Visual Exploration of a Series of Academic Conferences

Kazuo Misue
Faculty of Engineering, Information and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, Japan
Email:misue@cs.tsukuba.ac.jp
http://www.cs.tsukuba.ac.jp/˜misue/

***Abstract*—The trends in a research field, especially changes in the features over the years, are subjects of interest for many researchers. This paper reports an exploratory analysis of the changes of research topics in an academic field. The target data of the analysis are the author-keywords included in papers presented at a series of academic conferences, IEEE International Conference on Data Mining (ICDM). The analysis process consists of three phases: (1) frequency of keywords, (2) appearance of keywords in papers, and (3) relationships among keywords. In phase 1, bar charts were used to observe the ranking of frequencies. In phases 2 and 3, anchored maps were adopted. The anchored maps are based on the spring-embedder model, but they provide viewpoints by using fixed "anchors." The analysis process revealed the major topics in the field of data mining and some changes in the relationships among topics.**

***Keywords*-information visualization; visual exploration; graph drawing; academic conferences**

I. Introduction

The trends in a research field, especially changes in the features over the years are subjects of interest for many researchers. This paper reports an exploratory analysis of the changes of research topics in an academic field. Hypotheses are not prepared; instead, it was our aim to explore the data to discover interesting features.

Researchers unfamiliar with a field of research may find information about the features of the research field useful for improve their understanding of the field. Researchers who have been active in the field for many years may be familiar with such features. Nevertheless, these researchers may still benefit from a reconfirmation, especially if they previously overlooked a feature of their field of research.

The research involved an analysis of keywords that were used by the authors of papers presented at a series of academic conferences, IEEE International Conference on Data Mining (ICDM). Initially, we focused on the frequencies at which keywords appeared. However, frequencies do not provide useful information about the relationships between keywords. Our focus therefore was on networks representing the relationships between papers and keywords, and additionally on the relationships between keywords and the changes they undergo over the years. This was accomplished by adopting anchored maps to observe the networks. An anchored map is an information visualization technique based on a fundamental graph-layout method, the spring-embedder model, but provides viewpoints, i.e., anchors, that facilitate the observation of changes in the networks.

The contributions of this paper can be listed as follows:

1) Contributing to knowledge about changes in research fields related to data mining,
2) Understanding the extraction of graph structures from data of papers, and
3) Using anchored maps in exploratory analyses.

II. Related Work

A. Research Map

Many attempts have been made over the years to draw maps representing the topics of documents. Lin [1] proposed a method to visualize document space by using Kohonen's feature map. The method visually depicts a collection of documents on information retrieval. Wise et al. [2] presented ThemeScape, a visualization technique that provides 3D landscapes of topics and themes in the document corpora. Fujimura et al. [3] proposed a variation of large-scale tag clouds. Their method displays a landscape of relationships among tags (topics) of many documents. Fried and Kobourov [4] demonstrated an approach for the visual exploration of research papers.

Many researchers have attempted to visualize networks of authors and papers. Chen and Paul [5] aimed to visualize intellectual structures in a knowledge domain. They visualized author co-citation networks to represent knowledge landscapes. Elmqvist and Tsigas [6] presented CiteWiz, a platform for bibliographic visualization that provides the visualization of keyword and co-authorship networks. Perianes-Rodríguez et al. [7] proposed a method for detecting, identifying and visualizing research groups in co-authorship networks, which they visualized by using the Kamada-Kawai algorithm, a variation of the spring-embedder model. By focusing on relationships between authors and papers, a bipartite network can be obtained, for which Misue [8] presented a visual analysis tool that uses anchored maps to visualize networks involving academic papers and authors. Naud et al. [9] and Ito et al. [10] presented methods to visualize a bipartite network spherically in 3D space.

B. Anchored Map

An anchored map [11] is a node-link diagram based on the spring-embedder model [12]. Anchored maps were originally designed for drawing bipartite graphs (two-mode networks). The technique fixes the nodes in one of the sets, called "anchors," at predetermined positions on a circumference, while allowing the nodes in the other set, called "free nodes," to be arranged freely. Any graph can also be drawn as an anchored map by dividing the nodes into two sets, provided that edges with both of their endpoints in the same set are ignored. More specifically, an anchored map was visualized using two steps:

1) Arrange the anchors on the circumference at equal intervals prior to deciding the ordering of the anchors on the circumference.
2) Fix the anchors and use the spring embedder to position the free nodes such that they appropriately express their relationships to the anchors.

