Ph.D. Thesis

Visualization of Networks Including Categorized Nodes for Summarization and Comparative Analysis

Rina Nakazawa

Advanced Science

Graduate School of Humanities and Sciences

Ochanomizu University

March, 2020

Abstract

Various kinds of data in this world can be represented by networks or graphs. Such data structures consist of elements of data represented as nodes and relationships between the elements represented as edges. Networks in recent years include attributes of their nodes, and the nodes can be grouped by the attributes as categories. So we need to connect these two types of information, relationships between data elements and categories of the elements when we analyze network-structured data.

Network visualization is one of the ways to help network analysis including categorized nodes. Summarization of network topology is especially important to understand the whole structure of the data. Comparative view and interaction are also helpful to select a part of the data. This dissertation proposes visualization techniques of networks including categorized nodes for summarization and comparative analysis. Summary visualization bundles the edges to avoid visual clutters placing the nodes based on both similarities of their categories and relationships. Comparative visualization with interaction aims to correspond to the categories and the relationships between networks. The dissertation introduced applications to a gene network and a citation relationship as examples of summarization, and an application to citation and co-author relationships as a example of comparative analysis.

Keywords : Information Visualization, Visual Analytics, Edge Bundling, Gene Network, Citation Network, Co-author Network

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

There are various datasets which consist of relationships between elements like people or objects in the world, and they can be treated as network structures. The elements of these data are categorized based on their attributes. Such categorized network datasets have been larger-scaled and more complex. They cause difficulties in understanding the structure of the network. Information visualization works effectively when we want to grasp the whole structure of a large amount of information, extract important structures, find and predict new insights on it. Itoh argued that information visualization is effective for the following tasks [26].

1. **O**verview: Grasp and understanding the whole structure of network data
2. **C**larification: Validation of search results and clarification of new insights
3. **H**andling: Selection and comparison of information for analysis
4. **A**nnouncement: Explanation for people's communication

This dissertation focuses on visualization techniques of categorical network data for the first three tasks.

1.1.1 Definition of a Categorized Network

In this section, we define a categorized network that we focus on in this dissertation. Here, the state that the i-th node n_i has its attribute is expressed by using m

dimensional boolean variables b_i in the equation 1.1. m is the total number of categories.

$$n_i = \{b_1, \cdots, b_m\} \tag{1.1}$$

If a node n_i categorized into the j th category, b_j is true, and if not, it is false.

Although edges may sometimes have categorical attributes, this dissertation focuses on network datasets that only its nodes have categorical attributes. The examples of categorical network data include mutual interactions between genes and their functions, citation relationships between research papers and their topics, and social networks between people and communities. We expect such networks provide us many insights by summarization and observation of relationships between elements of the networks and their categorical information.

1.1.2 Tasks and Visualization Techniques for Analytics of a Categorized Network

Next, we describe types of visualization techniques of the categorical network for the three tasks, overview, clarification, and handling. Here, many famous studies on information visualization follow the Shneiderman' s mantra [121].

"*Overview first, zoom and filter, then details-on-demand.*"

This mantra corresponds to **O** and **H** in Itoh' s statement because the tasks for visualization **O** and **H** usually result in **C** and **A**. Based on them, we suppose the combination of the following two visualization techniques satisfies the statements presented by Itoh:

1. Summarization visualization of network data
2. Comparative visualization of multiple networks

Summarization visualization is usually used for **O** and **C** while Comparative visualization works for **C** and **H**.

1.1.3 Summarization Visualization of Network Data

Nodes and edges are the two targets of network summarization. The typical summarization techniques are for nodes of the categorical network, which has been

long years since they were studied [51] [101]. Meanwhile, the summarization of edges has also been studied in the last two decades [74]. The edge summarization techniques are often called as edge bundling. Most of edge bundling techniques are based on the directionality of edges.

We present visualization techniques of a categorical network summarizing a large number of edges by its categories. Suppose that we summarize categorical network and analyze elements (nodes) related to a particular category. In such a case, we expect the elements placed close to the focused element have similar features to the focused one. We define these four requirements, and propose visualization techniques to satisfy them.

1. Summarize elements with the same category
2. Place the elements with common categories more closer
3. Place the elements related to each other more closer
4. Summarize a large amount of relationships

1.1.4 Comparative Visualization of Network Data

There have been many studies on comparative techniques of multiple networks whose elements are common. They usually use temporal network datasets. On the other hand, there have been few studies on the comparison between networks whose types of components (nodes or edges) are different. Such networks are defined as follows: The i-th network consists of the set of m nodes $n_0, \cdots, n_{m-1}$ and the set of k edges $l_0, \cdots, l_{k-1}$. The sets of nodes in the networks are different, however, they have any relationships and construct n-partite graph structure of these sets. The number n is the number of networks.

Suppose that we visualize the related parts of multiple networks focusing on a particular category when types of elements are different. There are tasks to see relationships between categorical information of network and compare correspondence

relationships. We define the requirements for comparison visualization and propose a visualization technique to satisfy them.

1. Summarize different types of elements in multiple networks based on common categories
2. Show common relationships between networks focusing on a particular category
3. Show different relationships between networks focusing on a particular category
4. Filter related elements or relationships between networks

1.1.5 Background of Visual Analytics for Domain Data

Needs for Gene Network Visualization

Various kinds of gene information have been revealed, however, there have been still unknown generic functions and relationships. To clarify this, many visualization of gene information have been studied for long years [47] [92]. Also, the following frameworks are well used to share information on gene functions and interactions.

Gene network A data structure that a node denotes a gene and an edge denotes a relationship between two genes

***GO* (Gene Ontology) [40]** Comprehensive, computational model of biological systems from information on the functions of genes and the project to build it. *GO* assigns one or more terms (*GO* terms) to a gene whose functions are already clarified.

However, the scale of a gene network is often very large, and as shown in Figure 1.1, the relationships usually generates visual cluttering so that they are hard to understand while visualizing it in a simple way.

We redefine the requirements for visualization of a gene network as follows, and present a visualization satisfying them.

1. Summarize genes with the same gene functions

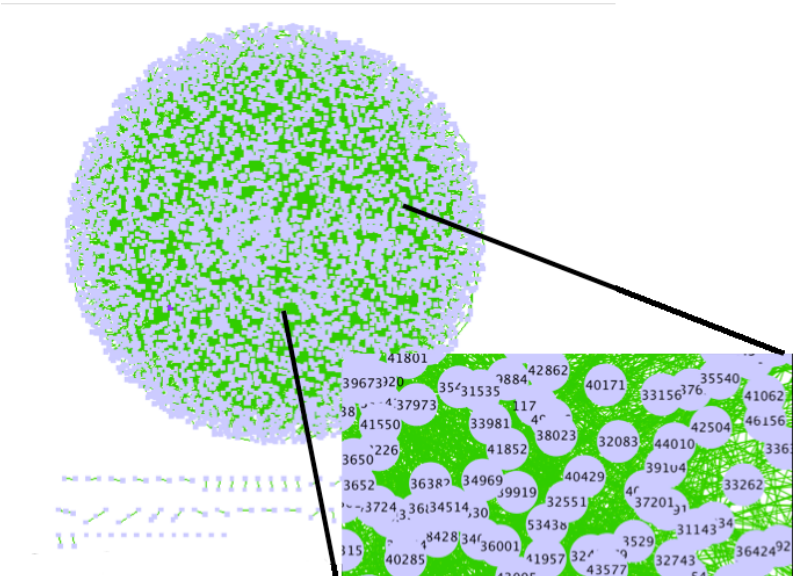

Figure 1.1: An example of overlapping gene networks.

2. Place genes with common gene functions more closer

3. Place genes which have mutual interactions more closer

4. Summarize gene interactions

Needs for Citation Network Visualization

Then, we describe the importance of citation visualization as the second application example. The survey of research papers is a very important task to find related work and grasp research trends when we study a research field. We usually survey research papers by using search engines such as Google Scholar [28] and ACM Digital Library [27], or investigating references of the papers we already read. However, it is difficult for novice researchers like under-grad students to grasp the focus paper' s position from the active search results as shown in Figure 1.2 or Figure 1.3. Moreover, novice researchers often miss related papers when they are not familiar with appropriate keywords or when the contents of papers they are looking for cover multiple research fields.

As we will describe in Chapter 2, many researchers have studied visualization techniques of citation relationships. The number of research papers in the world is still increasing and a large number of new research fields have been created. Since

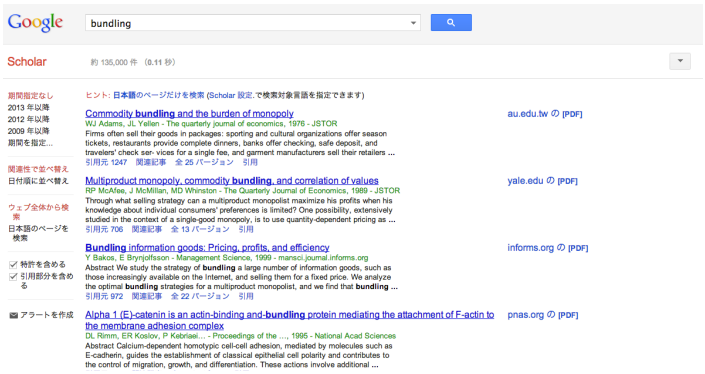

Figure 1.2: An example of search result by Google Scholar.

Figure 1.3: An example of search result by ACM Digital Library.

multidisciplinary research has been studied more and more, it is important to grasp and organize citation relationships across multiple research fields. Here, we estimate many users expect that papers including the similar contents of the focused paper are placed close to the focused paper while surveying papers on a research topic. We apply this to the four requirements in Chapter 4.

1. Summarize papers with the same topic
2. Place papers with common topics
3. Place papers with a citation relationship
4. Summarize citation relationships between topics

1.2 Contribution of This Research

We propose visualization techniques for summarization and comparison of categorical networks. The techniques include two cores. The one is applying the edge bundling between the clusters of the categorized nodes to a hybrid force-directed and space-filling layout. The second is switching bundles and edges based on the number of edges between the node clusters and the ratio of nodes connected to a bundle in two clusters interactively, while the existing bundling techniques usually require tuning parameters for the force of bundling edges. The techniques for summarizing a categorical network place the nodes applying a hybrid force-directed and space-filling algorithm based on both the edges and categories of the nodes and bundle the edges based on categories of the nodes. As for the comparison of multiple categorical networks, our technique places the nodes by reusing the positions of the node clusters which consist of the common combinations of the categories in the networks. To avoid visual clutters between the networks, it represents the relationships between the networks by filtering of the nodes in views.

We first explain the state of the arts about visualization techniques of a categorized network in Chapter 2. Chapter 3 describes a visualization of a gene network as an example of application to undirected network data. The next chapter 4 introduces a visualization of a citation network as an example of application to directed network data. Then, we present a comparative visualization of multiple categorical networks and introduces the visualization of citation and co-author networks at the same time as the example in Chapter 5. The contributions of this dissertation include implementation and overview, filter, and comparison operations for sophisticated visualization of categorical network data.

Chapter 2

Related Work

This chapter introduces related work regarding visualization techniques for network data in Section 2.1 and visual analytics using network data in Section 2.2.

2.1 Visualization Techniques for Network Data

This section summarizes the visualization techniques of a categorized network. There have been studied visualization techniques of networks for long years. The techniques could be roughly categorized into two types, summarization of networks and operations (or interactions) against networks for comparison. The targets of summarization are nodes and edges. Summarization of nodes generates groups based on categories or attributes of the nodes.

2.1.1 Network Layout of Categorized Network

Visualization techniques of a categorized network have been studied as shown in Hadlak's survey paper [69] from many perspectives. Some techniques represent categories as the representative node in an overview and show the individual nodes of the groups when users zoom in the view. Others, on the other hand, place the nodes of the same categories closer and the nodes of the different categories faraway. Itoh proposed a network visualization technique that expresses one or more attributes assigned to the nodes by colors, and places the nodes which have common attributes more closer [80]. This technique visualizes relationships between the network structure and the attributes of its nodes. Graphicle [95] tackles a trade-off between the network

structure and the attributes of nodes. It switches a mode to visualize the attributes by using color or size while placing the nodes in grids, circles or scatterplots or a mode to visualize the network structure by a force-directed algorithm. Bubble Sets [51] draws set groups using continuous and concaves iso-contours with minimizing cluster overlap. BranchingSets [104] is a visualization technique inspired by Kelp Diagrams [84] to assign hierarchical category information by colors of nodes and edges. In a case of multiple categories, the technique draws nodes with multiple lines side by side, one for each color, to represent the groups of the nodes. These visualization techniques decrease visual clutters caused by crossing of edges and overlapping of nodes, however, most of the techniques do not handle visual clutters caused by overlapping nodes and edges.

2.1.2 Edge Bundling Techniques

Edge is another element of a network, and summarization technique of edges is usually called edge bundling. It has become popular since Holten's technique [74] which covers hierarchical networks presented. The techniques give points of edges for the generation of splines, presume that Coulomb force and spring force attract edges at each point, and calculate the control points based on the forces to bundle edges [74] [75]. Muelder's technique [101] based on Treemap inherits the same idea. However, bundles generated by these techniques sometimes pass through over the nodes and it could decrease the readability of the network structure. Other techniques optimize the calculation of bundles based on energy or force-directed models [131] [75]. Ersoy [65] repeats a process that clusters edges, generates a framework of a network based on the clusters and attracts edges using display areas split by the framework. Selassie proposes a technique to bundle edges related the network structure by fixing the force-directed model [116] so that it repulses edges based on directionality of edges. These techniques have a problem of long computation time. Gansner's technique [68] is based on geographical information and fixes the position of each node by latitude and longitude. It needs to decrease the travel distance from the positions of the original edges in a case of generating bundles and long computation time to generate bundles that satisfy this constraint. Lambert' s technique creates a grid graph based on the positions of the

nodes and computes the shortest routes to bundle edges [86]. The existing bundling techniques visualize the main routes of a bunch of edges effectively, however, they do not consider to avoid overlapping nodes and bundles. An edge compression technique which replaces individual edges with edges connected to groups of nodes [61] addresses this problem. Many researchers reported the applications of edge summarization techniques. Zhou et al. categorized edge bundling techniques and introduced the visualization applications in their survey [132]. Visual analysis of categorical nodes and edges of social networks proposed by Crnovsanin et al. uses edge bundling technique [52]. Eiffel [79] is a visualization technique to support understanding of the evolutionary influence of a particular root node using an edge bundling technique. We tackle the problem of overlapping nodes and bundles while we keep the relationships between the categories of the nodes.

2.1.3 Visualization of Multilayer Networks

The visualization techniques of a simple network have been studied for a long time. In recent years, such networks are becoming more complicated. The real-world data are often expressed as sets of networks consisting of multiple different types of nodes or edges. Biological pathways and social networks are the examples [113] [107]. Some researchers defined the following networks as multilayer networks [85] [98]:

1. N-partite graph
2. Time information may also be considered as a layer
3. Multiple independent types of relationships between entities (nodes)

The visualization techniques of such complex networks have been focused in recent years. The existing related network models are also considered as multilayer networks in Figure 2.1. One of the models is an n-partite graph. All edges of the n-partite graph are between layers and there are no edges in the same layer. A multivariate graph is a network model whose nodes include multiple attributes. That is, node attributes divide the network into layers. Another model is a dynamic graph. It is also considered as

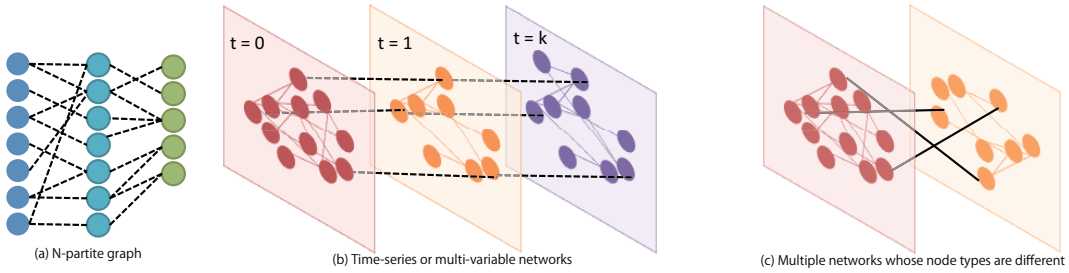

Figure 2.1: Types of Multilayer Networks.

multilayer network which consists of layers by time slices [81] [33]. The following tasks described in a paper [98] are considered in the visualization of such multilayer networks:

1. Cross-layer connectivity
2. Cross-layer entity comparison
3. Inter-layer comparison
4. Layer manipulation

To support these tasks, the number of researchers studying about visualization of multilayer networks is increasing. Detangler [111] provides an interaction technique to support multilayer network exploration using a multilayer metric. Like MuxViz [57], some systems for visual analytics of multi-layer networks are available to the public. Still, there is a design space to improve interactions including a comparison of different layers. The major interactions with networks are as follows:

1. Extraction of sub-graphs
2. Comparison of multiple networks

Extraction of sub-graphs is utilized for graph analysis such as community detection. Atlas [31] decomposes a large network into layers of sub-graphs using edge decomposition. The technique sorts these layers based on peel vertex score and visualizes them with representative metrics of graph analytics as an overview of a network. When a user selects a layer of a network, the technique visualizes a sub-graph in the selected layer by a node-link diagram in a different view. The users can

compare the layers with multiple views. Moreover, pivoting is also an operation against networks that a user hop from a root node to the connected target nodes. Jacob' s Ladder [38] repeats pivot of sets of nodes and filters the layers for sub-graph extraction.

We regard a set of citation and co-author relationships as a multilayer network and visualize them focusing on the interaction of comparison layers described in Chapter 5.

2.2 Visualization Analytics of Network Data

This section introduces visual analytics of network data, especially we handled in this thesis, gene networks, citation networks, and co-author networks.

2.2.1 Visual Analytics of Gene Network

There have been studied gene network visualization techniques. Cytoscape [118] is an open-source platform to visualize biological networks such as protein-protein and genetic interactions. It is widely used around the world and there are many plugins. BisoGenet [96] is one of the Cityscape plugins which build and visualize biological networks to represent multiple isoforms of a gene as results of alternative splicing or coding relations of two paralogous genes coding the same protein.

