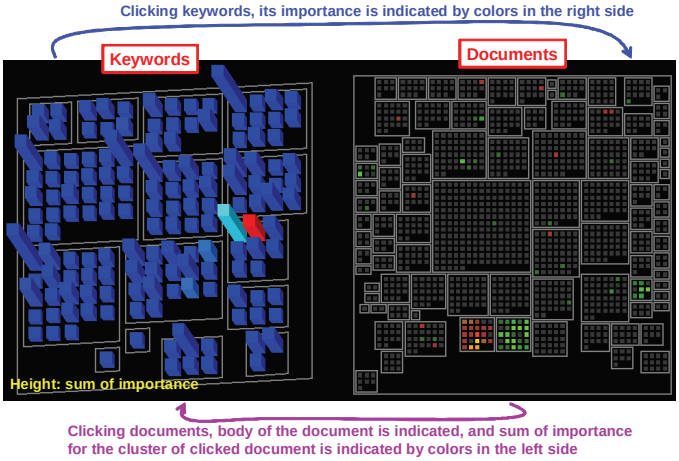

Figure 1: Overview of our dual hierarchical data visualization technique. The technique first generates clusters of keywords and documents, and then visualizes the two sets of clusters applying dual hierarchical data visualization technique.

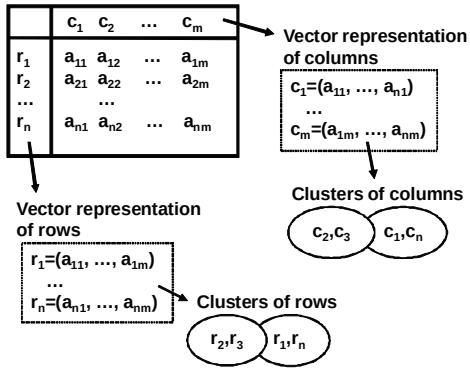

Figure 2: Clustering of columns and rows of matrix data.

ABSTRACT

This poster presents a technique for visualization of matrix data applying dual hierarchical data visualization technique. The presented technique first applies hierarchical clustering for rows and columns of the matrix data. It then visualizes the two sets of clusters applying dual hierarchical data visualization components. The poster also presents an application of the technique for visualization of keyword-document matrix generated from large-scale corpus data. The paper shows an example of visualization results of corpus data consists of thousands of Japanese newspaper articles, and introduces an interesting trends discovered from the result.

1 MATRIX DATA VISUALIZATION BY A DUAL HIERARCHICAL DATA VISUALIZATION TECHNIQUE

1.1 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 2. 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.

The technique generates clusters of columns by the following procedure. It first calculates Euclidian distances between every possible pairs of columns. Let the distance between c_i and c_j as d_{ij}, and the maximum distance of the all distances as D_{max}. Here the technique calculates the similarity value s_{ij} for d_i and d_j, as $s_{ij} = 1.0 - d_{ij}/D_{max}$, where 0 is the minimum value of s_{ij}, and 1 is the maximum value of s_{ij}. The technique then generates clusters of columns, applying agglomerative clustering method. 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.

Above processes are 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.

1.2 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, where colors, shapes, and heights of the bars vary according to application-oriented semantics.

1.3 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. Consequently, the technique highlights interesting columns in the right part. Similarly, when a user clicks c_j in

*e-mail: haruho@itolab.is.ocha.ac.jp
†e-mail: itot@computer.org

the right part, the technique calculates visual attributes of rows, and highlights interesting rows in the left part.

2 Implementation for Corpus Data Visualization

This section presents an implementation of corpus data visualization applying our dual hierarchical data visualization technique. 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, 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 shown in Figure 1, we modified visual attributes and interaction mechanism between the left and right parts to customize for corpus data visualization as follows: 1) Heights of icons in the left part are proportional to the sum of importance, $\sum_{j=1}^{m} c_{ij}$ for i-th keyword. 2) 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. 3) A user can choice one of the conditions to filter the articles while calculating the sum of importance of the keywords. 4) When a cursor points one of the icons of keywords, it indicates the keyword. 5) When a user inputs a keyword, it highlights the icon, which corresponds to the keyword, in the left part. 6) When a user clicks an icon or a cluster in the left part, it indicates a list of keywords inside the cluster. 7) 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. 8) 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. 9) When a user clicks an icon in the right part, it indicates the body of the document. 10) 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.

3 Experiments

3.1 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" [1], an open software for morpheme analysis of Japanese documents, and "termex" [2], also an open software for importance calculation of words in Japanese documents. 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, and visualized them.

3.2 Example of Visualization Results

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 yellow circle), and looked a list of keywords in the cluster. We clicked the two keywords "USA" and "Internet" in the list, and therefore

[1]Distributed at http://chasen.naist.jp/hiki/ChaSen/.

[2]Distributed at http://gensen.dl.itc.u-tokyo.ac.jp/.

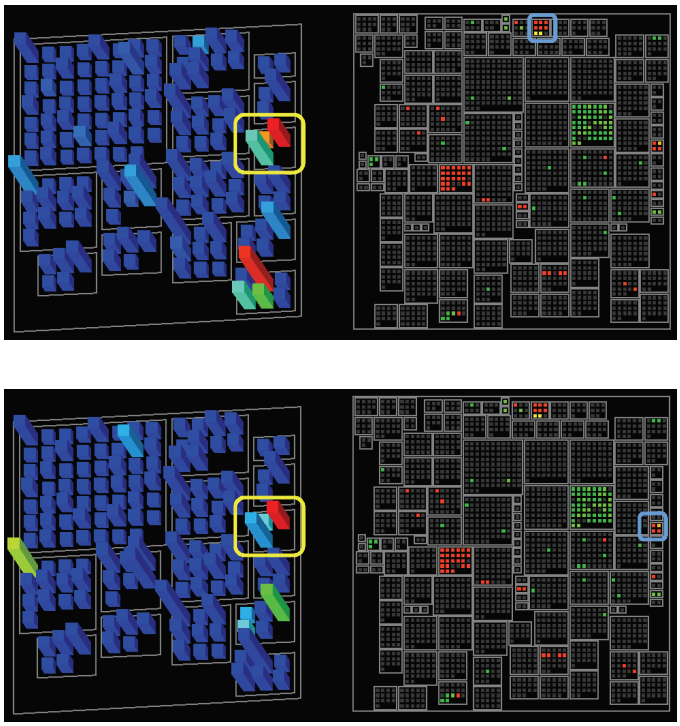

Figure 3: Example of visualization results.

several icons in the right part were highlighted. Here, red icons denote articles about Internet, green icons denote articles about USA, and their intensities denote the importance of these keywords.

From this visualization result, we observed two clusters (indicated by blue circles) that there were yellow icons inside. We clicked the cluster indicated by the blue circle in Figure 3(upper), and observed that several bars are then highlighted as 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 blue circle in Figure 3(lower), and observed that several bars are then highlighted as non-blue colors in the left part. 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 of companies.

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

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] 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.