A Visualization of Research Papers Based on the Topics and Citation Network

Rina Nakazawa, Takayuki Itoh, Takafumi Saito
Ochanomizu University, Tokyo University of Agriculture Technology
{leena@itolab.is.ocha.ac.jp, itot@is.ocha.ac.jp, txsaito@cc.tuat.ac.jp}

Abstract

Survey of research papers is not an easy task for novice researchers, because they are not always good at finding all appropriate keywords for the survey. Moreover, it is not easy for them to understand positions of papers in their research fields instantly, even when they use famous search engines like Google Scholar; it may often take a long time for them to find scholarly literature. On the other hand, many researchers have presented citation visualization techniques for surveying research papers. However, it is still often difficult to observe the complicated relations across multiple research fields or traverse the entire relations in their interest. In this paper, we proposed a visualization technique for citation networks applying topic-based paper clustering. Our technique categorizes papers applying LDA (Latent Dirichlet Allocation), and constructs clustered networks consisting of the papers.

Keywords--- **Citation network visualization, edge bundling, topic-based.**

1. Introduction

Survey of research papers is very important for research processes, understanding the trend of the research fields and finding the related work. Researchers use text-based portal Web sites such as Google Scholar [1] and ACM Digital Library [2], and look up for the references of papers they read. However, it is difficult for novice researchers to survey papers and grasp the position of the papers in the research fields with search results. Besides, they may miss papers in case they do not know all the appropriate keywords and in case papers they really want to survey straddle multiple fields.

There have been many researches on visualization of citation networks such as Mackinlay [3] and Small [4] which are useful for survey of research fields. However, we suppose still there are many open problems on visualization of citation networks. For example, researchers continuously trigger for new fusional fields, and therefore they need to organize and understand the relations of papers that cover multiple fields. To organize the open problems, we define the demands in visualization of citation networks for survey of papers as follows:

- Categorize papers that have similar topics to the same group
- Place papers that belong to the category in common closer
- Place citing and cited papers closer
- Summarize citation relationships

In this paper, we propose a visualization technique that satisfies these four demands. Users can follow up research fields and citations, and finally understand the relations among the related papers using this technique. Our technique categorizes papers based on their contents first. Then, it constructs a citation network by treating papers as nodes, and citations as directed edges. Our technique visualizes the relations of research papers with their contents and citations. It would help novice researchers to understand the differences between the tendencies of similar research fields.

2. Related Work

This section introduces existing visualization techniques for paper citation networks.

PaperLens [5] 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. Brandes et al. [6] presented a visualization technique for citation networks with topographic maps that places the hub papers cited by many papers higher. It also arranges the papers that have similar citation pattern closer. That enables us to easily find the hub papers and the groups of papers that have similar citation patterns. Citeology [7] orders papers by the number of their citations with respect to each year, and places them from the center of the display. It can visualize up to eight generation 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. Dunne et al. [8] proposed an integrated visualization of citation network and summary of papers. The users can look at the citation, ranking based on the citation count, and summary of papers in the cluster generated by graph clustering based on citation structure at the same time. It may require larger display spaces. Also, the network visualization shows only papers extracted by the keyword-based search, so the users may miss papers that are cited by several papers related to the keyword, because they do not include the particular keyword.

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 citation 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. Another 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 their fusion. However, there are few visualization techniques addressing this problem.

3. Proposed technique

This section describes the processing flow of the presented technique. We treats 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. It then applies our hierarchical network layout technique with an edge bundling algorithm. Our implementation also provides rendering and interaction techniques.

3.1. Paper classification

The proposed technique applies LDA (Latent Dirichlet Allocation) [9] to categorize papers based on the contents of papers. LDA is a generative topic model that considers each document as a mixture of various topics. It could solve the problem to categorize papers that straddle multiple research fields. The technique applies LDA to the sets of paper abstracts to estimate topics and calculate the topic distribution for each abstract. 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 classification 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 in the topic.

3.2. Network layout

Next, our technique arranges nodes applying a hybrid force-directed and space-filling graph drawing algorithm [10] to calculate the positions of nodes corresponding to the individual papers. 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.

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 bundle the edges or not. We have already implemented the edge bundling algorithm in our previous work [11]; however, it had a problem that straight bundles with which summarizes a lot of edges may lead to misconceptions of Gestalt principle (Figure 1) 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 2). 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 Figure2-(2). Then, it bundles the edges only when the number of edges between two clusters is larger than the threshold. (Figure2-(3)). The technique applies this process to all pairs of the node clusters (Figure2-(4)).

