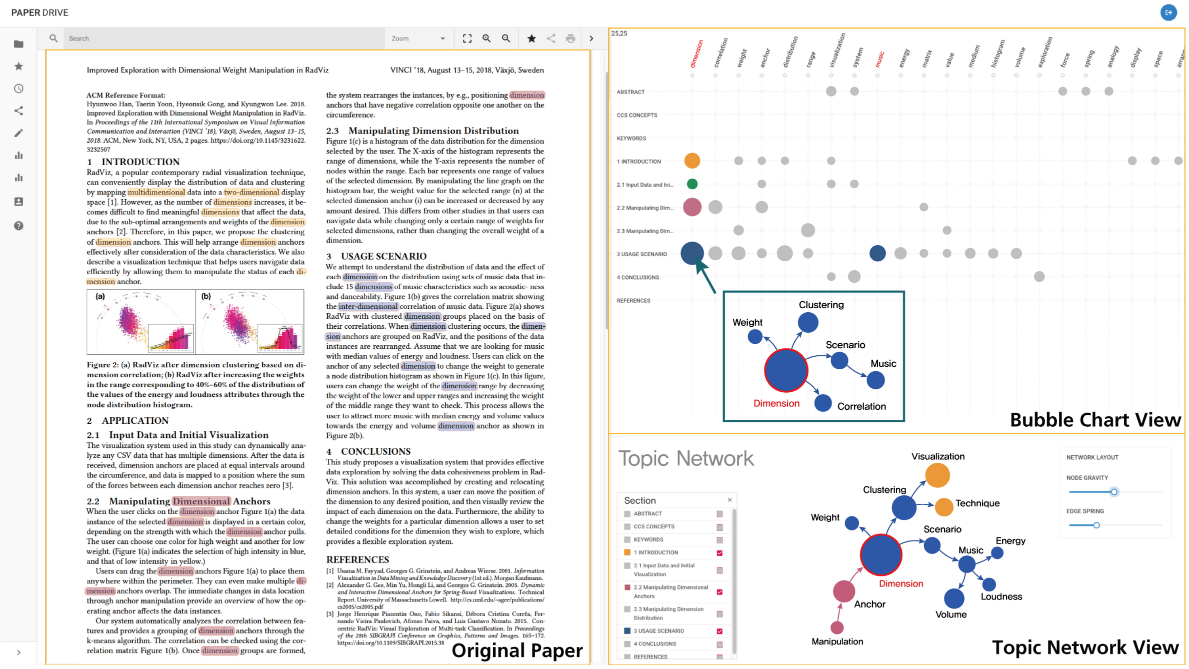
An Interactive Paper Summarization System through Topic Network Visualization

Hyunwoo Han, Hyoji Ha, Jaejong Ho, Hyeonsik Gong*
Lifemedia Interdisciplinary
Program, Ajou University

Junyup Hong†
Department of Cyber
Security, Ajou University

Soojung Lee, Juwon Hong, Kyungwon Lee‡
Department of Digital Media,
Ajou University



Improved Exploration with Dimensional Weight Manipulation in RadViz

VINCI '18, August 13–15, 2018, Växjö, Sweden

ACM Reference Format:
Hyunwoo Han, Taerin Yoon, Hyeonsik Gong, and Kyungwon Lee. 2018. Improved Exploration with Dimensional Weight Manipulation in RadViz. In *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction (VINCI '18), Växjö, Sweden, August 13–15, 2018.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3231622.3232507

1 INTRODUCTION

RadViz, a popular contemporary radial visualization technique, can conveniently display the distribution of data and clustering by mapping multidimensional data into a two-dimensional display space [1]. However, as the number of dimensions increases, it becomes difficult to find meaningful dimensions that affect the data, due to the sub-optimal arrangements and weights of the dimension anchors [2]. Therefore, in this paper, we propose the clustering of dimension anchors. This will help arrange dimension anchors effectively after consideration of the data characteristics. We also describe a visualization technique that helps users navigate data efficiently by allowing them to manipulate the status of each dimension anchor.