There are still challenges in visualization of gene networks in spite of such popular visualization system which can be extended. Nishiyama et al. [102] presented a technique to visualize network constructed based on gene expression values, which does not represent gene interactions or ontology. Breitkreutz et al. [42] represented gene functionality and network applying a variety of standard network visualization techniques, where it seems difficult to comprehensively represent scale-free network datasets. However, these techniques do not consider a process to avoid overlapping nodes and edges in the case of a large dataset. Dinkla et al. [56] visualized gene regulatory network by using matrix. While this technique can visualize sub-networks, it still has a problem that it consumes larger display space. The display space extends transversally or longitudinally in the case of a large network. Forbes et al. defined a model of protein-protein interactions as a directed network that nodes represent rules

of the model and edges represent influences between rules in a period [67]. Cruz et al. proposed an interactive tool with a timeline to represent the dynamic behaviors of molecular networks [53]. Itoh's hierarchical graph visualization technique [80] has been applied to visualize relationship and dependency between gene-gene interactions and gene expression values and discovered interesting features including an isolated subgroup of genes which satisfies specific conditions of gene expression values. However, it is not easy to explain the semantics of all features of the visualization results from the knowledge of gene-gene interactions and gene expression values.

2.2.2 Visual Analytics of Citation Network

This section introduces existing visualization techniques for paper citation networks. One of the goals of these techniques is understanding the trends of research topics. Citation network is one of the popular resources of data to help a survey of scholarly literature. Shahaf et al. [117] introduced their technique which visualizes the relationships between terms as a metric of survey papers so that the relations represent metro maps. In addition to citation relationships, some researchers utilized text information on scholarly literature. Citespace II [45] aims to visualize evolution and trend of research topics using hybrid networks of co-cited articles and terms citing the articles. Berger [37] also embeds citation relationships into text visualization. CiteRivers [72] visualizes the trend of the topics that papers include and the number of papers for each conference or journal by citation flow. Users can easily understand the trend of the topic they focus on and which conference or journal is related to it. These techniques support to understand the transition and trend of the research topics in a certain conference or a journal. However, they are only the visualization of topics and they do not support citation relations in the conference. Using only these techniques, we cannot find the related works that are not related to the user's focus topic directly but are cited by papers of the focus topics. Therefore, they are not for our purpose. Another goal is to help the literature review. Some researchers proposed visualization systems to help this task using citation and text information of literature [46] [109] [129].

There are some kinds of approaches to visualize citation network. One is the network topological approach and another is time-oriented one. One of these approaches to visualize citation patterns is to use the citation network topology [41] [45]. Brandes et al. [41] presents a visualization technique for citation networks with topographic maps that place the hub papers cited by many papers higher than the other papers. It also arranges the papers that have similar citation patterns closer. That enables us to easily find the hub papers and the groups of papers that have similar citation patterns. These topological techniques are easy to find well-cited papers and understand citation patterns. However, it is difficult to find papers by only citation topology when users do not know which paper they should use for a clue to track the citation relations.

Many researchers have also studied the time-ordered visualization technique for citation network other than topology-based ones [97] [112] [125] [127] . CitNetExplorer [127] applies a transitive reduction of a citation network and put them in chronological order. It assigns colors of nodes which denote publications to the attributes like successor and predecessor. Citeology [97] orders papers by the number of their citations for each year, and places them from the center of the display. It can visualize up to eight hops of the citations. This study represents structures of citation networks by placing nodes corresponding to papers in the time-series order. When a citation network has complicated relations across multiple research fields, it causes serious edge crossing and cluttering which bring bad impact on readability. Visualization results with heavy cluttering prevent the users from grasping the positions of papers, while the users want to understand the positions of the interested papers in the research fields. New papers always cite the older papers, so we think we do not need to use time-ordered visualization for citation network when we show the direction of the citations. CiteVis [125] visualizes citation relations by highlighting the citing nodes and the cited nodes when a user clicks the node. This technique reduces a visual clutter of edges and enables users to understand the citation relations of the clicked paper and the trend of the number of presented papers in the conference.

Our goal is to help novice researchers to survey papers that they want to read and

understand the positions of the papers in the research fields with search results instantly. To achieve this goal, we think both the citation relations and the topics of the papers are important as described in the previous section. Therefore, we propose a visualization technique that concerned both of the citation relations and the topics of the papers. Still, a few techniques have been considered both of them. Dunne et al. [60] proposed an integrated visualization of citation network and a paper summary description. Users can simultaneously look at the citations, ranking based on the citation count, and summary description of papers in the cluster generated by graph clustering based on citation structure. This representation has a bottleneck that it may require larger display spaces. Also, the network visualization shows only papers extracted by the keyword-based search. The users may miss papers if these papers do not contain the user-specified keywords or they are not cited by papers including the keywords. Though these novel visualization techniques have been presented, it is not still always easy to find important papers by using such techniques. One of the reasons is that these existing techniques often require users to manually specify the papers whose citations they want to figure out. It often happens that novice researchers do not know all the appropriate keywords, and therefore it is not easy for them to determine which papers they should read. The second reason is that it may happen to miss the papers which do not have citation relations but have similar contents when we only focus on citation relations to find papers. The third reason is that many recent new research fields have triggered as fusions of multiple research fields. Researchers need to organizationally understand the relations of papers that cover such multiple fields along with their fusion. However, few visualization techniques address this problem.

Visual Analytics of Co-author Network

Co-author relationship is also a popular resource to help a survey of scholarly literature or finding reviewers or collaborators and it has been studied for long years. Co-author network and citation network have been often analyzed individually [83]. Henry et al. [73] also analyzed publications and authors of four major HCI conferences in 20 years. In the analysis, they provided a visual exploration of the co-author network by

using matrix and node-link diagram. CiteWiz [64] provides interactive visualization system of influential researchers and publication impact. It cannot compare citation relationships and co-author relationships in an overview mode. Co-author network often becomes an example dataset that visualization techniques of network structure deal with. GraphPrism [82] used co-author network to show its ability to help the understanding of network structure by providing node-link diagram and heatmaps of graph metrics for nodes at the same time. Juniper [103] is an interactive table-based visualization system of topology and attributes of sub-networks. It is applied to co-relationships between characters in a novel of Game of Thrones and co-author relationships of ACM CHI conference and IEEE Transactions on Visualization and Computer Graphics. Its limitation is the size of sub-networks. As these studies have shown, few studies develop combinational visualizations of citation and co-author networks.

Visual Analytics of Other Scholarly Literature Data

In this section, we introduce the existing techniques of visualizing scholarly literature. One of the visualization targets of the existing works is the topics of text corpus [55] [120] [124]. TIARA [93] expresses patterns of topics and keywords in time series using a stack graph. PaperLens [90] is a visualization technique that applies the mixture distribution model to the titles and keywords, then estimates their topics, and finally shows papers by topics and publication years. Lee et al. [91] analyze visually trends of papers published in ACM CHI, one of the most famous international conference in the area of human-computer interaction. These techniques show that understanding the trends of research topics is a very important task.

Another target is the relationship between topics and researchers. PivotPaths [58] is a visual interface for searching for faceted information resources. It visualizes relationships between tripartite information spaces, people, resources, and concepts. This technique does not visualize relationships between the items in the same information space. Moreover, it requires users to input any keyword for focusing on a particular item. Suppose that we are not familiar with a research field to investigate.

It is probably difficult for us to cover the research field if we only use topic keywords, a citation relationship or a co-author relationship. Cartolabe [43] is a similar project which also visualizes relationships between researchers, documents, and topics. This system places both researchers and documents in the same display space based on topic similarities. These techniques help to find similar researchers to find documents or researchers in a particular topic based on the words that they use. On the other hand, they are not suitable to understand relationships between topics because they do not provide the other relationships like citation and co-author networks.

In Chapter 5, we combine text information of papers, citation relationships, and co-author relationships.

2.3 Conclusion

This chapter introduces related works regarding visualization techniques for summarization of categorical nodes, edges, and comparison multiplex networks. We also describe visual analytics for each domain data, gene network, citation network, and co-author network.

Chapter 3

Integrated Visualization of Gene Network and Ontology Applying a Hierarchical Graph Visualization Technique

3.1 Introduction

This chapter presents an integrated visualization technique of gene ontology and interactions. An overview of the technique is explained in Section 3.2. Section 3.3 describes its implementation in detail. Section 3.4 contains examples of Drosophila genes, and summarized in Section 3.6.

Recent bioinformatics techniques have realized the mapping of genetic information; however, we still have many open problems of unexplained genetic functions and relationships. As we mentioned in Chapter 1, GO terms and gene network are important to explain genetic functions and relationships. There has been studied about visualization techniques of gene networks, however, they remain the below problems.

1. Huge number of nodes and edges in gene networks

2. Complexity of edge structure

As for the first problem, therefore visualization results with such networks are not often comprehensive. The more clear structure of gene network is important to solve this problem. The other problem causes overlapping of nodes and edges in a display

space, and this decreases the visibility of gene interactions and relationships between gene functions. The goals of our visualization technique are as follows:

G1: Find pairs of clusters that have a lot of gene-gene interactions when clustering genes by their functions

G2: Find outlier relationships between a small number of genes in a cluster

3.2 Overview

This section presents an overview of our technique for integrated visualization of gene ontology and interactions. Figure 3.1 is a visualization example of gene network applying GO terms. We used FRUITSNet [80] to place a gene network. The dataset we used is the same as the one described in Section 3.4. A node corresponding to a gene is colored by 10 types of GO terms selected randomly. An edge represents an interaction between genes. The main problems of this visualization result are as follows:

1. Limitation of the number of GO terms that can be visualized caused by the number of colors that users can distinguish

2. Visual clutters by thousands of edges and nodes

As for the first problem, although hundreds of GO terms are used in this dataset, this figure expressed only 10 or so GO terms. Therefore, we group GO terms whose functions are similar to a cluster (we call it GO cluster in this chapter) and color genes by GO cluster. Our technique clustering GO terms based on their DAG (Directed Acyclic Graph) structure to generate GO clusters. Some GO terms have one or more parental GO terms because of the DAG structure. The technique firstly generated clusters of GO terms according to the DAG structure which arranges the terms and assigns the term clusters to nodes corresponding to the genes. It corresponds information on gene functions defined by GO terms to colors of the nodes in a gene network and visualized gene functions and relationships at the same time. This would support efficient prediction of unknown gene functions or relationships. The reason to assign

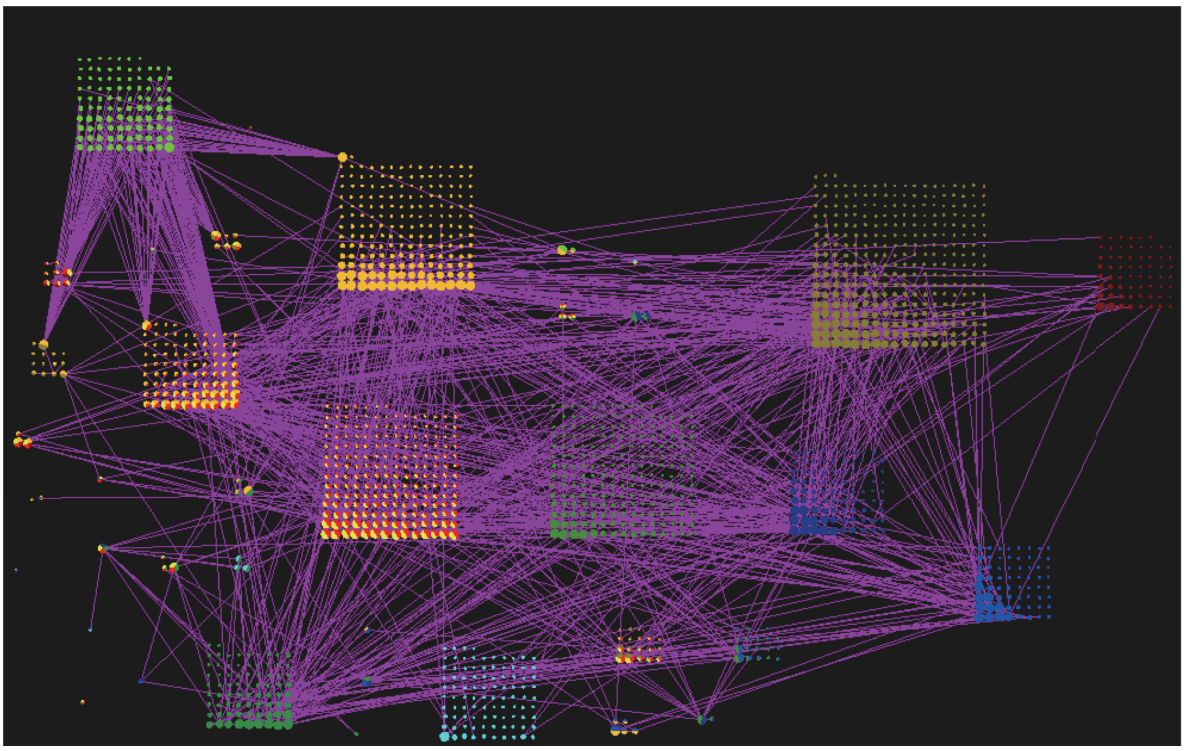

Figure 3.1 Visual clutter problems prevent grasp of relationships between genes and their functions.

GO terms to genes instead of gene expression levels is that experimental errors could have an impact on the levels because of their values from experiments.

In addition to that, assignment GO terms as attributes of nodes would work effectively for validation of the existing assignment of GO terms. We expected that it is fruitful and informative if we visualize the distribution of GO terms on the gene network, because we can discover new relationships between functions defined as GO terms and interactions represented as a network. We also expect this knowledge may assist in predictions of undiscovered genetic functions and relationships, which is an important process for independent and customized planning of experiments and analysis of genetic diseases. To address the second problem, we employ a hybrid force-directed and space-filling approach which is used in FRUITSNet [80] and apply edge bundling. Our technique generates clusters of genes (we call them gene clusters in this chapter) according to the GO clusters and calculates the positions of gene clusters. It then divides the nodes according to the commonality of the term clusters, and divides again according to their

connectivity. It calculates the positions of node clusters applying a hierarchical graph visualization technique, featuring hybrid force-directed and space-filling approach. The force-directed algorithm attempts to minimize total length and interactions of edges, and the space-filling algorithm it attempts to maximize the space utilization.

After that, the technique bundle the edges between two gene clusters when the number of edges between these two clusters is larger than a threshold. Each bundle is drawn with polylines so that it avoids overlapping nodes. This process reduces the overlaps of nodes and edges in display space and improves the readability of relationships between gene functions and interactions. This technique bundles edges connecting pairs of nodes belonging to the common node clusters and draws the bundles as thicker polygonal lines. This can express the whole structure of a gene network with gene functions clearly. The target of our technique in this chapter is gene network with non-directional edges.

3.3 Implementation

This section describes the implementation details of the presented technique.

3.3.1 Gene Ontology (GO) and Gene Clustering

A set of GO terms are usually used as a DAG. Our implementation marks GO terms used in the set of given genes, and clusters the terms according to the topological distances on the DAG structure. We generate clusters of GO terms in the following steps. The technique firstly selects one of the GO terms and extracts all of its children GO terms. We define such a GO cluster that the number of assigned genes to the clusters is larger than a threshold as a cluster.

We then extract 10 to 15 GO clusters from the result according to specific conditions, such as rougher clustering results, or numbers of genes which terms in a cluster are assigned. We think 10 to 15 is a good number to visually distinguish the terms by the colors of nodes. The implementation applies node clustering as a preprocessing of the graph visualization. The fast modularity community structure inference algorithm

firstly divides nodes to generate groups of nodes which completely the same sets of term clusters are assigned [48]. It then divides the nodes in each cluster according to the density of edges. The current implementation firstly refers to GO clusters assigned to each gene and labels each gene with the clusters. It then categorizes each gene into a combination of GO clusters. We define a group of genes whose combination of GO clusters are completely the same, as a single cluster of genes. After that, the technique groups genes in each cluster into multiple clusters based on the density of edges. We employ this algorithm as the clustering technique based on edge density.

3.3.2 Graph Layout

After the clustering process, the implementation applies two steps of data layout: force-directed and space-filling layout steps based on FRUITSNet [80]. It firstly generates a network structure corresponding gene clusters to its node. Iterative computation of the nodes using force-directed model decides the tentative position of gene clusters. This layout places gene clusters connected with edges closer and gene clusters including common GO clusters closer in a display space. During the space-filling layout step, the current implementation preserves pre-defined distance of blanks between adjacent clusters, so that bundled edges can pass through between them. Applying the space-filling layout with reference to this tentative placement, our technique attempts to maximize space utilization while it avoids overlapping gene clusters.

3.3.3 Edge Bundling

The technique applies edge bundling after placing all clusters and nodes for readability of edges. To visualize gene network, the geometric structure of node placement is not important but only topological information on gene relationships. We focus on this and employ more simple and fast edge bundling technique rather than the existing ones. As for the processing flow of edge bundling, the technique selects one of the gene clusters and counts the numbers of edges connecting two nodes belonging to arbitrary pairs of clusters. It bundles the edges and draws as thicker polygonal lines if the number exceeds the pre-defined threshold. This representation emphasizes major relationships

between pairs of genes which common sets of term clusters are assigned. We repeat this process against all pairs of gene clusters. Here, we adjust the width of a bundle W by a function of the number of edges between clusters $k(1)$. In our implementation, we defined $a = 0.7,\ \ b = 2.5$, and the number of pixels as W.