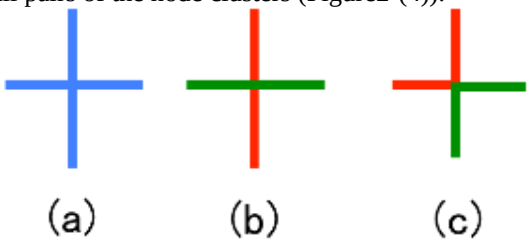

Figure 1 A misconception of Gestalt principle

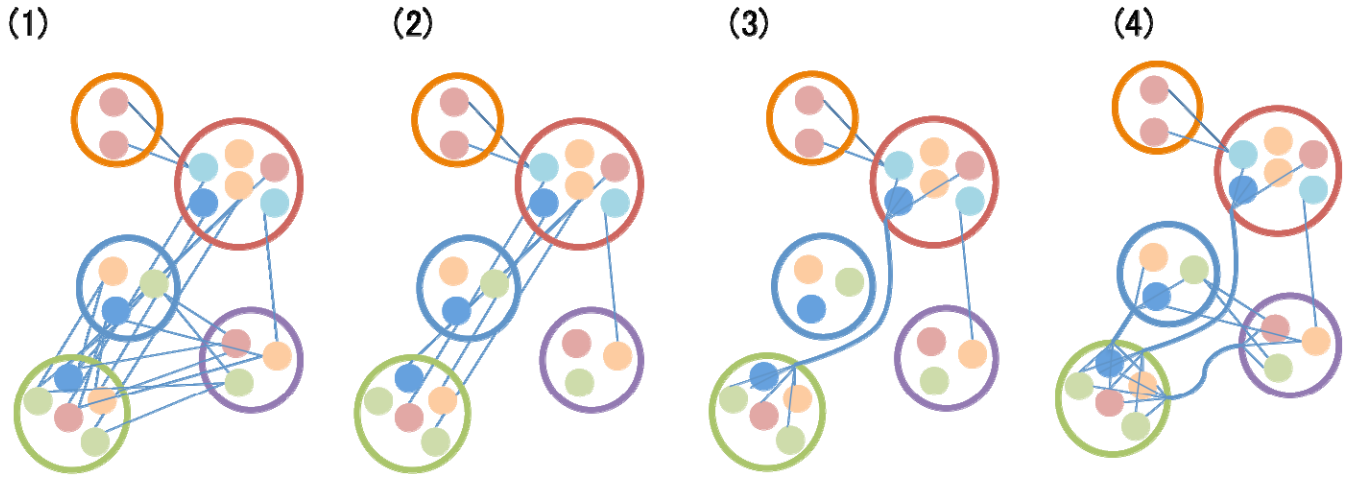

Figure 2 Edge bundling

3.3. 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. We can also draw arrows or assign different hues to the 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 arrows, 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 the edge direction. As Figure 3 shows, we draw nodes with the color scale corresponding to the publication years.

Figure 3 Color scaling (Upper): The node color, (Lower): The edge color

3.4. User Interface

Figure 4 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 (1)(3). When a user clicks a node corresponding to a particular paper, the technique displays the details of the paper such as ACM identifier, 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 is applicable 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. 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. Also, when the user enters a keyword in the text input widget shown in Figure 4 (2), the technique displays only the papers whose titles include the keyword. When users want to survey 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 users want to look into respective papers, they can also narrow down the focus paper in the same procedure. If they click a paper node, its citation edges are highlighted. They can follow these edges and trace them.

4. Results

We implemented the proposed technique with Java Development Kit (JDK) 1.6.0. We applied a citation network dataset consisting of 1072 full papers published in the SIGGRAPH conferences during 1990 to 1994, and during 2000 to 2010, provided by the ACM Digital Library [2]. We extracted the title, publication year, abstract, references, and authors from html files of the papers. We did not apply the paper information during 1995 to 1999, because we could not extract the abstracts from ACM Digital Library.