Figure 2: (a) RadViz after dimension clustering based on dimension correlation; (b) RadViz after increasing the weights in the range corresponding to 40%–60% of the distribution of the values of the energy and loudness attributes through the node distribution histogram.

2 APPLICATION

2.1 Input Data and Initial Visualization

The visualization system used in this study can dynamically analyze any CSV data that has multiple dimensions. After the data is received, dimension anchors are placed at equal intervals around the circumference, and data is mapped to a position where the sum of the forces between each dimension anchor reaches zero [3].

2.2 Manipulating Dimensional Anchors

When the user clicks on the dimension anchor Figure 1(a) the data instance of the selected dimension is displayed in a certain color, depending on the strength with which the dimension anchor pulls. The user can choose one color for high weight and another for low weight. (Figure 1(a) indicates the selection of high intensity in blue, and that of low intensity in yellow.)

Users can drag the dimension anchors Figure 1(a) to place them anywhere within the perimeter. They can even make multiple dimension anchors overlap. The immediate changes in data location through anchor manipulation provide an overview of how the operating anchor affects the data instances.

Our system automatically analyzes the correlation between features and provides a grouping of dimension anchors through the k-means algorithm. The correlation can be checked using the correlation matrix Figure 1(b). Once dimension groups are formed, the system rearranges the instances, by e.g., positioning dimension anchors that have negative correlation opposite one another on the circumference.

2.3 Manipulating Dimension Distribution

Figure 1(c) is a histogram of the data distribution for the dimension selected by the user. The X-axis of the histogram represents the range of dimensions, while the Y-axis represents the number of nodes within the range. Each bar represents one range of values of the selected dimension. By manipulating the line graph on the histogram bar, the weight value for the selected range (n) at the selected dimension anchor (i) can be increased or decreased by any amount desired. This differs from other studies in that users can navigate data while changing only a certain range of weights for selected dimensions, rather than changing the overall weight of a dimension.

3 USAGE SCENARIO

We attempt to understand the distribution of data and the effect of each dimension on the distribution using sets of music data that include 15 dimensions of music characteristics such as acoustic- ness and danceability. Figure 1(b) gives the correlation matrix showing the inter-dimensional correlation of music data. Figure 2(a) shows RadViz with clustered dimension groups placed on the basis of their correlations. When dimension clustering occurs, the dimension anchors are grouped on RadViz, and the positions of the data instances are rearranged. Assume that we are looking for music with median values of energy and loudness. Users can click on the anchor of any selected dimension to change the weight to generate a node distribution histogram as shown in Figure 1(c). In this figure, users can change the weight of the dimension range by decreasing the weight of the lower and upper ranges and increasing the weight of the middle range they want to check. This process allows the user to attract more music with median energy and volume values towards the energy and volume dimension anchor as shown in Figure 2(b).

4 CONCLUSIONS

This study proposes a visualization system that provides effective data exploration by solving the data cohesiveness problem in RadViz. This solution was accomplished by creating and relocating dimension anchors. In this system, a user can move the position of the dimension to any desired position, and then visually review the impact of each dimension on the data. Furthermore, the ability to change the weights for a particular dimension allows a user to set detailed conditions for the dimension they wish to explore, which provides a flexible exploration system.