The implementation is somewhat similar to Cerm et al. [44] which avoids passing through rectangular nodes. The implementation firstly extracts a pair of node clusters that edges connecting two nodes belonging to the pair of the clusters (Fig.3(a)). It then generates a segment which connects the centers of the rectangles corresponding to the pair of node clusters, drawn as a thick dotted line in Figure 3.2(a). Let the numbers of nodes in each of the clusters which are connected by edges to be bundled as m_1 and m_2, and positions of nodes as $\{p_{11}, ..., p_{1j}, ..., p_{1m_1}\}$ and $\{p_{21}, ..., p2k, ..., p2m_2\}$. Also, let the positions of centers of rectangles corresponding to the clusters as r_1 and r_2. Here, the technique calculates the endpoints of the segment e_1 and e_2 which connects the rectangles as follows:

$$e_1 = (1-t) * r_1 + t * r_2$$
$$e_2 = t * r_1 + (1-t) * r_2 \qquad (0 < t < 0.5)$$

The implementation then detects the collision between the segment e_1e_2 and other rectangles. If the segment intersects with other rectangles, the implementation folds the bundle segment to avoid the intersections, as shown in Figure 3.2. Finally, it draws thinner segments $p_{1j}e_1$ and $p_{2k}e_2$, and a thicker polygonal line connecting e_1 and e_2. Applying this process recursively generates polylines that go around gene clusters as shown in Figure 3.2(b). This process avoids overlapping bundles and clusters of the nodes. However, users could miss bundles visually in a case that bundles follow the same route along a ridgeline of a rectangle corresponding to a gene cluster with applying only this process. Therefore, when overlapping bundles, we address this problem by moving the positions of bundles as shown in Figure 3.2(c). Bundles still overlap where the display space is packed with gene clusters. In such a case, we draw a single polyline between two gene clusters considering crossing the bundle and gene clusters is

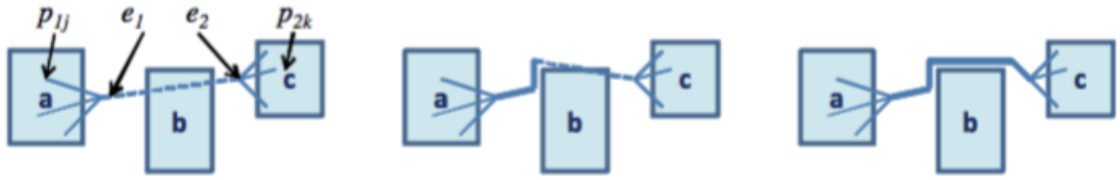

Figure 3.2 Edge bundling process that avoids rectangles corresponding to node clusters. (a) Generation of line segments (b) Generation of polygonal lines. (c) Avoid overlapping.

unavoidable. Users can interactively control the change of edge bundling mode and the threshold for the number of edges k with the components of GUI described in detail in the next section.

3.3.4 User Interface

This system has one main drawing space and an interactive panel in a window, and there are some types of graphical user interface components on the interactive panel. The system can zoom, pan and shift the main view. The panel consists of buttons for selection of GO clusters to highlight the nodes and a button for edge bundling. The threshold for bundling edges has a big impact on a visualization result. To compare visualization results with different thresholds, users can set the minimum number of edges for applying edge bundling and change the opacity of unbundled edges to draw discreetly with range sliders. When users click a node, the panel shows the identifiers of the gene corresponding to the node and the detailed information on GO clusters assigned to the gene. The drawing space provides the detailed relationships of the clicked node by highlighting edges that are connected with the node at once. Highlighting edges is applicable for two nodes at the same time, and this enables users to compare the relationships of both of the two nodes.

3.4 Experiments

This section shows and discusses the results. We prepared a Drosophila gene network dataset provided by iRefIndex [110], containing 8,945 nodes, 32,703 edges, and 259 non-clustered GO terms. iRefIndex provides gene interactions of the same pair of genes reported in each paper and experiment respectively. Therefore, we treated them as a single edge removing the duplications. Also, we used protein information provided in NCBI Entrez Gene database [30] as nodes. We used only the edges that both of the nodes corresponding to protein information in Entrez Gene because all of the nodes provided by iRefIndex do not correspond to those of Entrez Gene. The number of clusters of GO terms assigned to the nodes is 12. The presented technique is implemented with JDK (Java Development Kit) 1.6 and executed on a personal computer (CPU 2.7GHz Dual Core, RAM 8.0GB) with Windows 7(64bit). Figure 3.3(A) shows an example of a visualization result before applying edge bundling. We compared the visualization result after edge bundling shown in Figure 3.3(B). Interactions between nodes in the upper left cluster indicated as (a) and nodes in the other clusters are well-represented in this example. However, it is difficult to follow major interactions between generic functions from this example, because of the high density of the edges. On the contrary, it is much easier to follow major interactions between commonly featured genes by looking at Figure 3.3(B). Here, rectangular regions indicated as (a) to (e) in Figure 3.3(B) are node clusters which nodes are annotated by the following term clusters:

(a) protein-DNA complex (GO: 0032993)

(b) intracellular organelle part (GO: 0044446), non membrane bounded organelle (GO: 0043228), organelle part (GO: 0044422)

(c) non membrane bounded organelle (GO: 0043228)

(d) cell projection (GO: 0042995)

(e) intracellular organelle part (GO: 0044446), organelle part (GO: 0044422)

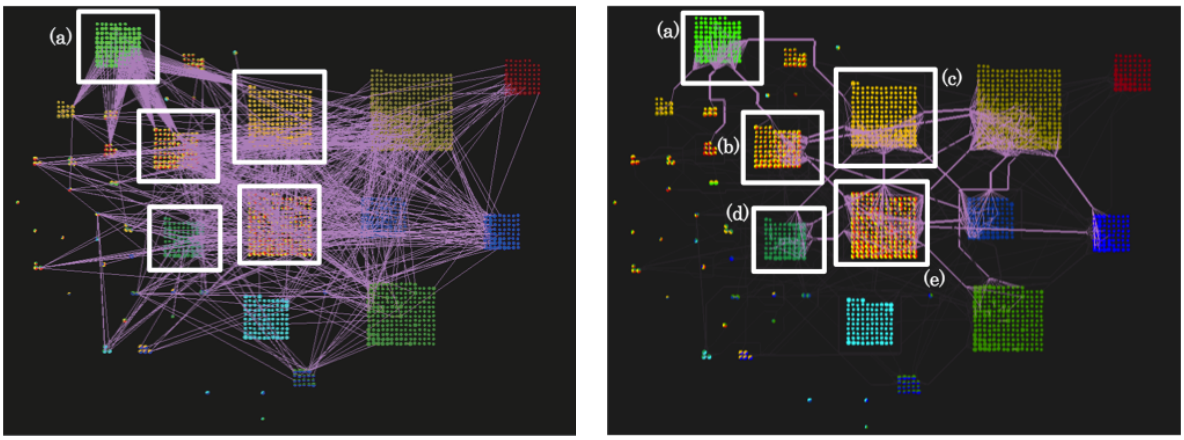

Figure 3.3 Visualization results of a Drosophila gene network. (A) Before edge bundling. (B) After edge bundling.

Term clusters are described as above by the terms and IDs of the cluster which are placed at a relatively higher level on the DAG structures. After applying edge bundling, we found that there were many edges between pairs of clusters (a) and (b), (a) and (c), (b) and (e), (d) and (e).

Next, we focused on the relationships between two clusters **A** and **B** shown in Figure 3.4(Left). Here, Figure 3.4 shows the result before edge bundling, and Figure 3.4(Right) shows the result after edge bundling. We found that Figure 3.4(Left) shows two groups of relationships between two clusters **AB1** and **AB2**. One yellow node at the upper left of the cluster **B** connects with several light green nodes of the cluster **A**, whereas it is hard to follow how other nodes of the cluster **B** connect to the nodes of other clusters while looking at Figure 3.4(Left). We also observed the same interaction in Figure 3.4 (Right) as well. In contrast to Figure 3.4(Left), these two relations are brought together, however, we could follow that the cluster **B** had a few-to-many relationship with the cluster **A**. According to the one-to-many relationship by the result before edge bundling, we could estimate the relations between other yellow nodes are also one-to-many. We clicked one of these yellow nodes (b2) and got a visualization result in Figure 3.5. The node (b1) at the upper left end of the cluster **B** does not connect with the other clusters while other nodes (b2) in the cluster **B** connected with. From this result, we can assume that the different GO terms which the nodes (b2) do not have, or other GO terms influence the common relationship. We compared the proposed

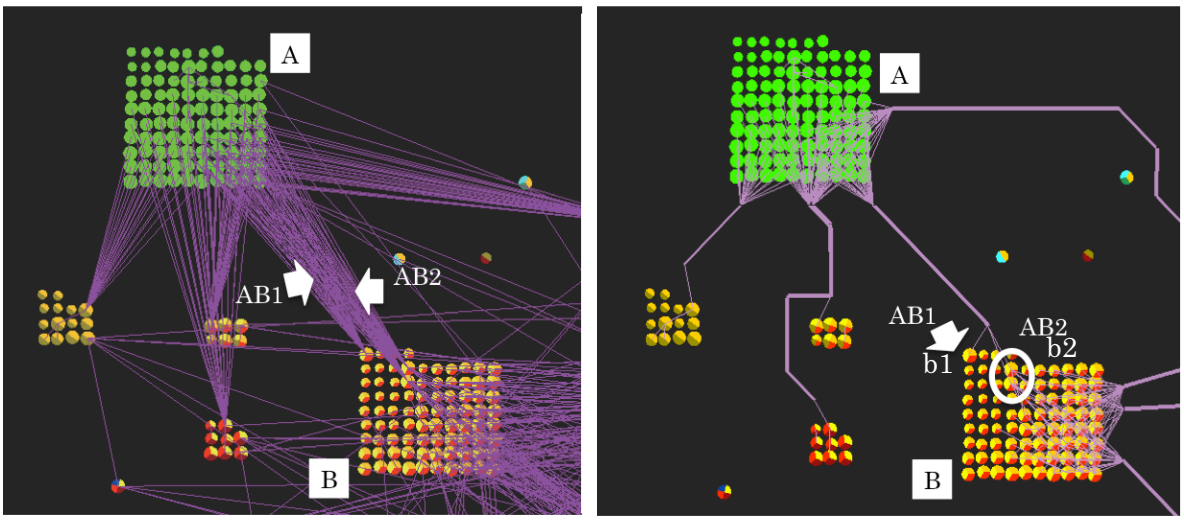

Figure 3.4 Close up views. (Left) Before edge bundling. (Right) After edge bundling.

technique with the existing techniques. As we mentioned in Section 2.2.1, many of existing techniques do not overlap nodes and bundles of edges. In such a case, it would be difficult to grasp relationships between nodes whose bundles are overlapping. Our technique avoids overlapping bundles and cluster of the nodes. On the other hand, we found some issues with these results. The issue is caused by drawing bundles as polylines to avoid overlapping bundles and nodes.

In Figure 3.4(Right), the upper left cluster **A** colored in green has edges between a cluster consisting of 6 nodes that are placed on the left of the cluster **AB1**, a cluster consisting of 8 nodes that are placed on the left of the cluster **B**. However, both of the bundles are overlapping each other and it could lead users to miss the existence of two bundles.

Moreover, we examine whether another result that the technique is applied to in Figure 3.6 corresponds to biological knowledge. There are overlapping nodes and edges in Figure 3.6(Left). On the other hand, Figure 3.6(Right) reveals relationships between nodes inside (c) and the role of (a) as a hub for other functionalities. The gene functionalities assigned to (a)-(h) are as follows:

(a) Regulation of biological quality

(b) Determination of adult life span

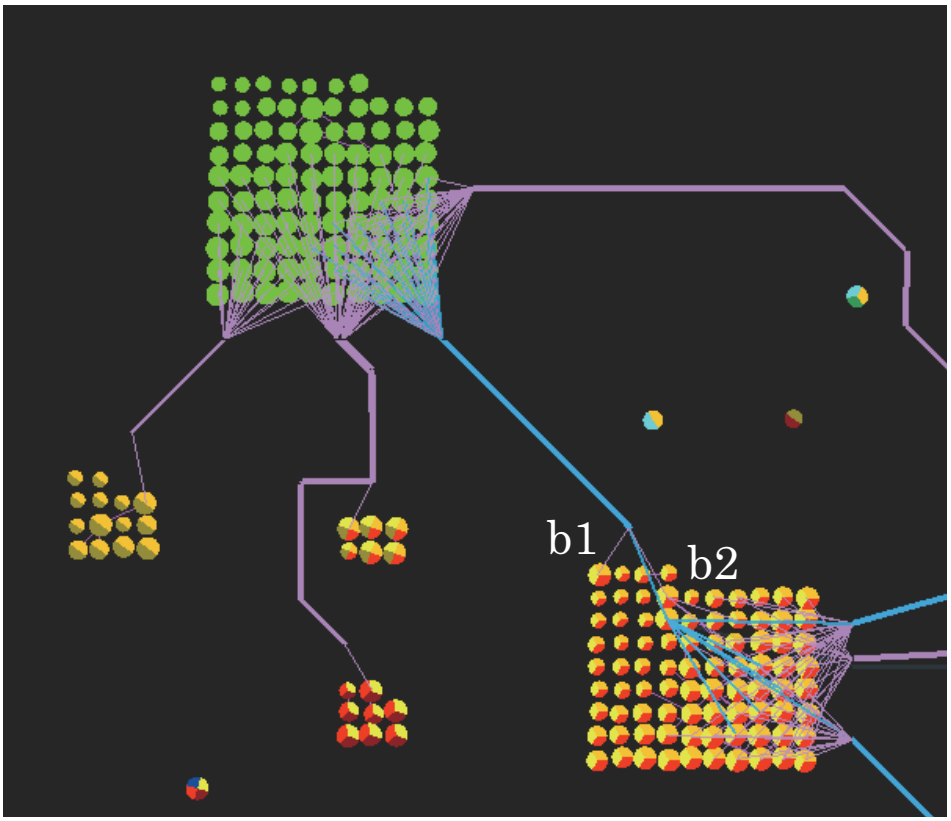

Figure 3.5: A visualization result when clicking a node.

(c) Cell cycle process

(d) Embryonic development

(e) Cell proliferation

(f) Cell division

(g) Cell adhesion

(h) Cell death

The visualization result that (a) works as a hub in Figure 3.6 corresponds to the insight that the gene function (a) works as a regulator. The genes which belong to (a) regulate something against the genes which belong to (b)-(h). Therefore, by applying edge bundling, the technique allowed us to find interesting nodes which seem to relate to the new gene functions and relationships. It also grasped the connectivities of nodes in the same gene function cluster which are hard to read without edge bundling. Applying

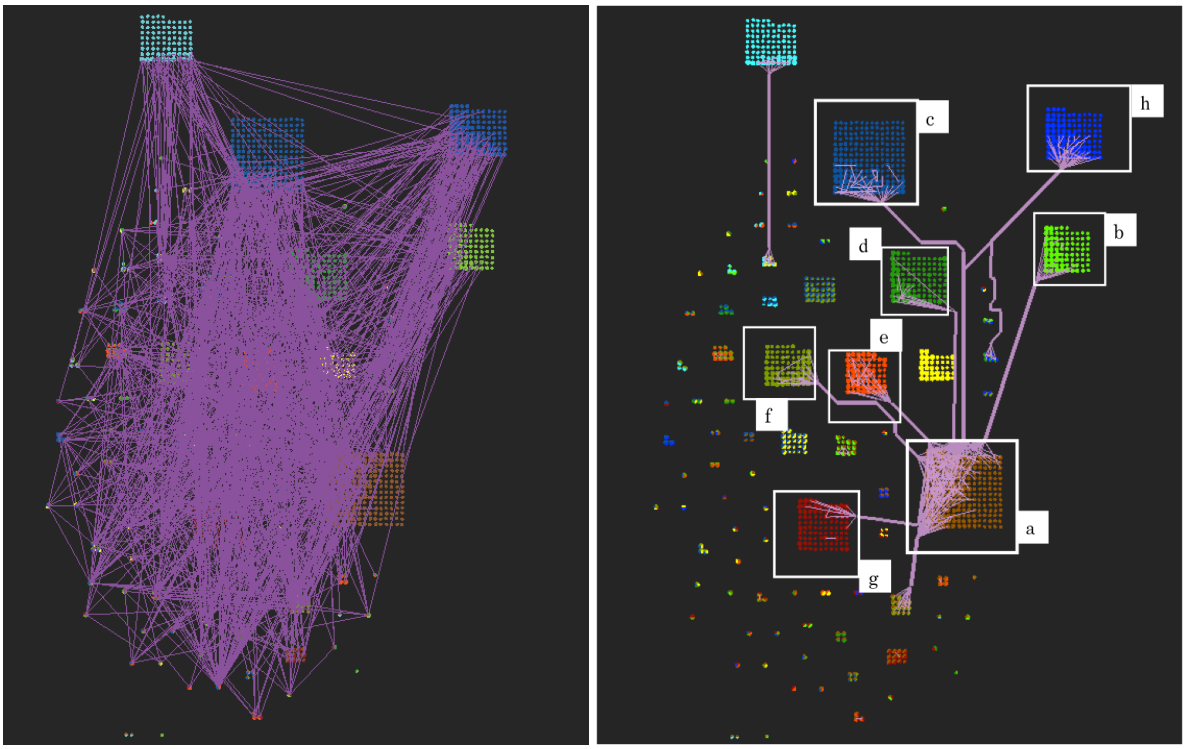

Figure 3.6 Visualization results corresponding to biological knowledge: (Left) Before edge bundling. (Right) After edge bundling.

edge bundling technique improves visibility of gene networks would help to grasp the relationships between genes and assume gene functionalities effectively.

3.5 Discussion

The visualization results showed that the technique supports an user to grasp the relationships between gene functions easily. Moreover, these relationships correspond to the biological insights into the gene functions which are already revealed. This correspondence demonstrated that our technique works effectively in a case of gaining consensus from users quickly by visualizing a number of insights on a single display. We verified visualization results reflect summarized insights on gene functions by assignment of GO terms at the upper levels to the nodes. The results present several interesting structural features of the gene network (G1). When zooming into the relationships between a pair of the clusters, we found an outlier node with different relationships between another cluster (G2).