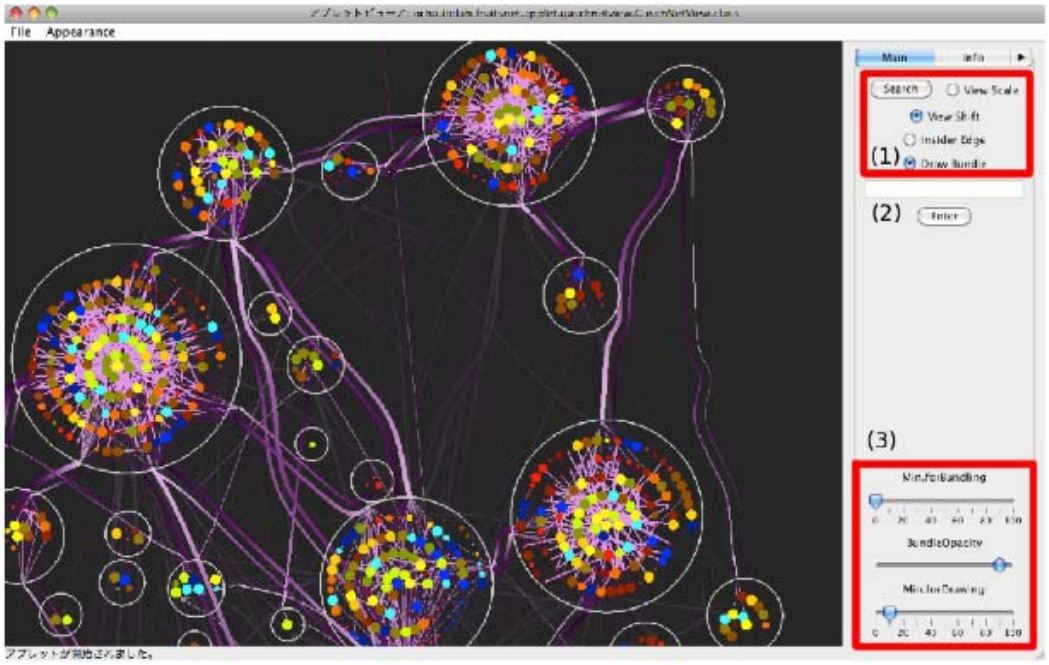

Figure 4 User Interface

4.1. Example of image processing

Figure 5 shows a visualization result when a user selected the "image processing" category. The cluster containing the papers categorized only to "image processing" appeared in the center of the view. Most of nodes of the cluster are warm colored. This denotes the tendency that most of papers on image processing are published by SIGGRAPH after 2000.

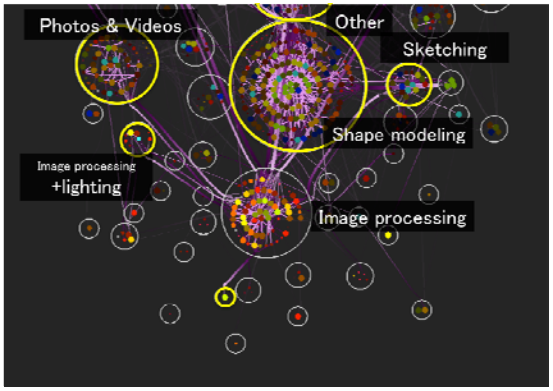

Figure 5 Example of image processing

4.2. Example of hardware

Suppose that a user surveys for research papers on “hardware and GPU.” Figure 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. 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.

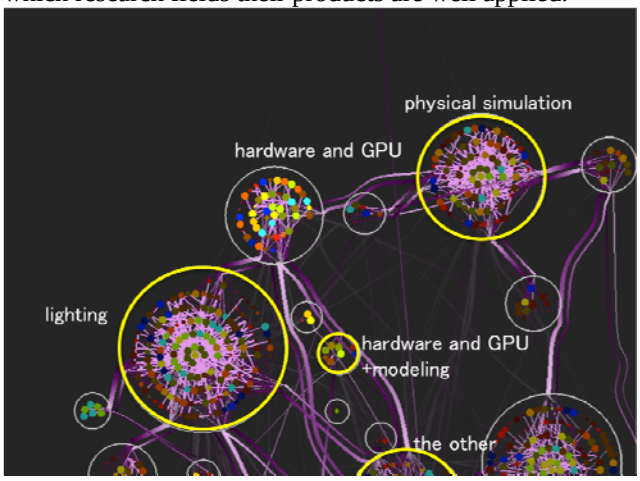

Figure 6 Example of hardware

4.3. Example of lighting and CG algorithm

Next, we supposed that a user searched for papers related to lighting. Figure 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)
- The irradiance Jacobian for partially occluded polyhedral sources (in 1994)
- A clustering algorithm for radiosity in complex environments (in 1994)
- Illuminating micro geometry based on precomputed visibility (in 2000)
- Efficient image-based methods for rendering soft shadows (in 2000)
- Conservative volumetric visibility with occluder fusion (in 2000)

We especially suppose the above papers published in 2000 might have triggered the invention of PRT (Precomputed Radiance Transfer).