REFERENCES

[1] Usama M. Fayyad, Georges G. Grinstein, and Andreas Wierse. 2001. *Information Visualization in Data Mining and Knowledge Discovery* (1st ed.). Morgan Kaufmann.
[2] Alexander G. Gee, Min Yu, Hongli Li, and Georges G. Grinstein. 2005. *Dynamic and Interactive Dimensional Anchors for Spring-Based Visualizations.* Technical Report. University of Massachusetts Lowell. http://cs.uml.edu/~agee/publications/cs2005/cs2005.pdf
[3] Jorge Henrique Piazentin Ono, Fabio Sikansi, Débora Cristina Corrêa, Fernando Vieira Paulovich, Afonso Paiva, and Luis Gustavo Nonato. 2015. Concentric RadViz: Visual Exploration of Multi-task Classification. In *Proceedings of the 28th SIBGRAPI Conference on Graphics, Patterns and Images.* 165–172. https://doi.org/10.1109/SIBGRAPI.2015.38

Figure 1: A paper inquiry system based on topic network visualization. Users are able to see their original uploaded paper at the left part of the system, and they can identify the summarization of the paper with the visualization in the right part. In the bubble chart view, they can grasp the distributions of topic keywords by each section. The topic network view allows users to figure out the network relations among the nodes selected above the bubble chart view.

Abstract

This study proposes a web system with visualization tools to help users easily explore and summarize topic keywords and contents of specific papers in which they are interested. The system is composed of three views; 1) A view that shows the original paper uploaded by user, 2) A bubble chart view that displays the distribution of topic keywords by each section of the paper, 3) A topic network view that shows the relation of keywords related to the selected nodes from the bubble chart. Through these views, users can not only understand what topic keywords are important in each section of the paper but can also quickly identify the content of the paper by checking the flow of the keywords' network.

Keywords: Topic Network Visualization, Automatic Summarization, Paper Archive.

Index Terms: Information interfaces and presentation—User Interfaces—Graphical user interfaces

1 Introduction

The process of finding relevant papers from a number of papers plays an important role in helping researchers to conduct their own research. Researchers read a lot of papers to get a reference for their research and to propose a new contribution to the related fields.

However, because of the massive amount of paper containing various research methods, researchers have to spend a lot of time searching for a proper paper they need. So there has been an emphasis on a system to solve this inefficient researching process [1]. In order to meet these demands, a method of extracting important keywords from text data is required. Therefore, this paper aimed to create a visualization system to make users easily and quickly understand the content of the papers they want to identify by using Latent dirichlet allocation (LDA) [2], an effective topic modeling method that can effectively summarize text data and identify topics. There are some previous studies about topic visualizations such as LDAvis [3], helping users understand text data by allowing users to inspect the terms associated with each topic. In this paper, we did not only focus on the previous studies' purposes, but also aimed to create a system that presents the distribution and relation in each

*e-mail: {ainatsumi, hjha0508, hapsoa, overholic10}@ajou.ac.kr
†e-mail: {wnsduqdl321}@ajou.ac.kr
‡e-mail: {skdufh, penguin373,kwlee}@ajou.ac.kr

section of topic keywords for making users understand the content of a paper more specifically. Our study focus on three objectives:

- Understanding the distribution of topic keywords by section using bubble chart view, and identifying flows among each keyword and summary of paper through topic network view.
- Analyzing and comparing networks created by each section or topic keywords via interactions.
- Applying an example paper to the system to demonstrate the usefulness of the study.

2 Methods

2.1 Data Processing

Once a research paper with PDF file format is uploaded, it will be saved in the database and converted into paper data in the mark-up language form. The paper most optimized for analysis in the web system is the VIS proceeding paper published by IEEE.

2.2 Visualization

2.2.1 Bubble Chart View

This view shows the distribution of the topic keywords of the uploaded paper for each section. The titles of the sections are arranged in rows, and the keywords are arranged in each column by summing up the importance of the topic keywords after the topic modeling. The size of each node is proportional to the frequency of appearance of the corresponding topic keyword for each section.

The user can hover each node and grasp the node-centric topic network. The network consists of keywords extracted by topic modeling in the section containing the hovered nodes. The edges among each network node consist of the dependency value of each word with syntax analysis At this time, the direction of the edge is formed toward the higher dependency value between two keywords. In addition, corresponding nodes may be selected through an interaction of clicking a node or dragging an area.