The potential future issues include the following:

(1) Overlapping bundles

(2) Difficulties of identification of similar colors

(3) Trade-off between edge crossing and clusters of gene functions

(4) Relationships between genes which are lacking of gene functions

3.6 Conclusion

This chapter presented a visualization technique for gene network featuring gene interaction and gene ontology (GO). The technique firstly generates clusters of GO terms according to DAG structure of the terms, and annotate genes the clusters which terms of the genes belong to. It then clusters genes according to the GO clusters and connectivity and visualizes the clustered gene network applying hybrid force-directed and space-filling node layout. Force-directed algorithm places the nodes with common GO terms closer. The GO-based edge bundling alleviates the problems of readability that the number of edges is large and that there are many overlapping edges and nodes. These bundles are depicted as polygonal lines which avoid overlapping clusters of the nodes. This chapter also introduced the visualization results of a Drosophila gene network provided by iRefIndex and showed that the technique supports users to grasp the relationships between gene functions easily.

Considering the issues we discussed in the previous section, we would like to alleviate the issues of readability from (1) to (3).

(1) Improvement of the edge bundling process so that the technique can repulses the bundles completely overlapped on the drawing spaces

(2) Improvement of expressions for items (e.g., gene expressions defined as Gene Ontology in this section) with colors

(3) Increase of readability for the inside of clusters

In practical use of gene analysis, GO terms at the lower levels are more important when we observe the detailed differences between genes. Users would require a more interactive visualization system that depicts an overview of gene network with GO terms at the upper levels and then shows the detailed information on gene functions applying GO terms at the lower levels to a user-selected part of the network. It is also important to compare genes lacking of gene function information with genes whose functions are already revealed. The additional function which emphasizes relationships between genes lacking of expressions would be helpful for users.

Chapter 4

A Visualization of Citation Network Applying Topic-based Clustering

4.1 Introduction

This chapter presents a visualization technique for citation networks applying a topic-based paper clustering. Section 4.2 introduces an overview of the system. Section 4.3 describes the implementation of the system in detail. Section 4.4 contains use cases the results of some evaluations. Finally, Section 4.5 summarizes this chapter.

Finding research papers is a very important task to understand trends in research fields and find related papers. Researchers use text-based portal Web sites such as Google Scholar [28], ACM Digital Library [27], and IEEE Explore Digital Library [29]. Researchers look up for the references of papers they have read. However, it is not always easy for novice researchers to survey papers they want to read and instantly understand the positions of the papers in the research fields with their search results. Moreover, young researchers may miss papers in case they do not find the appropriate keywords, or in case that papers they really want to survey straddle multiple research fields. We define *keywords* in this section as the terms which consist of a topic, not terms which the authors annotate. A topic includes multiple keywords.

A citation network contains documents as nodes and citation relationships between documents as edges. When a document d_i cites another document d_j, the document d_i usually represents as a source node while the document d_j as a target node. There have been many studies on visualization of citation networks, including Mackinlay et al. [94]

and Small et al. [122], which aimed to alleviate these difficulties. However, we suppose still there are many open problems on visualization of citation networks. For example, researchers continuously trigger for new fusions of multiple fields, and therefore they need to organize and understand the relations of papers that cover multiple research fields. Another problem is while surveying papers in unfamiliar research fields. Papers in the unfamiliar research fields sometimes do not include the terms well-used in a research field which the users are familiar with. Conversely, terms may have very different meanings depending on the research field. In such cases, we find that the papers are not what we expected after we read them. Understanding the positions of the papers in the research fields is important for researchers to identify whether the papers are related to the topics which they want to survey.

To organize these open problems, we define the requirements in visualization of citation networks for a survey of papers as follows:

R1: Find much-cited papers which include user-specified topics.

R2: Find papers whose contents are similar to the focused papers, such as the papers which do not contain the user-specified keywords but belong to the user-interested topics.

R3: Find the contents of papers which have citation relations between the papers using user-specified keywords or belonging to user-interested topics.

R4: Find tightly related pairs of topics.

4.2 Overview

We propose a visualization technique that satisfies the requirements in the previous section, by implementing the following solutions.

S1: Categorize papers that have similar topics to the same group. (Section 4.3.1)

S2: Place papers that belong to the category in a circular region. (Section 4.3.2)

S3: Place citing and cited papers closer by applying a force-directed layout algorithm. (Section 4.3.2)

S4: Summarize citation relations by applying an edge bundling method. (Section 4.3.3)

Solutions **S1** and **S2** satisfy **R1** and **R2** because users can easily find papers belonging to user-interested topics by looking at the particular circular regions of the topics. Solution **S3** satisfies **R3** because users can follow the citation relations between closely placed papers. Solution **S4** satisfies **R4** because bundled edges effectively represent tight relations between pairs of topics.

This technique applies a general purpose graph layout technique, against many studies on visualization of citation relations apply time-oriented visualization design. We aimed to represent the positions of user-interested papers in a set of topics, and relations between pairs of topics. Therefore, the system applied our own algorithm based on clustered graph layout [4] [80], not time-oriented visualization. The thesis presents our subjective evaluation comparing with an existing time-ordered technique. The contributions to this section are as follows:

- Proposed a technique to visualize citation network based on their topics.
- Demonstrated that our technique could be helpful to grasp the positions of papers in research fields with the visualization results.
- Compared with the time-ordered visualization technique.

This technique would help novice researchers to understand the differences between the tendencies of similar research fields.

4.3 Implementation

This section describes the processing flow and the implementation of the presented technique. We treat the papers as nodes and citations as directed edges of a network. The technique classifies the papers based on their contents to construct a hierarchical network. The technique then applies our hierarchical network layout technique with an

edge bundling algorithm. The implementation also provides rendering and interaction techniques.

4.3.1 Clustering Papers

The proposed technique applies LDA (Latent Dirichlet Allocation) [39] to categorize papers based on the contents of papers. LDA is a generative topic model which allows a document to include various topics. It is generally used because it can avoid overfitting the data. As a paper can include multiple topics, we think LDA is appropriate for our purpose. It could solve the problem to categorize papers that straddle multiple research fields. The technique applies LDA to the set of all the paper abstracts to estimate topics and calculate the topic distribution for each abstract. LDA needs to be given the number of topics, so we determine the number heuristically. We regard these topics as research fields and categorize all papers based on them. The technique supposes a paper is related to the particular topic if a value of the topic distribution is larger than the threshold. We removed unnecessary words from the abstract as a preprocessing to improve the quality of clustering results. The removed words included non-important words such as prepositions, or too frequently used terms such as *propose* and *technique*. Then, we presumed the contents of the topic from 20 words whose probability is highest on the topic. The reason we do not use keywords assigned in papers because the words vary widely and it is difficult to categorize papers by only the keywords. Our clustering allows a paper to belong to multiple topics. A cluster in this paper can include multiple topics. For example, there is a cluster which including only topic A whereas papers about topic A and topic B belong to another cluster.

4.3.2 Network layout

Next, our technique arranges nodes applying Itoh's hybrid force-directed and space-filling graph drawing algorithm [80] to calculate the positions of nodes corresponding to the individual papers. We define a dataset of papers has a hierarchical structure of conferences and topics. In this paper, conferences are upper-level clusters and topic clusters are lower-level ones to show the trend of topics and citation relationships in

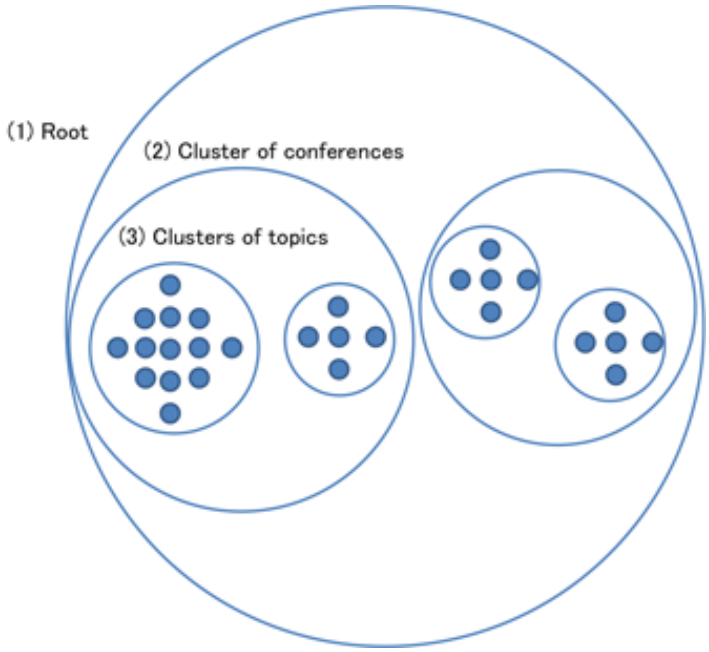

Figure 4.1 Space-filling hierarchy layout for our system. The algorithm calculates positions of nodes or clusters from (3) the lowest level clusters to (1) the top-level of the dataset.

conferences.

We calculate positions of papers under topic clusters shown as (3) in Figure 4.1, positions of topic clusters under conference clusters shown as (2) in Figure 4.1, and positions of conference clusters under the root node using Itoh's algorithm. The technique displays the nodes supposing that their sizes are proportional to the number of citations. The force-directed algorithm enables to place papers that belong to the same research category closely, and also, papers that have citation relations closely. Then, the space-filling algorithm enables to avoid the node cluttering and improve the display space utilization.

4.3.3 Edge Bundling

After the above process, the technique summarizes the edges corresponding to citations by applying an edge bundling algorithm. Our implementation of the edge bundling enables users to adjust the threshold controlling whether it bundles the edges or not. We have already implemented the edge bundling algorithm in our previous work [4]; however, it had a problem that straight bundles with which summarizes a lot of edges

may lead to misinterpretations of Gestalt principle as Figure 4.2 shows when the bundles avoid nodes and bend at a right angle.

To prevent the misconceptions, we place nodes in circular and bundle the citation edges with Catmull-Rom spline curves (Figure 4.3). Our technique firstly calculates the shapes of all bundle paths so that they do not overlap the node clusters. According to the threshold the user sets, the technique determines whether the number of the edges of one cluster with the others is larger than the threshold like Figure 4.3-(b). Then, it bundles the edges only when the number of edges between the two clusters is larger than the threshold. (Figure 4.3-(c)). The technique applies this process to all pairs of the node clusters (Figure 4.3-(d)). The width of a bundle represents the number of citations between two node clusters.

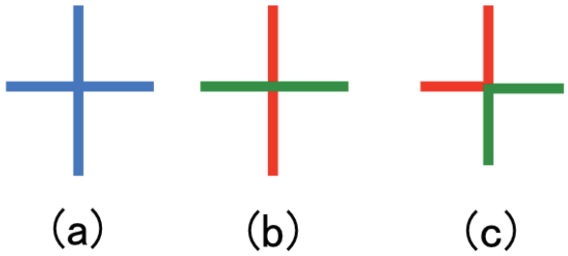

Figure 4.2 A misinterpretation of Gestalt principle: (a) The appearance of two bundles. (b) It usually looks like two bundles crossing. (c) The two bundles actually bent at a right angle.

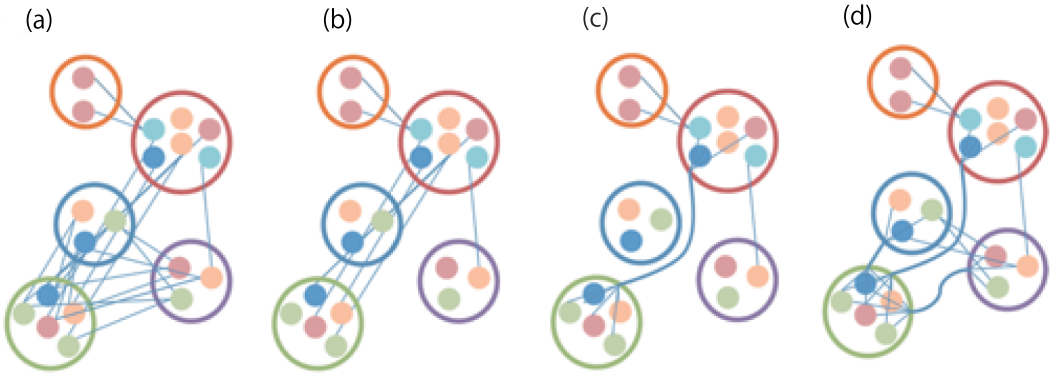

Figure 4.3 The processes of edge bundling: (a) Calculate the shapes of all bundle paths. (b) Count the number of edges between two clusters. (c) Bundle the edges only when the number of edges between two clusters is larger than the threshold. (d) Apply the processes (b)(c) to all pairs of the node clusters.

4.3.4 Color Scaling for Network Rendering

Since citation networks have directionality so-called *cited* and *citing*, our technique draws the cited side of the edges in bright pink, and the citing side of the edges in dark pink, to represent the directionality of the edges. When a user clicks a node without edge bundling technique, the system highlights the links of the clicked node with blue or green color. We chose the colors based on the following requirements, so the other colors are also appropriate if they satisfy them.

- Use the color except pink to compare the edges of the clicked node and the other edges
- Use two different colors to compare the nodes which are clicked at the first time and the next time

We can also draw arrows or assign different hues to each side of the edges for the representation of directionality of the edges. However, these representations are not always adequate for large-scale networks and networks in which there are many hubs. When we represent the direction by an arrow, heavy cluttering may happen around hub nodes or dense regions, which would degrade the readability. Besides, we assign hues to the nodes, and our technique controls brightness to represent directions of edges. The brightness "Dark-to-Light" is a better representation of directed-edges than an arrow except a tapered one as Holten et al. studied representations of directed-edge using combinations of shape and color [76]. We have assigned the width of an edge to the amount of relationships between two node clusters, and it is reasonable to employ brightness for a representation of direction of edges. As Figure 4.4 shows, we draw nodes with the color scale corresponding to the publication years.

Figure 4.4: Color scaling: (a) The node color, (b) The edge color.

4.3.5 User Interface

Figure 4.5 is a snapshot of the user interface we implemented. The left side of the window features the drawing space, while the right side features two tabs. One of the tabs features various GUI widgets. Users can scale and shift the view, switch the edge bundling mode, and set its threshold, by using the GUI widgets shown in Figure 4.5 (a)(c). When a user clicks a node corresponding to a particular paper, the technique displays the details of the paper such as the digital object identifier (DOI), title, authors, year, and abstract, on the panel featured by the other tab. At the same time, it highlights the edges of the clicked node and those of the nodes that are connected to the clicked node. This edge highlight function can be applied to two nodes together, and this enables to compare the citations of each paper.

By the way, it is not always easy for the novice researchers to find the paper that they should read first, just by observing the citation networks. Such users can filter papers on the display by selecting a research category or entering a keyword. When the user enters a keyword in the text input widget shown in Figure 4.5 (b), the technique displays only the papers whose titles include the keyword. Also, When selecting a research category that the user is interested in, the node cluster that has only the research category is magnified in the center of the display in Figure 4.6 and Figure 4.7. It is useful to firstly overview, and then narrow down the focus cluster by selecting a category or entering a keyword when users want to survey the whole contents of the conference or research fields. Users can track bundles of the focus cluster, and then move to focus on other clusters. In case users want to look into respective papers, they can also narrow down the focus paper in the same procedure. If users click a paper node, its citation edges are highlighted. Users can follow these edges and trace them. VIGOR [108] is an integrated visualization technique to show the query, the result of joining all query matches, subgraphs of the search result with clustering, and summarization of each cluster' s feature distributions. We think visualizing the words in the user-focused cluster by tag cloud or visualizing its feature distribution like VIGOR will also enable novice researchers to understand papers and topics for their survey.

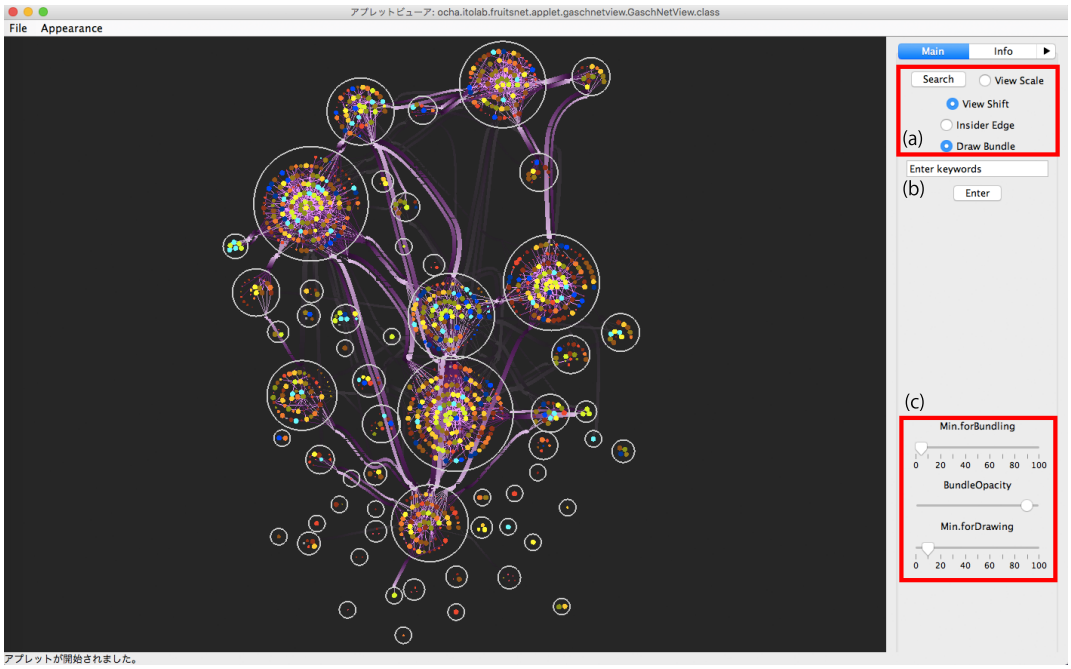

Figure 4.5 User Interface: (a) Scale and shift the view, switches the edge bundling mode. (b) Enter the keyword to display only the papers whose titles include it. (c) Set its threshold.

4.4 Examples and Evaluation

This section introduces some use cases of the visualization technique presented in the previous section. We implemented the proposed technique with Java Development Kit (JDK) 1.6.0. In this section. First, the section shows some visualization results of the proceedings of a single conference. It then shows the results in a case of applying our technique to the dataset of multiple conferences. In addition to that, we asked the questionnaires to the 21 graduate students as an example of novice researchers and discuss the effectiveness of the technique with the results.