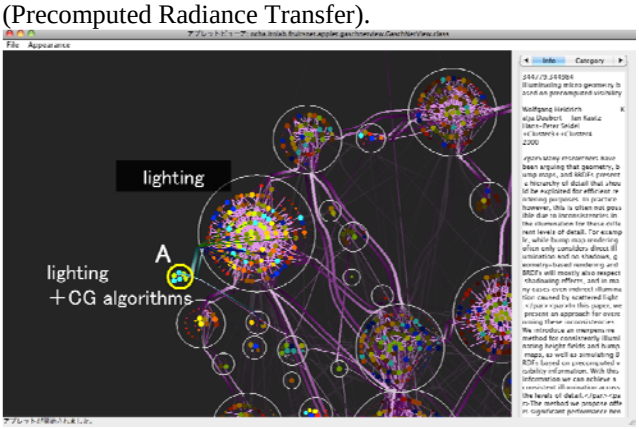

Figure 7 Example of lighting and CG algorithm

4.4. Example with a keyword

Figure 10 (Upper) 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 10 (Lower). The technique highlights the edges connected to the clicked nodes and the citations of the cited and citing nodes.

Figure 10 (Upper) 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 10 (Lower) connect with either blue or green edges.

We listed all the titles and figures (see Figures 8 and 9) of the papers classified into these two groups. The papers connected with green edges as follows:

- Continuous capture of skin deformation (in 2003)
- Building efficient, accurate character skins from examples (in 2003)
- Capturing and animating skin deformation in human motion (in 2006)
- Data-driven modeling of skin and muscle deformation (in 2008)

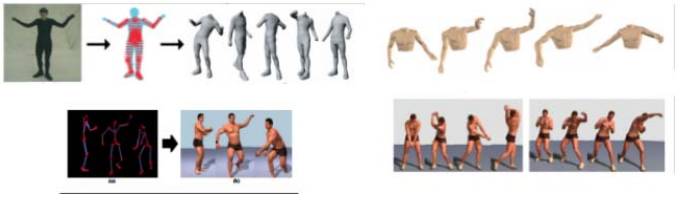

Figure 8 Pictures in papers of the green stream

We listed the papers that belong to the blue stream.

- Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin (in 2003)
- Analysis of human faces using a measurement-based skin reflectance model (in 2006)

Figure 9 Pictures in papers of the blue stream

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, 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.

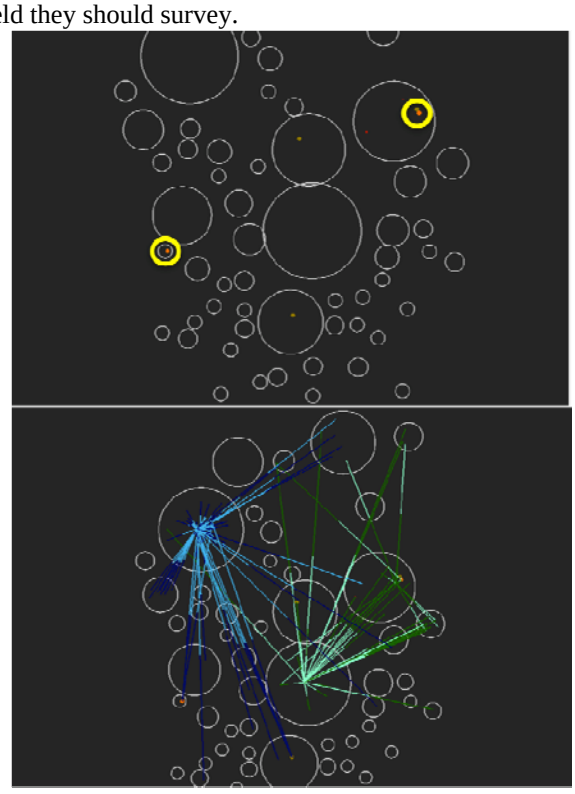

Figure 10 (Upper) Result with a keyword “skin”, (Lower) Result when an user click two nodes

5. Evaluation

5.1. Preliminary questionnaires

There have been a lot of citation visualization techniques as we mentioned in Section 2. As against our technique applies a general purpose graph layout technique, typical existing techniques places nodes corresponding to the papers in time-ordered. However, we assume the time-ordered layout policy is not mandatory, because it is sufficient in many use cases 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 the 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 suggests 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 as 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 papers.