2.2.2 Topic Network View

This view shows a visualization of the relationships among the nodes selected in the bubble chart view. After merging each of the networks that appeared when hovering the selected nodes in the bubble chart view, the network is updated with some keywords that have high connection weights. The color of each node is colored by the color of the section which has the most corresponding keywords. With the left controller, users can select the sections and filter the information by the selected section to update network. In addition, when a node is clicked, the labeled keyword is highlighted in the left original paper view, allowing users to find out the actual usage of the term.

3 Usage Case

In this section, we introduce a scenario that assumes a user who tries to conduct multidimensional Radviz visualization has uploaded work by Han et al. [4], a related article, into the system.

First, the user can grasp the distribution of topic keywords by section of the uploaded paper through the Bubble chart view. If the user is interested in the topic 'dimension', the user can hover the 'dimension' node for each section to see the keyword network centered on that node. In the process, when the user hovers the topic keyword 'dimension' in the section 'Usage Scenario', the user can identify that three keywords 'dimension', 'scenario', and 'music' are connected in the order (Figure 2-Left).

Therefore, the user selects nodes corresponding to the keywords 'dimension' and 'music' to identify the topic network. The user checks only three sections - 'Introduction', 'Manipulating Dimensional Anchors' and 'Usage Scenario' - at the left controller to see a network containing only the selected section (Figure 2-Right).

Through the network, the user can identify that 'clustering', 'weight' and 'scenario' nodes are connected additionally around the 'dimension' node. As two nodes - 'visualization' and 'technique' are connected around the 'clustering' node, the user can grasp that the paper used a cluster visualization technique. Also, the 'scenario' node is connected to the 'music' node, which is connected to 'volume', 'loudness', and 'energy' nodes. Through this information, the user is able to identify that the paper used music data to write a scenario with three variables - volume, loudness, and energy.

Also, when the user clicks the node labeled with 'dimension', the 'dimension' keywords corresponding to the three sections checked before in the controller are highlighted in the left original paper view. In this way, users can freely search for topic keywords, set desired conditions, and then summarize specific contents of a paper by identifying topic maps.

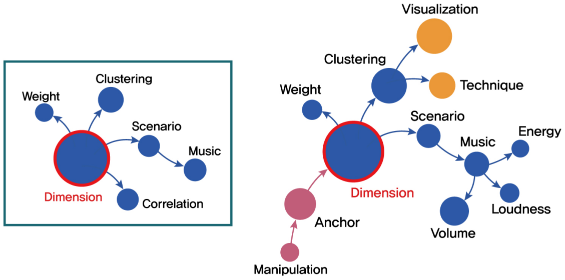


Figure 2: (Left) A keyword network which appears on a bubble chart view when hovering a node 'dimension' in the section 'usage scenario'. (Right) A keyword network on a topic network with the nodes selected from the bubble chart view.

4 Conclusion

The goal of this study is to develop a visualization system that helps users to reduce the time for finding relevant papers so that they can manage the research process efficiently. Through the visualization, they are able to identify topic keywords of paper by each section and can get summarized information of specific sections of the paper with some interactions which make a topic network from selected nodes. It also highlights where the keyword of the node appears in the actual paper so that the user can easily find the relevant contents and read them in detail. Furthermore, this system can improve the performance of topic modeling by reinforcing taxonomies considering options such as synonyms in preprocessing of text data, so that a more advanced system can be created later making the users can actually save the paper and use the system. Additionally, the study will perform a validation process for confirming whether the end users can actually use this system usefully.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5B6075104)

References

[1] International Society for Optics and Photonics. *Automatic paper summary generation from visual and textual information*, volume 11041, 2018.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[4] Hyunwoo Han, Taerin Yoon, Hyeonsik Gong, and Kyungwon Lee. Improved exploration with dimensional weight manipulation in radviz. In *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction*, pages 118–119, 2018.