4.4.1 An Example of a Conference Proceeding

We applied a citation network dataset consisting of 1072 full papers published in the SIGGRAPH conferences from 1990 to 1994, and from 2000 to 2010, provided by the ACM Digital Library [27]. We extracted the title, publication year, abstract, references, and authors from HTML files of the papers. We did not apply the paper information from 1995 to 1999, because we could not extract the abstracts from ACM Digital

Library.

Example of Hardware and GPU

Suppose that a user survey for research papers on *hardware and GPU*. Figure 4.6 is an example when the user selected the *hardware and GPU'* category. We could observe that the cluster in the center contained papers categorized only to *hardware and GPU* had dense relationships between the *physical simulation*, *lighting*, and *shape modeling* categories. We also found that the cited bundles of the *hardware and GPU* cluster are thicker than the citing ones, which means many papers in these research fields *physical simulation*, *lighting*, and *shape modeling* refer to the papers in the *hardware and GPU* cluster, and the researches in these fields have often evolved based on the researches in the *hardware and GPU* category. Among these relationships, especially, the relation between the *hardware and GPU* and *lighting* clusters clearly shows the above fact. Therefore, we expect that the *hardware and GPU* cluster could give a clue to the research team that develops hardware systems when they want to know which research fields their products are well applied.

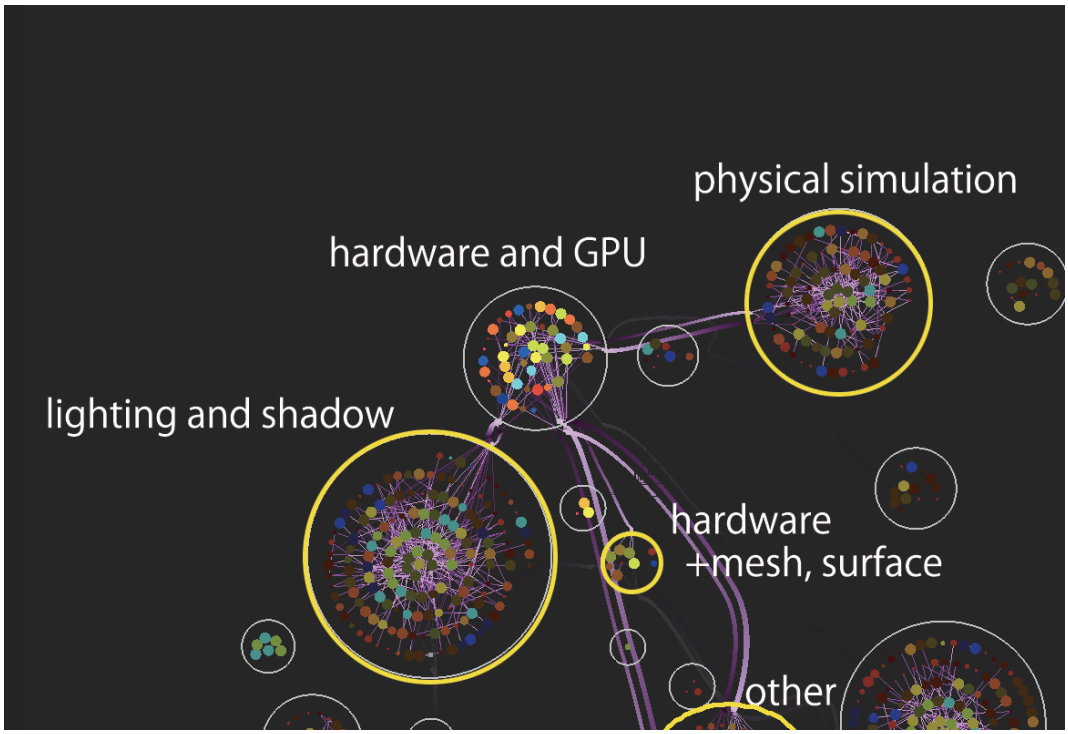

Figure 4.6: Example of *hardware and GPU*.

Example of lighting and CG algorithm

Next, we supposed that a user searched for papers related to lighting. Figure 4.7 shows an example of visualization under this supposition. The cluster A is a group that categorized into *lighting and CG (Computer Graphics) algorithm*. We found the nodes in this cluster were colored in light blue or yellow-green, where the colors depicted that the papers corresponding to the nodes were published in 1994 and 2000. Although this cluster is small, the problems in this research field were addressed once in 1994 and discussed again in 2000. The papers in the cluster A are as follows:

- A fast shadow algorithm for area light sources using back projection (in 1994) [59]
- The irradiance Jacobian for partially occluded polyhedral sources (in 1994) [35]
- A clustering algorithm for radiosity in complex environments (in 1994) [123]
- Illuminating micro geometry based on precomputed visibility (in 2000) [70]
- Efficient image-based methods for rendering soft shadows (in 2000) [32]
- Conservative volumetric visibility with occluder fusion (in 2000) [119]

Example with a keyword

Figure 4.8 (a) shows an example that a user entered the keyword *skin*. When the user did not apply the edge bundling and clicked the two orange nodes, many edges are drawn as shown in Figure 4.8 (b). The technique highlights the edges connected to the clicked nodes and the citations of the cited and citing nodes. Figure 4.8 (a) demonstrates that we can classify the research papers whose titles contain the term *skin* into two research fields. Therefore, we clicked two orange nodes, one in a larger cluster, and the other categorized in the different cluster far from the first one. As a result, we could grasp the two streams containing each of the clicked nodes because all the displayed nodes in Figure 4.8 (b) connect with either blue or green edges. We listed all the titles and figures (see Figure 4.9 and Figure 4.10) of the papers classified into these two groups. The papers connected with green edges as follows:

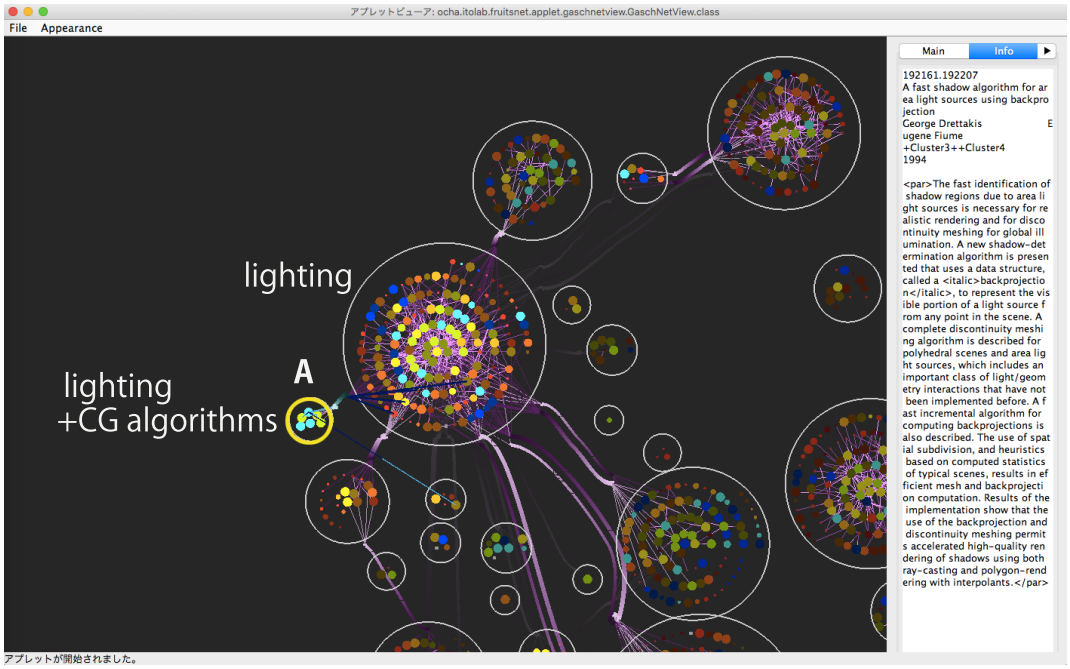

Figure 4.7: Example of *lighting and CG algorithm.*

(a) Continuous capture of skin deformation [115] (in 2003)

(b) Building efficient, accurate character skins from examples [100] (in 2003)

(c) Capturing and animating skin deformation in human motion [105] (in 2006)

(d) Data-driven modeling of skin and muscle deformation [106] (in 2008)

We listed the papers that belong to the blue stream.

(e) Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin [126](in 2003)

(f) Analysis of human faces using a measurement-based skin reflectance model [130] (in 2006)

As we could understand from these pictures of the papers, our technique demonstrated that researches of SIGGRAPH related to *skin* could be divided into two groups, based on their topics and citations. One of the topics is related to human animation generation using motion capture systems, and the other discusses generation or analysis of human face skins. This result demonstrates that the technique enables the novice researchers,

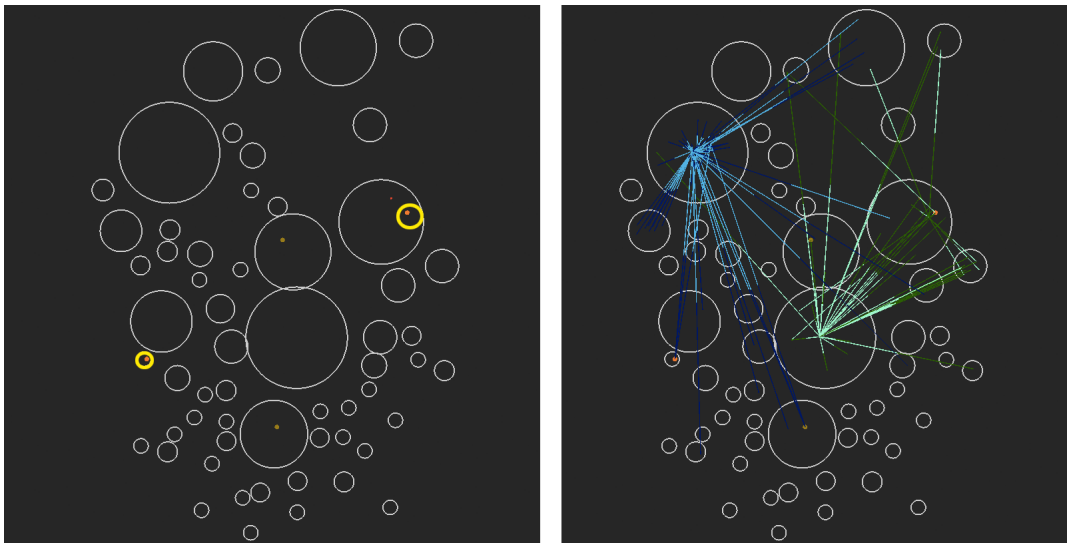

Figure 4.8 Example with a keyword: (a) Result with a keyword *skin*, (b) Result when an user click two nodes. The edge bundling is not applied.

who study computer graphics and want to read papers related to skin, to understand that there are two research fields related to skin and to choose which field they should survey.

4.4.2 Examples of Journals

As an example of multiple journals, we applied a citation network dataset consisting of 3604 full papers published in IEEE Transactions on Visualization and Computer Graphics (TVCG) and IEEE Computer Graphics and Applications (CG&A) from 1981 to 2015 provided by the IEEE Explore Digital Library [29]. We removed papers whose abstract cannot be provided in the library from the dataset. Figure 4.11 shows a visualization result of our technique. The papers of TVCG are placed on the left side of Figure 4.11 and the papers of CG&A are placed on the right side of Figure 4.11. This result shows that there are many citation relationships between TVCG and CG&A. The following is a list of topics of the papers published by TVCG and CG&A.

1. VR and AR
2. Simulation
3. Geometry and modeling

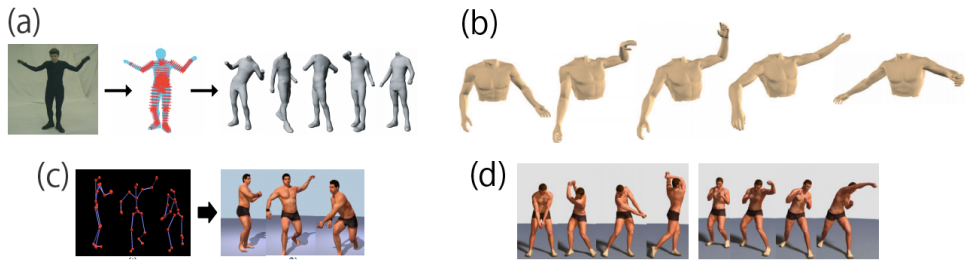

Figure 4.9 Pictures in papers of the green stream. (a) Continuous capture of skin deformation [115]. (b) Building efficient, accurate character skins from examples [100]. (c) Capturing and animating skin deformation in human motion [105]. (d) Data-driven modeling of skin and muscle deformation [106].

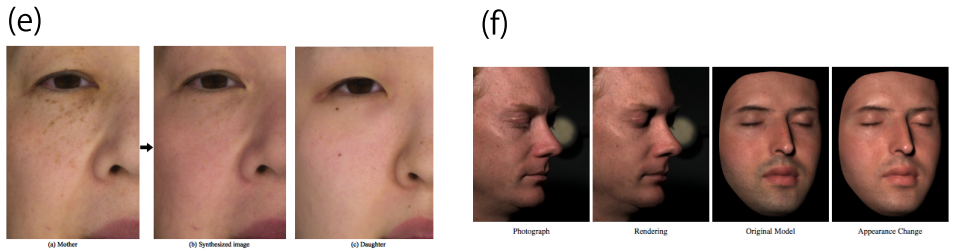

Figure 4.10 Pictures in papers of the blue stream. (e) Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin [126]. (f) Analysis of human faces using a measurement-based skin reflectance model [130].

4. Animation and motion

5. Interactive system

6. Lighting and rendering

7. Volume rendering

8. Software and environment

9. Art

10. Color and projection

11. GPU and hardware

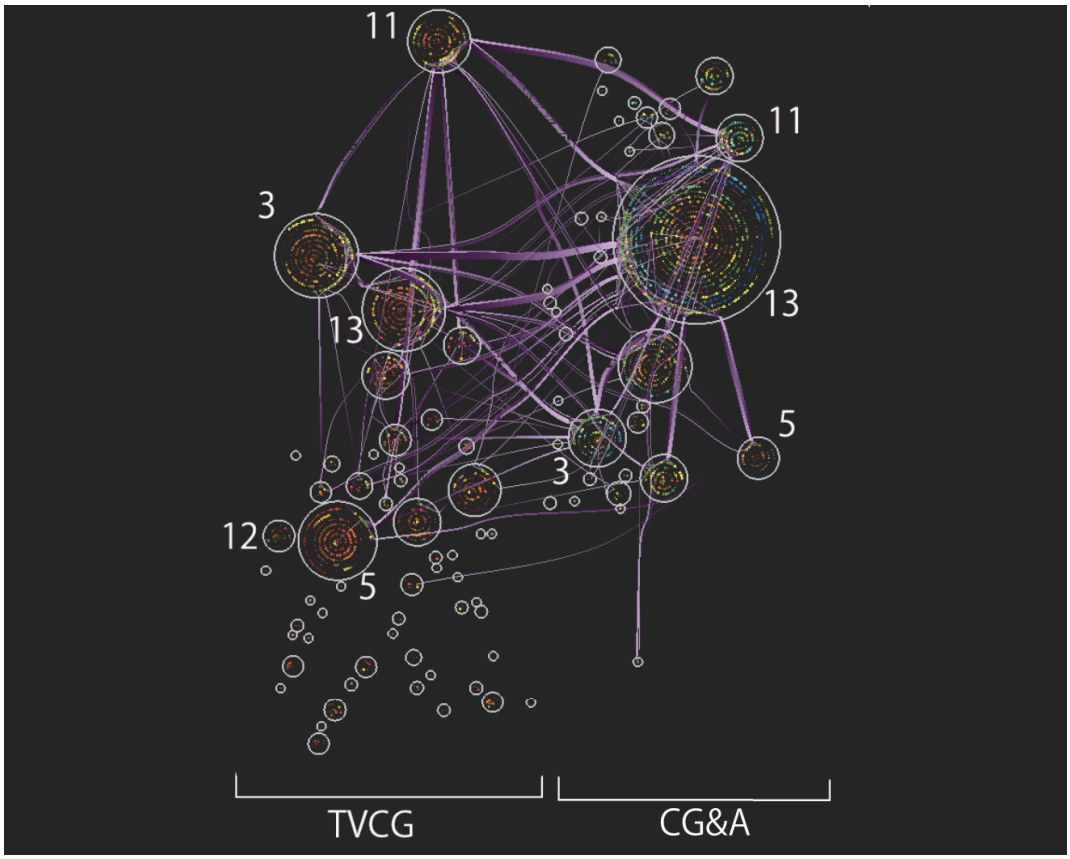

Figure 4.11 Example of IEEE Transactions on Visualization and Computer Graphics (TVCG) and IEEE Computer Graphics and Applications (CG&A). The numbers correspond to the itemization of the topics.

12. Graph visualization

13. Others

As the trend of topics in TVCG and CG&A, papers which include only topic 12 are published in TVCG, but not much published in CG&A. The papers in cluster 12 are colored in brown and red. They do not cite papers in CG&A. This explains that CG&A would not be helpful for users who are interested in *graph visualization*. Then, we introduce another feature between TVCG and CG&A. Figure 4.11 shows that papers in the clusters which include only topic 3 or 11 have dense relationships between the two journals. They are related to *geometry and modeling* and *GPU and hardware*. The papers of these two topics cite each other. This result would give a clue to find papers about the topics as the visualization result of SIGGRAPH dataset shown in the

previous section. On the other hand, papers in the clusters which include only topic 5 do not cite each other much between the two journals. However, the papers in these clusters have some common relationships to papers in the largest cluster of CG&A. The papers in the clusters 13 are not categorized in any of the above 12 topics. In this way, our proposed technique can help to observe the trend of papers published in multiple journals.

4.4.3 Evaluation

Preliminary questionnaires

There have been a lot of citation visualization techniques as we mentioned in Section 2.2.2. As against our technique applies a general purpose graph layout technique, typical existing techniques places nodes corresponding to the papers in time-ordered. We assume the time-ordered layout policy is not mandatory since it is sufficient for many users to recognize each of the visualized papers are old or new. For example, we often just want to know whether the paper is the oldest one as the roots in the research field or the newest one. To prove our hypothesis, we conducted a subjective evaluation to compare our technique and the time-oriented visualization technique. Before the evaluation, we had a questionnaire to define what we carefully observe while surveying papers. We asked three questions to ten graduated students majoring computer science.