Regarding the question 3, we roughly divide the answers into three categories, “the transition of research fields”, “the citation relationships”, and “both research fields and citation relations, or what they reveal in combination”. It demonstrates the demands to understand both research fields and citations.

5.2. 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 11 (called “A” in this section) with the time-oriented citation visualization shown in Figure 12 (called “B” in this section), 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 1 was a strong agreement with A, and 5 was a strong agreement with B. The following are questions we asked to participants:

1. The transition of papers amount published in the conference every year.
2. The main topic in 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.
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).

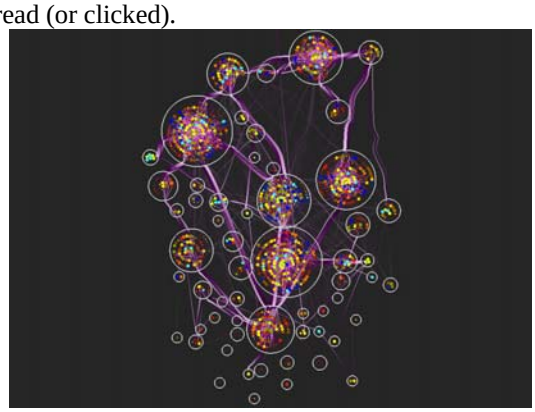

Figure 11 Our technique (A)

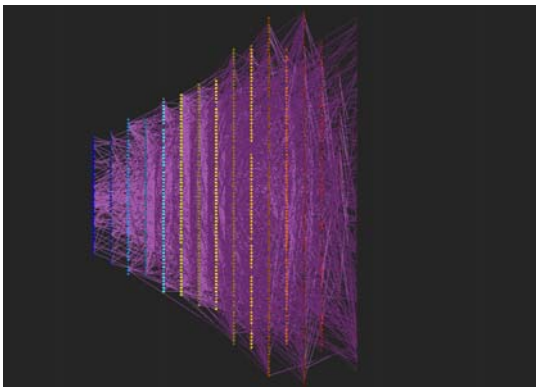

Figure 12 Time-oriented technique (B)

Figure 13 shows the evaluation result. The X-axis denotes the sequential number of questions, and the Y-axis denotes the quantity 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.

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

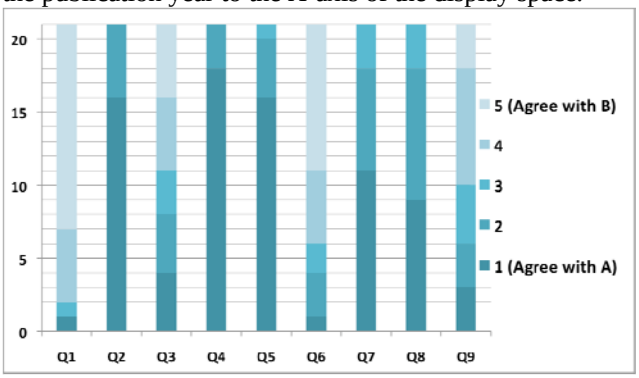

Figure 13 Result of the evaluation

Conclusions

We presented a visualization technique of citation networks for survey of research papers, and discussed the results. Our technique applies topic-based paper clustering to construct hierarchical network. It then applies a hybrid force-directed and space-filling network layout algorithm, and an edge bundling technique with Catmull-Rom spline curve. This paper also introduced the results and the user evaluation.

As a future work, we would like to apply the technique to larger citation datasets and conduct user evaluations. Also, we would like to also visualize more complex datasets combining co-citation and co-author networks. We expect it helps to understand the citation relations across multiple research fields more easily.

References

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

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

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

[4] H. Small, Visualizing science by citation mapping, Journal of the American society for Information Science, 50(9), 799-813, 1999.

[5] B. Lee, M. Czerwinski, G. Robertson, B.B. Bederson, Understanding research trends in conferences using PaperLens, CHI'05 extended abstracts on Human factors in computing systems (pp. 1969-1972), 2005.

[6] U. Brandes, T. Willhalm, Visualization of bibliographic networks with a reshaped landscape metaphor, the symposium on Data Visualisation 2002.

[7] J. Matejka, T. Grossman, and G. Fitzmaurice, Citeology: visualizing paper genealogy, CHI'12 Extended Abstracts on Human Factors in Computing Systems, pp. 181-190, 2012.

[8] 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, 63(12), 2351-2369, 2012.

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

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

[11] 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, 2012