1. What do you want to know when you search for papers?
2. What technique do you want for surveying papers well?
3. What do you want to know if you look into the citation network visualization in a particular conference for twenty years?

Regarding the question 1, a half of the students answered that they would like to know whether the papers are similar to their researches. In other words, it is important to define criteria of similarity of research topics and papers. Other answers are regarding citations and research topics or fields of papers. These answers suggest the usefulness of visualizing topic-based structures of papers and citations. We also suppose the

structures of topics and citations can be used to determine the similarity among papers. Several students answered they wanted to know the differences (e.g. advantages and disadvantages) among the techniques presented in the papers. We would like to solve this issue as a future work because both our technique and the existing techniques cannot represent the concrete contents only with the visualization results.

Regarding the question 2, more than half of the students mentioned that word-based smart search techniques are important for paper survey processes, including synonym recommendation and search refinement. This result proves that novice researchers including graduated students had troubles while selecting keywords to search for papers. Regarding the question 3, we roughly divide the answers into three categories, *the transition of research fields*, *the citation relations*, and *both research fields and citation relations*, or *what they reveal in combination*. It demonstrates the demands to understand both research fields and citations.

Evaluation: comparison with time-oriented visual representation

According to the result of the questionnaire, we asked 21 graduate students majoring computer science to compare our technique shown in Figure 4.12 (A) with the time-oriented citation visualization shown in Figure 4.12 (B), and evaluate which visualization is proper to know the contents below. We implemented the time-oriented technique mimicking Citeology.

We asked participants to answer the questions as 5-level scores, where 5 represents a strong agreement with A, and 1 represents a strong agreement with B. The following are the contents for which we asked the participants which visualization is more suitable:

1. The transition of papers amount published in the conference every year.
2. The main topic of the conference.
3. The trend of a research topic by year.
4. The research fields that seem to have a strong relationship with a field you focus on.

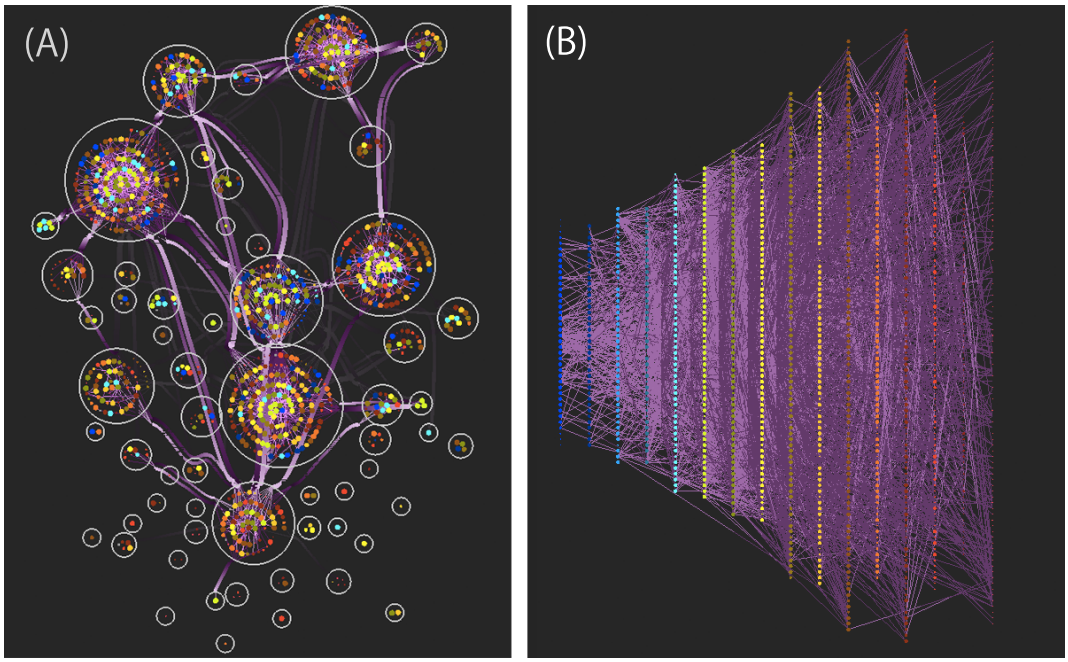

Figure 4.12 Comparison of visualization techniques: (A) Our technique. (B) Time-oriented technique.

5. Much-cited papers on a certain topic.

6. The latest paper on a certain topic.

7. The content trends of papers citing the paper you read (or clicked).

8. Papers whose contents are similar to the paper you read (or clicked).

9. Papers that had a great influence on the paper you read (or clicked).

Figure 4.13 shows the evaluation result. The X-axis denotes the sequential number of questions, and the Y-axis denotes the number of responses. Our technique was evaluated as more beneficial in the questions 2, 4, 5, 7, and 8, while the time-oriented visualization B was evaluated as more effective in the questions 1, 6, and 9. The questions 4, 5, 7, 8 correspond to the requirements **R1-4**. Therefore, the result of this evaluation demonstrated that our technique satisfies the requirements.

Although we expected the time-oriented technique B has an advantage on the questions 3 and 9, the figure demonstrates their rates varied widely. The result denotes our technique are also effective for the questions 3 and 9. Especially, the rate of the

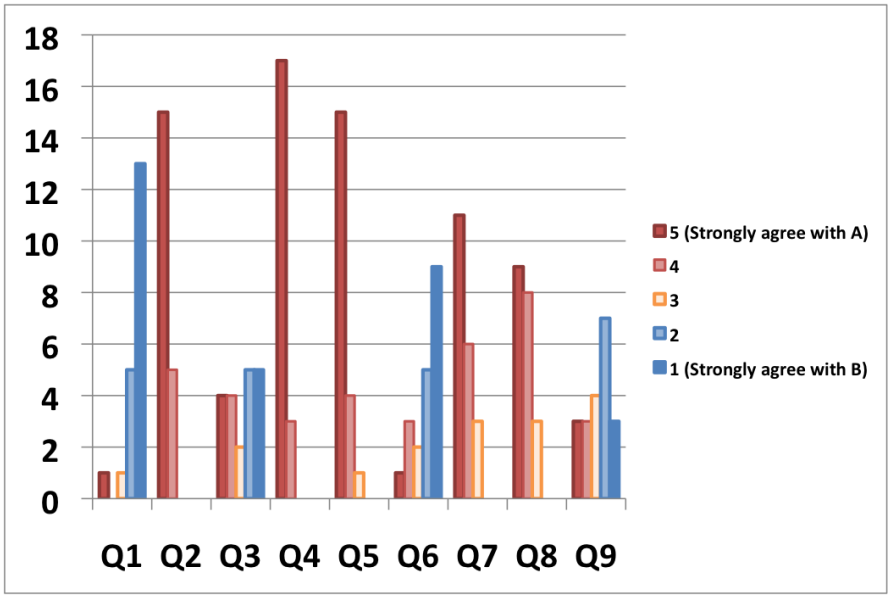

Figure 4.13 Result of the evaluation. The participants answered the questions as 5-level scores. 5 represents a strong agreement with A (our technique), and 1 represents a strong agreement with B (time-oriented technique).

questions 9 resulted in the variation because we did not need to know the publication year strictly to distinguish papers that had a great influence. This result proves that we do not need to assign the publication year to the X-axis of the display space.

4.5 Conclusion

This chapter defined four requirements of a survey of research papers, presented a visualization technique of citation networks based on them and discussed the visualization results. The technique applies topic-based paper clustering to construct a hierarchical network. It then applies a hybrid force-directed and space-filling network layout algorithm, and an edge bundling technique with Catmull-Rom spline curves. Our GUI design realizes the requirement **R1**, as a function of topic selection filtering. We applied datasets of publication of ACM SIGGRAPH, IEEE Transactions on Computer Graphics and Visualization, and IEEE Computer Graphics and Applications. These results showed our technique could help to understand the positions of papers in research fields and find papers even when users do not know all appropriate keywords. Also,

our technique could show the trends of topics and citation relations in a particular conference or journal. The case of visualization with a keyword *skin* demonstrated our technique satisfies the requirements **R2** and **R3** and the case of hardware topic of ACM SIGGRAPH showed for the requirement **R4**. This chapter also introduced the results of the user evaluation comparing with a time-oriented visualization technique. The result demonstrated that our technique was more helpful for novice researchers like students to find papers.

We determined the number of topics for LDA heuristically as mentioned in 4.3. The optimization of this parameter is important for more practical use. In addition to that, we plan to improve the layout of nodes inside the clusters to reduce visual clutters inside the node clusters. While our technique summarizes the relationships between the node clusters, the readability of the relationships of the nodes inside a cluster still remains a difficulty of understanding citation relationships. We would like to tackle these problems as future work.

Chapter 5

CoCoa: A Linked Network Visualization System of Co-citation and Co-author Relationships

5.1 Introduction

This chapter presents a visualization technique for citation networks applying a topic-based paper clustering. Section 5.2 introduces an overview of the system. Section 5.3 describes the implementation of the system in detail. Section 5.4 contains use cases the results of some evaluations.

The most famous metric of co-author network is Erdös number. The distance of co-authorship from the famous mathematics researcher Paul Erdös. This metric is not useful in the other research fields. We expect that visualization of co-author network would help survey of research papers rather than using such a metric. In recent years, Huang et al. [78] and Ebesu et al. [63] use neural network models for citation recommendation. In terms of understanding the detailed recommendations or search results, researchers visualize publication data in many ways [36] [72] [114] [87]. Federico et al. [66] reported a large number of visual approaches to scholarly literature and categorized them according to their data type and tasks. These studies still require users to input keywords for a query. In such a case, novice researchers sometimes miss papers when query keywords are not appropriate to survey their targeted research fields. Other techniques visualize citation relationships or co-author relationships to help survey of scholarly literature [83] [73]. However, few studies develop combinational visualizations

of citation and co-author networks. PivotPaths [58] is a visual interface for searching for faceted information resources. It visualizes relationships between tripartite information spaces, people, resources, and concepts. This technique does not visualize relationships between the items in the same information space. Moreover, it requires users to input any keyword for focusing on a particular item. Suppose that we are not familiar with a research field to investigate. It is probably difficult for us to cover the research field if we only use citation relationship or co-author relationship.

Therefore, we aim to help more efficient survey by combining text data of papers, citation relationships, and co-author relationships. We use a co-author network to represent the author information. Here, a co-author relationship between authors **A** and **B** indicates one or both of the following things:

1. An author **B** works in the research field **A'** where an author **A** works
2. A research field **B**' where an author **B** works can collaborate with the research field **A'** where an author **A** works

These can be clues to find related research fields and techniques. It would be effective for the help of survey to combine the citation relationships and the co-author relationships. The goal of this chapter is to help users to understand relationships among research fields, find influential authors in the research fields, and survey research papers casually.

5.1.1 Scenario

The previous chapter presented a visualization technique of a citation network applying topic-based clustering. We define the target users in this chapter as follows. The users we expect are not familiar with the research field. They are expected to have the following characteristics:

S1: They do not know all appropriate query keywords in that research field

S2: They do not know influential researchers in the research field

S3: They are not familiar with the relationships between the field and others

The target task of this study is to investigate a research field by exploratory search starting from papers or researchers in the field. Students who start studying their research field are typical target users. Such students may need to specify a direction of a future thesis and find any researchers to ask for advice of their study as a part of their tasks. We also expect our study can help young researchers who have fewer experiences on PC chairs or finding collaborators. For example, REMatch [77] is one of the systems which visualizes research publications to search for research experts for future collaborations. The tasks include finding researchers in unfamiliar research fields, and they have a similar feature in terms of *non-expert* for this task.

Based on this scenario, we redefine the requirements for visualization of citation and co-author networks.

R1: Understand the relationships between the focus research field and the others

R2: Find a paper related to the focus research field

R3: Find a researcher who can be a clue to find related research papers or an advisor of a research

5.2 Overview

This chapter presents a linked visualization system of citation and co-author networks named CoCoa. Our system first shows the overviews of citation and co-author relationships. When the users select one of the topic categories or input some keywords, this system filters the nodes of both the networks based on the users' selection.

Shown in Figure 5.1, the system provides a citation network on the left side of its view and a co-author network on the right side of the view by node-link diagrams. We suppose the papers and the authors as nodes, citations and co-author relationships as edges. The color of each paper node corresponds to the publication year of the paper, and the color of each author node corresponds to the publication year of the first paper that the author wrote. The size of each node is proportional to the number of its

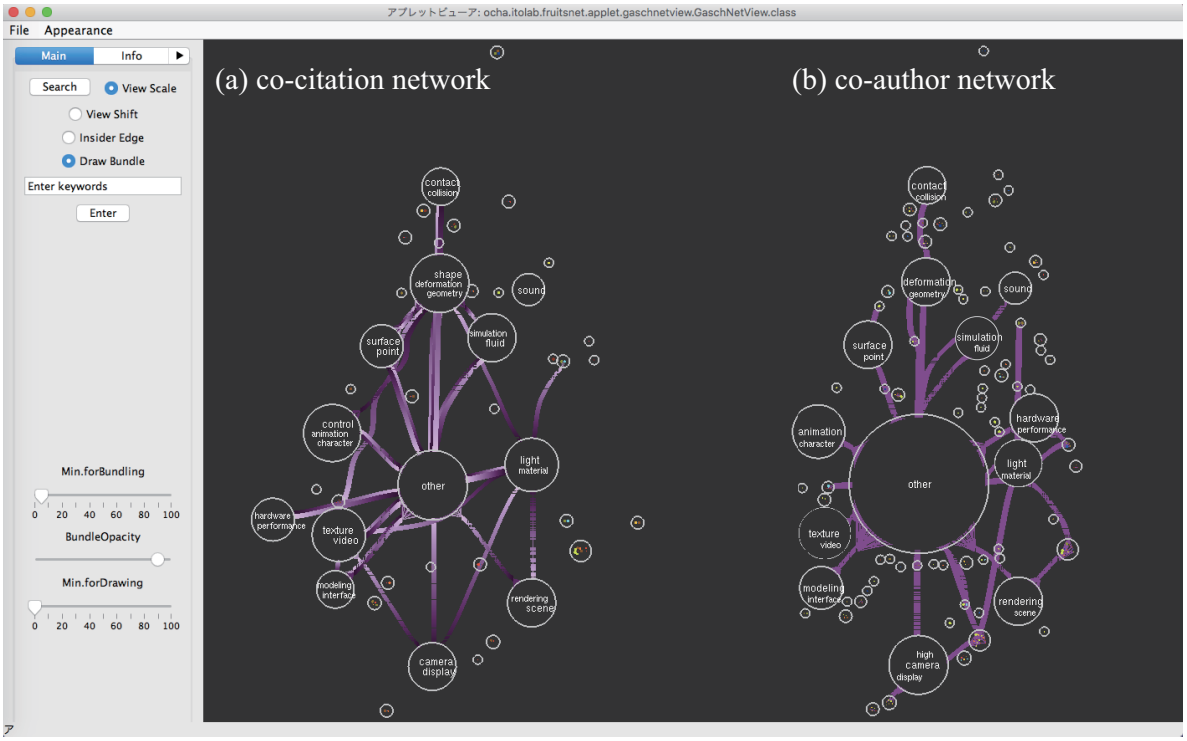

Figure 5.1 The overview of the system. The left view (a) shows a co-citation network and the right one (b) shows a co-author network. Each circle represents a cluster of nodes and labels of clusters indicate their topics. When a user zooms in, the labels disappear and the nodes of papers and authors appear.

citation count. Our system visualizes the directionality of citation as the brightness of the edge color while co-author relationships do not have the directionality. The system clusters nodes of networks by topics and the labels in the large clusters and represents the contents of the topics in the overview mode. Nodes in the clusters appear when a user zooms in. A user can filter nodes and edges in the one network by selecting a node of another network. Instead of showing relationships between two views like VisLink [49], this interaction presents relationships between two networks.

5.3 Implementation

This section describes the processing flow and the implementation of the presented technique. We also treat the papers and authors as nodes and citations and co-author relationships of networks.

5.3.1 Clustering Papers and Authors

Meanwhile, we define an author as a set of papers the author wrote. That is, we also represent an author as bags of words in the abstracts of their papers the author wrote. We define the k-th paper p_k as a set of m words in its abstract $S_k = \{w_1, w_2, \cdots, w_m\}$. We describe the i-th author a_i, who wrote n papers $p_1, p_2, ..., p_n$, as follows:

$$a_i = \bigcup_{k=1}^{n} S_k \tag{5.1}$$

We mix the collection of papers and authors to categorize them based on the same metrics. Our system applies LDA to the collection of words for papers and authors, estimates their topics, and then calculates the topic generative probability distribution for each paper and each author. We define a topic estimated by LDA as a research field. By using the distributions, the system categorizes both of papers and authors by their contents. A paper or an author belongs to a topic if the generative probability distribution of the topic for the paper or the author is larger than the user-specified threshold. Here, the clustering technique could generate many clusters that include only one paper or one author if there are too many combinations of topics. To avoid this situation, we limit the maximum combinatorial number of topics. We limit the combinatorial number to three in our implementation. In case that, the number of topics whose probability is larger than the threshold is more than three, we select the top three topics in terms of probability as representatives. Papers and authors are classified as *other* topic when their probabilities of all topics are lower than the threshold.

5.3.2 Topic Labelling

Next, we select the words to represent the contents of the cluster as follows:

1. List top words in the higher order of their generative probabilities
2. Compare the words in the k-th highest generative probability of each topic starting from k=1 and, if the words duplicate, employ the word as the representative of the topic where the generative probability is highest

3. Repeat step 2 until the maximum number of representative words in all topics becomes n

Many techniques on topic visualization apply labels or tag cloud histogram of appearance frequency for representative words. These techniques show top words as features of topics. The number of words ranges from 50 words to the user-defined number [50] [54] [128] [71]. Another technique is so-called auto topic labeling techniques [88] [89] [34], which generate labels by combining words with more importance. We first tried the technique Mei presented [99] and found that it may generate the same label candidates. Since visualizing appearance frequency of representative words requires larger display space, we visualize topics of clusters using the selected multiple words as labels.

The system selects the labels to visualize the contents of topics in the node clusters generated in the previous step. The number of labeling words is proportional to the size of clusters. The maximum number is three in our implementation. The size of clusters including multiple topics is smaller than that of clusters for a single topic. These clusters of multiple topics are placed near clusters including a common topic as described in the next section. Therefore, we label only clusters which include a single topic.

5.3.3 Network Placement

After clustering nodes of papers and authors, our system arranges the nodes based on the categorization. We determine the node positions of a citation network by applying Nakazawa' s technique [3] based on the result of the clustering described in Section 4.3. The system visualizes a citation network and a co-author network next to each other to make relationships between the two networks easier to understand. The clusters in different networks containing the same set of topics are separately placed when we place a citation network and a co-author network individually. This would destroy the users' mental map [62]. Therefore, we calculate the positions of the clusters in one network and then determine the positions of the clusters in another network. We place

the networks in the following steps:

1. List the clusters whose combination of the topics are common in both networks
2. Calculate the positions of the clusters in a citation network applying a hybrid force-directed and space-filling algorithm [3]
3. Apply the positions of the clusters listed in Step 1 as the initial positions of the clusters in a co-author network
4. Calculate the positions of the clusters in the co-author network similar to Step 2

After calculating the nodes of the networks, we summarize the edges applying the edge bundling. Edge bundling techniques highlight patterns to make easier to compare relationships between two networks [74].

5.3.4 Interaction

The CoCoA system supports interactions such as zooming, panning, and searching for the publication titles with a keyword. When a user wants to survey the whole contents of the conference or research fields, it is useful to firstly overview, and then narrow down the focus cluster by selecting a category or entering a keyword. They can track bundles of the focus cluster, and then move to focus on other clusters.

In case that a user wants to look into respective nodes, they can also narrow down the focus paper or author. The system shows the labels of the topic cluster, the paper title or author name of the node with a mouse hover. In addition to that, clicking a node enables a user to get the paper or the author shows detailed information on the left panel of the window. It includes the title or author name of the clicked node, ACM identifiers of the papers, authors of the papers, publication year, and abstracts. Clicking a node in the one network filters nodes in another network and presents only the nodes related to the clicked one. When a user selects an author, only paper nodes that have the selected author are shown in another network. The edges of the clicked node are also highlighted at the same time. A user can follow these filtered nodes and highlighted edges and trace them to find the next target node.

5.4 Use case

We applied a dataset of citation and co-author networks consisting of 1200 full papers published in the ACM SIGGRAPH conferences from 1990 to 2010 provided by the ACM Digital Library [27]. Note that this dataset does not include papers whose abstracts are not provided on the web pages. The total number of the authors is 2031.

Suppose that we are bachelor students who have just started to study Computer Graphics for our future thesis. Our interest is an application of modeling techniques, and we also need to study related areas. First, we select *modeling and interface* topic and the system magnifies the topic category *modeling and interface* in the center of the display in Figure 5.2. We can find an influential researcher related to the topic by clicking the large node **A** in the *modeling and interface* cluster of co-author network in Figure 5.2(b). The green highlighted edges denote the co-author relationships of the clicked node. The large node in the co-author network denotes the author published many papers in SIGGRAPH. The system shows the node **A** collaborated with researchers in several clusters. The labels of the clusters are *texture and video*, *modeling and interface*, *deformation and geometry*, and *others*. Also, the system filters nodes in the other view of the clicked one and provides only the publications of the researcher **A** in SIGGRAPH shown in the blue circles of Figure 5.2(a). The papers are categorized into topics *modeling and interface*, *deformation and geometry*, and *others*. Compared to co-author relationships, there were no publication related to *texture and video* topics. A wide bundle between these two clusters in Figure 5.2(a) shows that topic related to *texture and video* has a strong relationship between *modeling and interface*. The system leads us to think a collaborator of the researcher **A** is more familiar with the topic *texture and video*.

This result suggests two ideas. If we study modeling especially for *texture*, we often need to ask for advice from the co-author of the researcher **A** in this topic. We can study papers which the researcher **A** or other researchers in the clusters related to *modeling and interface* published, in a case of focusing on the user interface. Thus, our system helps exploratory search of scholarly literature and understanding trends across

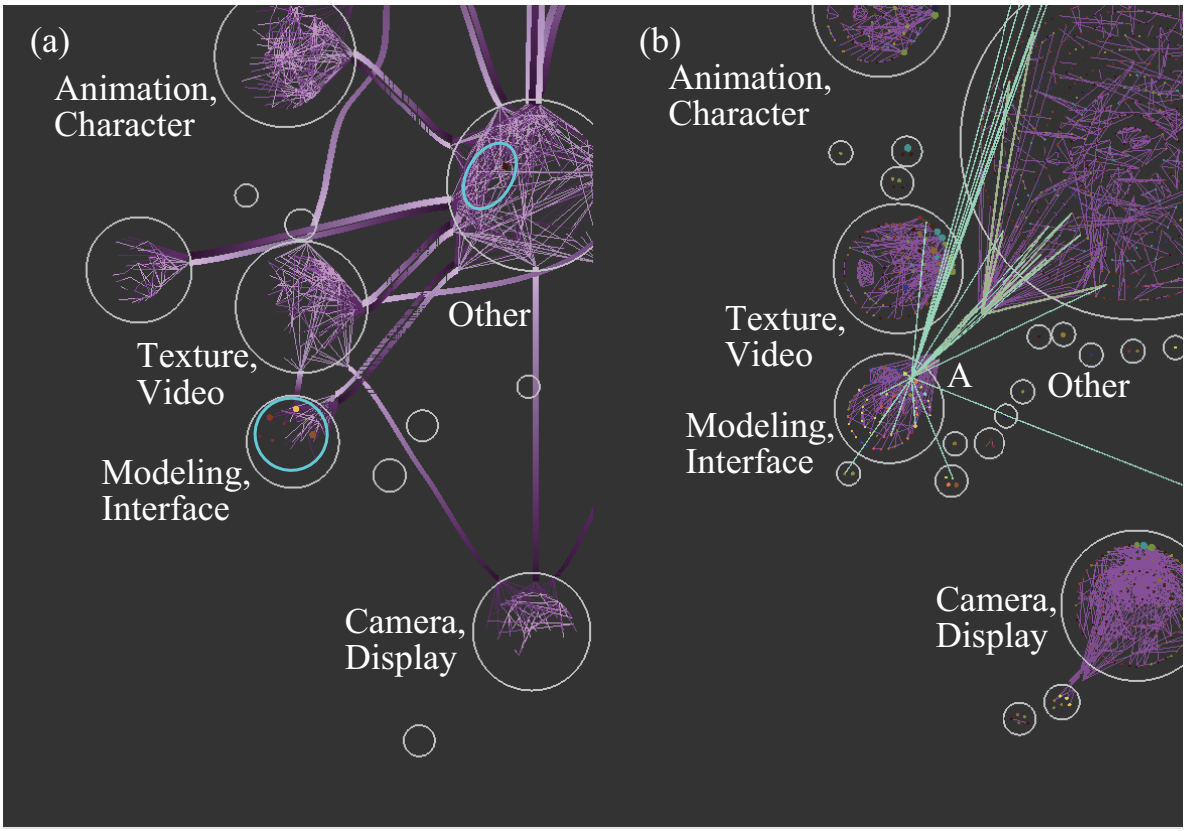

Figure 5.2 When selecting *modeling and interface* topic first, our system shows the clusters including the topic in the center of each view. The green highlighted edges denote the co-author relationships of the clicked node A in the view of (b). The system filters the nodes in another view (a) whether they are publications of the clicked node A or not. The nodes in blue circles in (a) are the publications of the clicked node A in (b).

multiple research fields by both citation and co-author relationships.

5.5 Conclusion

Both a co-citation network and a co-author network help search for scholarly literature. This chapter introduced a linked visualization system for these networks named CoCoa for the survey of scholarly literature. CoCoa treats the nodes of the two networks as bags of words and clusters them into topic groups together. The system places two networks next to each other. In addition to this, the system reuses the cluster position of a citation network as the cluster positions of a co-author network. Such a linked placement is expected to allow users to compare them easily. The example of citation

and co-author networks in the ACM SIGGRAPH conferences showed that the overview of the visualization system using topic clustering satisfies the requirements R1 and R2. Filtering nodes in the other view of the user's clicked node and comparison of the relationships gave us visual clues to find the related papers or advisors, which is the requirement R3. Thus, the system would help novice researchers to understand the relationships among research fields and find desired authors or papers.

As for future work, clustering granularity is one of the challenges. The current clustering process using LDA treats a node consisting of a small number of words at very low generative probabilities against all topics. Young researchers including students is an example of such nodes because the number of their publication in the conference is much smaller than senior researchers. Such nodes are grouped into the *other* topic. In result, the number of nodes in the *other* topic tends to become much larger than others, especially in the co-author network. We plan to apply the additional clustering process to the nodes whose generative probabilities against all topics are very low and add a user-defined categorization of them. Handling much larger dataset including multiple conference papers would be necessary for more practical use. In this case, we need to tackle a problem to determine the number of topics and clustering granularity. We plan to do clustering with LDA at a larger granularity once, and when the number of nodes in a cluster is large, we repeat clustering against the nodes in the cluster hierarchically. We are also planning to design and conduct user evaluations as future work.

Chapter 6

Conclusion

This chapter summarizes the main contributions presented in this dissertation, and offers suggestions for future work.

6.1 Summary

Networks in recent years include attributes of their nodes, and the nodes can be grouped by the attributes as categories. We need to connect these two types of information, relationships between data elements and categories of the elements, when we analyze network-structured data. Summarization view of categories and network topology, comparative view and interaction of visualization results help such network analysis including categorized nodes visually.

This dissertation proposes the visualization techniques of categorized networks for summarization and comparative analysis. The contributions of the visualization technique for network summarization include preserving both node similarities based on their categories and the edge structure to some extent, and avoiding visual clutters of nodes and edges. As for the proposed technique for comparative analysis, its contributions include coordinated placement of node clusters based on the major categories and the representation of a relationship between two networks with interactions of the users. The dissertation also introduces applications of the proposed techniques to a gene network and a citation relationship as examples of summarization and an application to citation and co-author relationships as an example of comparative analysis.

Acknowledgments

I would like to thank my supervisor, Professor Takayuki Itoh, for teaching, guiding, and encouraging me. Without his support and understanding of my situation, this dissertation would not exist.

I am also greatly appreciative of my dissertation's committee members, Professor Itiro Siio of Ochanomizu University, Professor Hiroaki Yoshida of Ochanomizu University, Professor Ichiro Kobayashi of Ochanomizu University, and Professor Mariko Hagita of Ochanomizu University, for their insightful comments and constructive criticisms that have helped to improve this dissertation.

My gratitude also goes out to Tamiya Onodera, Hiroshi Horii, and Miki Enoki of IBM Research - Tokyo, who always encouraged me and supported me to write this dissertation while working; Jun Sese and Aika Terada of Humanome Lab, who provided a dataset of a gene network and advised me about visualization of a gene network; Professor Kei Yura of Ochanomizu University and Sachiyo Aburatani of National Institute of Advanced Industrial Science and Technology, from who advised me about analysis of gene networks; Professor Takafumi Saito of Tokyo University of Agriculture and Technology, who provided a dataset of ACM SIGGRAPH papers; Akiko Eriguchi of Microsoft Research, who advised me about implementation of LDA technique; Professor Kwan-Liu Ma of University of California Davis, who advised me a lot about network visualization when I visited his laboratory.

Finally, I would like to extend my indebtedness to my family for their understanding, support, and encouragement throughout my study.

Bibliography

Publication

Journal Papers

[1] 中澤, 伊藤, 瀬々, 寺田, エッジの束化を用いた遺伝子ネットワークの可視化, 可視化情報学会論文集, Vol. 33, No. 11, pp. 25-32, 2013.

[2] N. Toeda, R. Nakazawa, T. Itoh, T. Saito, D. Archambault, Convergent Drawing for Mutually Connected Directed Graphs, Journal of Visual Languages and Computing, Vol. 43, pp. 83-90, 2017.

[3] R. Nakazawa, T. Itoh, T. Saito, Analytics and Visualization of Citation Network Applying Topic-based Clustering, Journal of Visualization, Vol. 21, No. 4, pp. 681-693, 2018. (可視化情報学会第 30 期（平成 30 年度）論文賞)

Conference talks in English

[4] R. Nakazawa, T. Itoh, J. Sese, A. Terada, Integrated Visualization of Gene Network and Ontology Applying a Hierarchical Graph Visualization Technique, 16th International Conference on Information Visualization (iV2011), 2012.

[5] R. Nakazawa, T. Itoh, J. Sese, A. Terada, Integrated Visualization of Gene Network and Ontology Applying a Hierarchical Graph Visualization Technique, IEEE VisWeek, Poster Session, 2012.

[6] R. Nakazawa, T. Itoh, T. Saito, A Citation Network Visualization Applying Topic-Based Paper Clustering, IEEE Pacific Visualization 2014, Poster Session, 2014. (Best Poster Award)

[7] R. Nakazawa, T. Itoh, T. Saito, A Visualization of Research Papers Based on the Topics and Citation Network, 18th International Conference on Information Visualisation (iV2015), pp. 283-289, 2015. (Best Paper Award)

[8] N. Toeda, R. Nakazawa, T. Itoh, T. Saito, D. W. Archambault, On Edge Bundling and Node Layout for Mutually Connected Directed Graphs, 20th International Conference Information Visualisation (iV 2016), pp. 94-99, 2016.

[9] R. Nakazawa, K. Ogata, S. Seelam, T. Onodera, Taming performance degradation of containers in the case of extreme memory overcommitment, In Proceedings of 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), pp. 196-204, 2017. (Acceptance rate: 18%)

[10] R. Nakazawa, T. Ueda, M. Enoki, H. Horii, Visualization Tool for Designing Microservices with the Monolith-First Approach, In Proceedings of 2018 IEEE Working Conference on Software Visualization (VISSOFT), pp. 32-42, 2018.

[11] T. Chiba, R. Nakazawa, H. Horii, S. Suneja, S. Seelam, ConfAdvisor: A Performance-centric Configuration Tuning Framework for Containers on Kubernetes, In Proceedings of 2019 IEEE International Conference on Cloud Engineering (IC2E), pp. 168-178, 2019.

[12] R. Nakazawa, T. Itoh, T. Saito, CoCoa: A Linked Network Visualization System of Co-citation and Co-author Relationships, In Proceedings of 21th EG/VGTC Conference on Visualization (EuroVis 2019), Short paper session, 2019.

Conference talks in Japanese

[13] 中澤, 伊藤, 瀬々, 寺田, 遺伝子ネットワークと Gene Ontology の統合可視化の一手法, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM 2012), C10-3, 2012.

[14] 中澤, 伊藤, 瀬々, 寺田, 遺伝子ネットワークと Gene Ontology の統合可視化の一手法, 情報処理学会第 74 回全国大会, pp. 105-106, 2012.

[15] 中澤, 伊藤, 瀬々, 寺田, エッジの束化を用いた遺伝子ネットワークの可視化, 2012年度人工知能学会全国大会, 2B1-R-3-3, 2012.

[16] 中澤, 伊藤, 瀬々, 寺田, エッジバンドリングを用いた遺伝子ネットワークと Gene Ontology の可視化の一手法, 第 41 回可視化情報シンポジウム, 2013.

[17] 中澤, 伊藤, 瀬々, 寺田, 遺伝子機能間の関係を明示する遺伝子ネットワークの束化と可視化, 情報処理学会第 35 回バイオ情報学研究会, BIO-35-2S, 2013.

[18] 中澤, 伊藤, 斎藤, 論文参照関係可視化の一手法 —SIGGRAPH 発表論文を例として, 画像電子学会ビジュアルコンピューティングワークショップ, 2013.

[19] 中澤, 伊藤, 斎藤, 研究分野と参照関係からたどる文献可視化インタフェース, 第 21 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2013), デモ/ポスター発表, 2013.

[20] 中澤, 緒方, 小野寺, メモリオーバーコミット環境下における Docker コンテナの性能変動の低減, 2015 年並列／分散／協調処理に関するサマー・ワークショップ (SWoPP2015), 2015.

[21] 十枝, 中澤, 伊藤, 斎藤, 有向グラフ可視化のためのバンドリングとノード配置, 情報処理学会第 78 回全国大会, pp. 299-300, 2016.

[22] 緒方, 中澤, 小野寺, 物理メモリ・オーバーコミット環境下における Docker コンテナのクラウドサーバ向けメモリ割当制御, 情報処理学会第 111 回プログラミング研究会 (PRO), 2017.

[23] 千葉, 中澤, 堀井, 動的なコンフィグチューニングによるコンテナ型アプリケーションの性能最適化, 2018 年並列／分散／協調処理に関するサマー・ワークショップ (SWoPP2018), 2018.

[24] 中澤, 岡, 堀井, コールスタックの要約とフレームの重要度にもとづく可視化の一手法，第 46 回可視化情報シンポジウム，2018.

[25] 千葉, 中澤, 堀井, Elastic Scheduling に向けたマイクロサービス性能モデルの検討, 2019 年並列／分散／協調処理に関するサマー・ワークショップ (SWoPP2019), 2019.

Bibliography

[26] 伊藤, 意思決定を助ける 情報可視化技術- ビッグデータ・機械学習・VR/AR への応用, コロナ社, 2018.

[27] ACM Digital Library, http://dl.acm.org/

[28] Google Scholar, http://scholar.google.co.jp/

[29] IEEE Xplore Digital Library, https://ieeexplore.ieee.org

[30] NCBI Entrez Gene, http://www.ncbi.nlm.nih.gov/gene

[31] J. Abello, F. Hohman, V. Bezzam, and D. H. Chau, Atlas: Local graph exploration in a global context, In *24th ACM International Conference on Intelligent User Interfaces*, IUI, 2019.

[32] M. Agrawala, R. Ramamoorthi, A. Heirich, L. Moll, Efficient image-based methods for rendering soft shadows, In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 375-384, 2000.

[33] A. Ahmed, X. Fu, S. Hong, Q. Nguyen and K. Xu, Visual Analysis of History of World Cup: A Dynamic Network with Dynamic Hierarchy and Geographic Clustering, In *Proceedings of Visual Information Communication (Proceedings of VINCI'09)*, Springer, pp. 25-39, 2010.

[34] N. Aletras, T. Baldwin, J. H. Lau, and M. Stevenson, Representing topics labels for exploring digital libraries, In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 239-248, 2014.

[35] J. Arvo, The irradiance Jacobian for partially occluded polyhedral sources, In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 343-350, 1994.

[36] F. Beck, S. Koch, and D. Weiskoph, Visual analysis and dissemination of scientific literature collections with SurVis, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 22, pp. 180-189, 2016.

[37] M. Berger, K. McDonough, L. Seversky, cite2vec: Citation-driven document exploration via word embeddings, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 23, No. 1, pp. 691-700, 2017.

[38] A. Bigelow, M. Monroe, Jacob' s Ladder: The User Implications of Leveraging Graph Pivots, In *Proceedings of the 2019 IEEE Pacific Visualization Symposium*, 2019.

[39] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[40] D. Botstein, J. M. Cherry, M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, Gene Ontology: tool for the unification of biology, *Nature Genetics*, Vol. 25, pp. 25-29, 2000.

[41] U. Brandes, T. Willhalm, Visualization of bibliographic networks with a reshaped landscape metaphor, *Konstanzer Schriften in Mathematik und Informatik*, 2002.

[42] B. Breitkreutz, C. Stark, and M. Tyers, Osprey: a network visualization system, *Genome Biology*, 2003.

[43] The Cartolabe project, https://cartolabe.fr/, 2019.

[44] M. Cerm, J. Dokulil, and J. Katreniakov, Edge routing and bundling for graphs with fixed node positions, In *Proceedings of 2011 15th International Conference on Information Visualisation*, pp. 475-481, 2011.

[45] C.Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for information Science and Technology*, Vol. 57, No.3, pp. 359-377, 2006.

[46] J. K. Chou, and C. K. Yang, PaperVis: Literature review made easy, *Computer Graphics Forum*, Vol. 30, No. 3, pp. 721-730, 2011.

[47] P. Ciccarese, S. Mazzocchi, amd F. Ferrazzi, L. Sacchi, Genius: a new tool for gene networks visualization, In *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP)*, pp. 107-111, 2004.

[48] A. Clauset, M. E. J. Newman, and C. Moore, Fining Community Structure in Very Large Networks, *Physical review E*, 70(6), 2004.

[49] C. Collins, and S. Carpendale, VisLink: Revealing relationships amongst visualizations, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 6, pp. 1192-1199, 2007.

[50] C. Collins, F. B. Viegas, and M. Wattenberg, Parallel tag clouds to explore and analyze faceted text corpora, In *Proceedings of 2009 IEEE Symposium on Visual Analytics Science and Technology*, pp. 91-98, 2009.

[51] C. Collins, G. Penn, and S. Carpendale, Bubble sets: Revealing set relations with isocontours over existing visualizations, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, No. 6, pp. 1009-1016, 2009.

[52] T. Crnovrsanin, C. Muelder, R. Faris, D. Felmlee, K. L. Ma, Visualization techniques for categorical analysis of social networks with multiple edge sets, *Social Networks*, Vol. 37, pp. 56-64, 2014.

[53] A. Cruz, J. P. Arrais, and P. Machado, Interactive Network Visualization of Gene Expression Time-Series Data, In *Proceedings of 2018 22nd International Conference Information Visualisation (iV)*, pp. 574-580, 2018.

[54] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, Context preserving dynamic word cloud visualization, In *Proceedings of the 2010 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 121-128, 2010.

[55] M. Delest, T. Munzner, D. Auber and J. P. Domenger, Exploring InfoVis Publication History with Tulip, *IEEE Symposium on Information Visualization*, 2004.

[56] K. Dinkla, M. A. Westenberg, and Jarke J. van Wijk, Compressed Adjacency Matrices, Untangling Gene Regulatory Networks, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp. 2457-2466, 2012.

[57] M. De Domenico, M. A. Porter, and A. Arenas, MuxViz: a tool for multilayer analysis and visualization of networks, *Journal of Complex Networks*, 3. 2, pp. 159-176, 2015.

[58] M. Dörk, N. H. Riche, G. Ramos, S. Dumais, PivotPaths: Strolling through Faceted Information Spaces, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp. 2709-2718, 2012.

[59] G. Drettakis, E. Fiume, A fast shadow algorithm for area light sources using backprojection, In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 223-230, 1994.

[60] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, B. Dorr, Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization, *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 12, pp. 2351-2369, 2012.

[61] T. Dwyer, N. Henry Riche, K. Marriott, and C. Mears, Edge Compression Techniques for Visualization of Dense Directed Graphs, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2596-2605, 2013.

[62] P. Eades, L. Wei, K. Misue, and K. Sugiyama, Preserving the mental map of a diagram, *Technical Report IIAS-RR-91-16E*, Fujitsu Laboratories, 1991.

[63] T. Ebesu, and Y. Fang, Neural citation network for context-aware citation recommendation, In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1093-1096, 2017.

[64] N. Elmqvist, and P. Tsigas, CiteWiz: a tool for the visualization of scientific citation networks, *Information Visualization*, Vol. 6, No. 3, pp. 215-232, 2007.

[65] O. Ersoy, C. Hurter, F. V. Paulovich, G. Cantareiro, and A. Telea, Skeleton-based edge bundling for graph visualization, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12, pp. 2364-2373, 2011.

[66] P. Federico, F. Heimerl, S. Koch, and S. Miksch, A survey on visual approaches for analyzing scientific literature and patents, *IEEE transactions on visualization and computer graphics*, Vol. 23, No. 9, pp. 2179-2198, 2017.

[67] A. G. Forbes, A. Burks, K. Lee, X. Li, P. Boutillier, J. Krivine, and W. Fontana, Dynamic influence networks for rule-based models, *IEEE transactions on visualization and computer graphics*, Vol. 24, No. 1, pp. 184-194, 2018.

[68] R. Gansner, Y. Hu, S. North, and C. Scheidegger, Multilevel Agglomerative Edge Bundling for Visualizing Large Graphs, In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, pp. 187-194, 2011.

[69] S. Hadlak, H. Schumann, and H. J. Schulz, A Survey of Multi-faceted Graph Visualization, In *Proceedings of EuroVis (STARs)*, pp. 1-20, 2015.

[70] W. Heidrich, K. Daubert, J. Kautz, H. P. Seidel, Illuminating micro geometry based on precomputed visibility, In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 455-464, 2000.

[71] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, Word cloud explorer: Text analytics based on word clouds, In *Proceedings of 2014 47th Hawaii International Conference on System Sciences*, pp. 1833-1842, 2014.

[72] F. Heimerl, Q. Han, S. Koch, and T. Ertl, CiteRivers: Visual analytics of citation patterns *IEEE Transactions on Visualization and Computer Graphics*, Vol. 1, pp. 190-199, 2016.

[73] N. Henry, H. Goodell, N. Elmqvist, J. D. Fekete, 20 years of four HCI conferences: A visual exploration, *International Journal of Human-Computer Interaction*, Vol. 23, No. 3, pp. 239-285, 2007.

[74] D. Holten, Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data, *IEEE Transactions On Visualization And Computer Graphics*, Vol. 12, No. 5, pp. 741-748, 2006.

[75] D. Holten and J. J. Van Wijk, Force ⊠ Directed Edge Bundling for Graph Visualization, *Computer Graphics Forum*, Vol. 28, No. 3, pp. 983-990, 2009.

[76] D. Holten, and J. J. Van Wijk, A user study on visualizing directed edges in graphs, In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2299-2308, 2009.

[77] M. I. Hossain, S. Kobourov, H. Purchase, and M. Surdeanu, REMatch: Research Expert Matching System, In *Proceedings of 2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, 2018.

[78] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. Lee Giles, A Neural Probabilistic Model for Context Based Citation Recommendation, In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*, 2015.

[79] Y. Huang, L. Shi, Y. Su, Y. Hu, H. Tong, C. Wang, and S. Liang, Eiffel: Evolutionary Flow Map for Influence Graph Visualization, *IEEE transactions on visualization and computer graphics*, 2019.

[80] T. Itoh, C. Muelder, K. Ma, and J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, In *Proceedings of 2009 IEEE Pacific Visualization Symposium*, pp. 121-128, 2009.

[81] M. Itoh, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa, Analysis and visualization of temporal changes in bloggers' activities and interests, In *Proceedings of the 2012 IEEE Pacific Visualization Symposium*, pp. 57-64, 2012.

[82] S. Kairam, D. MacLean, M. Savva, J. Heer, Graphprism: compact visualization of network structure. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 498-505, 2012.

[83] W. Ke, K. Borner, L. Viswanath, Major information visualization authors, papers and topics in the acm library, *IEEE Symposium on Information Visualization*, 2004.

[84] K. Dinkla, M. J. Van Kreveld, B. Speckmann, and M. A. Westenberg, Kelp diagrams: Point set membership visualization, *Computer Graphics Forum*, Vol. 31, No. 3pt1, pp. 875-884, 2012.

[85] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Moreno, and M. A. Porter, Multilayer networks, *Journal of complex networks*, 2. 3, pp. 203-271, 2014.

[86] A. Lambert, R. Bourqui, and D. Auber, Winding roads: Routing edges into bundles, *Computer Graphics Forum*, Vol. 29, No. 3, pp. 853-862, 2010.

[87] S. Latif and F. Beck, VIS Author Profiles: Interactive Descriptions of Publication Records Combining Text and Visualization, *IEEE transactions on visualization and computer graphics*, Vol. 25, No. 1, pp. 152-161, 2019.

[88] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, Best topic word selection for topic labelling, In *Proceedings of the 23rd International Conference on Computational Linguistics (Posters)*, pp. 605-613, 2010.

[89] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, Automatic labelling of topic models, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 1536-1545, 2011.

[90] B. Lee, M. Czerwinski, G. Robertson, B.B. Bederson, Understanding research trends in conferences using PaperLens, *Extended Abstracts on the 2005 CHI Conference on Human factors in computing systems*, pp. 1969-1972, 2005.

[91] C. Lee, A. Garbett, J. Wang, B. Hu, D. Jackson, Weaving the Topics of CHI: Using Citation Network Analysis to Explore Emerging Trends, *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, p. LBW0115, 2019.

[92] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg, Comparative analysis of multidimensional, quantitative data, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No.6, pp. 1027-1035, 2010.

[93] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, Tiara: Interactive, topic-based visual text summarization and analysis, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 3, No. 2, pp. 25, 2012.

[94] J. D. Mackinlay, R. Rao, S. K. Card, An organic user interface for searching citation links, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 67-73, 1995.

[95] T. Major, and R. C. Basole, Graphicle: Exploring Units, Networks, and Context in a Blended Visualization Approach, *IEEE transactions on visualization and computer graphics*, Vol. 25, No. 1, pp. 576-585, 2019.

[96] A. Martin, M. E.Ochagavia, L. C. Rabasa, J. Miranda, J. Fernandez-de-Cossio, R. Bringas, BisoGenet: a new tool for gene network building, visualization and analysis, *BMC bioinformatics*, Vol. 11, No. 1, pp. 91, 2010.

[97] J. Matejka, T. Grossman, and G. Fitzmaurice, Citeology: visualizing paper genealogy, *Extended Abstracts of the 2012 CHI Conference on Human Factors in Computing Systems*, pp. 181-190, 2012.

[98] F. Mcgee, M. Ghoniem, G. Melançon, B. Otjacques, and B. Pinaud, The State of the Art in Multilayer Network Visualization, *Computer Graphics Forum*, 2019.

[99] Q. Mei, X. Shen, and C. X. Zhai, Automatic Labeling of Multinomial Topic Models, In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2017.

[100] A. Mohr, M. Gleicher, Building efficient, accurate character skins from examples, *ACM Transactions on Graphics (TOG)*, Vol. 22, No. 3, pp. 562-568, 2003.

[101] C. Muelder and K. L. Ma, A Treemap Based Method for Rapid Layout of Large Graphs, In *Proceedings of the 2008 IEEE Pacific Visualization Symposium*, pp. 231-238, 2008.

[102] K. Nishiyama, and T. Itoh, Visualization of Hierarchical Gene Network using Heiankyo View, *The Journal of Society for Art and Science*, Vol. 6, No. 3, pp. 106-116, (in Japanese), 2007.

[103] C. Nobre, M. Streit, and A. Lex, A, Juniper: A Tree+ Table Approach to Multivariate Graph Visualization, *IEEE transactions on visualization and computer graphics*, Vol. 25, No. 1, pp. 544-554, 2019.

[104] F. Paduano, R. Etemadpour, and A. G. Forbes, BranchingSets: Interactively visualizing categories on node-link diagrams, In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, pp. 9-16, 2016.

[105] S. I. Park, J. K. Hodgins, Capturing and animating skin deformation in human motion, *ACM Transactions on Graphics (TOG)*, Vol. 25, No. 3, pp. 881-889, 2006.

[106] S. I. Park, J. K. Hodgins, Data-driven modeling of skin and muscle deformation, *ACM Transactions on Graphics (TOG)*, Vol. 27, No. 3, p. 96, 2008.

[107] N. Pezzotti, J. D. Fekete, T. Höllt, B. P. Lelieveldt, E. Eisemann, and A. Vilanova, Multiscale visualization and exploration of large bipartite graphs, *Computer Graphics Forum*, Vol. 37, No. 3, pp. 549-560, 2018.

[108] R. Pienta, F. Hohman, A. Endert, A. Tamersoy, K. Roundy, C. Gates, N. Shamkant, D. H. Chau, VIGOR: interactive visual exploration of graph query

results, *IEEE transactions on visualization and computer graphic*, Vol. 24, No. 1, pp. 215-225, 2017.

[109] A. Ponsard, F. Escalona, and T. Munzner, PaperQuest: a visualization tool to support literature review, In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2264-2271, 2016.

[110] S. Razick, G. Magklaras, and I. M. Donaldson, iRefIndex: A consolidated protein interaction database with provenance, *BMC Bioinformatics*, 9(1), 405, 2008.

[111] B. Renoust, G. Melançon, and T. Munzner, Detangler: Visual analytics for multiplex networks, *Computer Graphics Forum*, Vol. 34. No. 3, 2015.

[112] B. Renoust, V. Claver, and J. F. Baffier, Flows of knowledge in citation networks, *International Workshop on Complex Networks and their Applications*, pp. 159-170, 2016.

[113] B. Renoust, V. Claver, and J. F. Baffier, Multiplex flows in citation networks, *Applied Network Science*, 2.1, 23, 2017.

[114] A. Rind, A. Haberson, K. Blumenstein, C. Niederer, M. Wagner, W. Aigner, PubViz: Lightweight visual presentation of publication data, In *Proceedings of Proc. Eurographics Conference Visualization (EuroVis)*–Short Paper, 2017.

[115] P. Sand, L. McMillan, J. Popović, Continuous capture of skin deformation, *ACM Transactions on Graphics (TOG)*, Vol. 22, No. 3, pp. 578-586, 2003.

[116] D. Selassie, B. Heller, and J. Heer, Divided edge bundling for directional network data, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2354-2363, 2011.

[117] D. Shahaf, C. Guestrin, E. Horvitz, Metro maps of science, In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1122-1130, 2012.

[118] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, and T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome research*, Vol. 13, No. 11, pp. 2498-2504, 2003.

[119] G. Schaufler, J. Dorsey, X. Decoret, F. X. Sillion, Conservative volumetric visibility with occluder fusion, In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 229-238, 2000.

[120] K. Shiratori, T. Itoh, Journal Visualization by a Dual Hierarchical Data Visualization Technique, In *Proceedings of the 2009 IEEE Pacific Visualization Symposium*, Poster Session, 2009.

[121] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, In *Proceedings 1996 IEEE symposium on visual languages*, pp. 336-343, 1996.

[122] H. Small, Visualizing science by citation mapping, *Journal of the American society for Information Science*, Vol. 50, No. 9, pp. 799-813, 1999.

[123] B. Smits, J. Arvo, D. Greenberg, A clustering algorithm for radiosity in complex environments, In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 435-442, 1994.

[124] J. Stasko, C. Görg, and Z. Liu, Jigsaw: supporting investigative analysis through interactive, *Information visualization*, Vol. 7, No. 2, pp. 118-132, 2008.

[125] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, C. D. Stolper, CiteVis: Exploring Conference Paper Citation Data Visually, *IEEE InfoVis*, Poster Session, 2013.

[126] N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima,Y. Miyake, Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin, *ACM Transactions on Graphics (TOG)*, Vol. 22, No. 3, pp. 770-779, 2003.

[127] N. J. Van Eck, and L. Waltman, CitNetExplorer: A new software tool for analyzing and visualizing citation networks, *Journal of Informetrics*, Vol. 8, No. 4, pp. 802-823, 2014.

[128] P. Venetis, G. Koutrika, and H. Garcia-Molina, On the selection of tags for tag clouds, In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 835-844, 2011.

[129] Y. Wang, D. Liu, H. Qu, Q. Luo, and X. Ma, A Guided Tour of Literature Review: Facilitating Academic Paper Reading with Narrative Visualization, In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, pp. 17-24, 2016.

[130] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, M. Gross, Analysis of human faces using a measurement-based skin reflectance model, *ACM Transactions on Graphics (TOG)*, Vol. 25, No. 3, pp. 1013-1024, 2006.

[131] H. Zhou, X. Yuan, W. Cui, H. Qu, and B. Chen, Energy-Based Hierarchical Edge Clustering of Graphs, In *Proceedings of 2008 IEEE Pacific Visualization Symposium*, pp. 55-61, 2008.

[132] H. Zhou, P. Xu, X. Yuan, H. Qu, Edge bundling in information visualization, *Tsinghua Science and Technology*, Vol. 18, No. 2, pp. 145-156. 2013.

Advanced Science,
Graduate School of Humanities and Sciences,
Ochanomizu University,

Rina Nakazawa