The number of edge crossings is strongly influenced by how the anchors are ordered. Deciding the ordering is therefore the most critical problem. The tool offers numeric indices as substitutes for the aesthetic criteria of graph drawing. Our work involved the use of the sum of the distances between every free node and the center of the circumferences, referred to as the eccentricity of the nodes, as an index.

III. Target Data

Data was obtained from papers that were accepted as regular papers for the ICDM between 2002 and 2013, with the condition that a paper had to contain one or more keywords. For example, a paper that was presented at ICDM 2012 contained six keywords: differential privacy, histogram, lossy compression, Fourier transform, and clustering.

The occurrence of orthographical variants required us to take some preparatory steps, including the conversion of all keywords to lowercase and the unification of some hyphenated words, for example, "time series" and "time-series", to remove variants. Singular and plural forms were not unified.

In our analysis, every regular paper was assigned an ID, which was used together with the year of publication and the keywords that were listed by the author. Our analysis did not use the paper title, abstract, author names, or affiliations.

IV. Frequency of Keywords

The total number of keywords that were found was 3058, and the number of unique keywords was 2048. The mean of the frequency of appearance of keywords was 1.49. There were 341 keywords that appeared more than once and 1707 keywords that appeared only once, although these numbers would depend on processing for orthographical variants.

Next, the frequency at which each keyword appeared between 2002 and 2013 was determined. Figure 1 displays

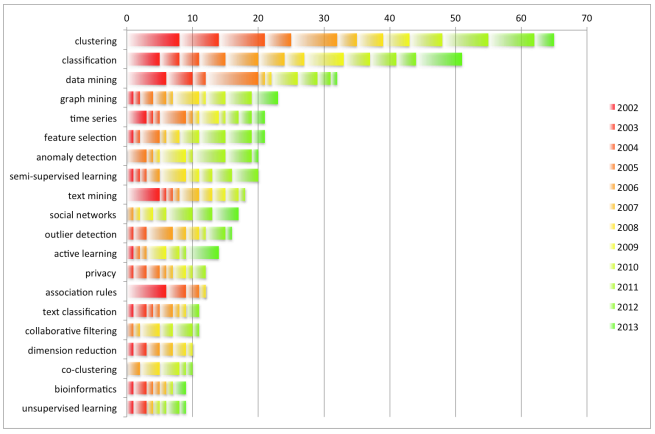


Figure 1. Frequencies of top 20 keywords.

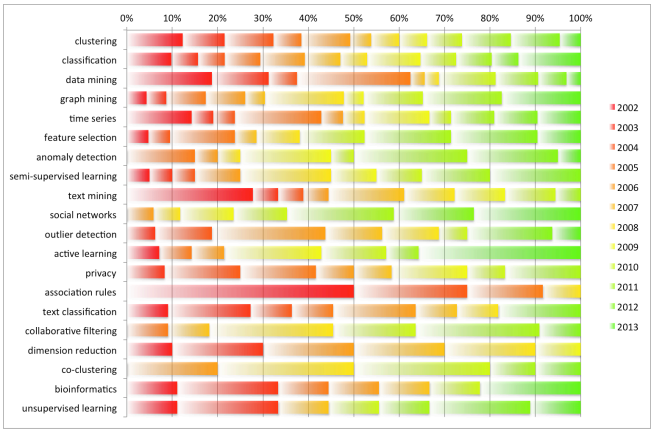


Figure 2. 100% stacked bars of top 20 keywords.

a stacked bar chart representing the frequencies of the top 20 keywords. The total length of each bar represents the cumulative total of 12 years. In the most of ordinal stacked-bar charts, layers are presented in significantly different colors to enhance the visibility of the borders of the layers. In Figure 1, however, we assigned a color along the gradation from red to green to each layer, because the layers represent years of ordinal scale. This caused the borders between layers to become unclear. We used another gradation consisting of white and a vivid color to clarify the borders.

During the 12 years under consideration, the keyword that was used most frequently was "clustering." It was followed by "classification" and "data mining" in the second and third places, respectively, although these two keywords were used much less frequently than "clustering". The frequency of use of "data mining" was about half that of "clustering." The fourth-most used keyword was "graph mining", followed by "time series" and "feature selection" in the fifth place and "anomaly detection" in the seventh place, etc.

Figure 2 presents the same data as a 100% stacked bar

chart, as this was considered to be a more suitable way of comparing the annual distribution of frequencies.

Two of the keywords, "clustering" and "classification," appeared with constant frequency. Against them, the frequency of the use of "data mining" was not constant; it was found to be low between 2006 and 2009. Although it is not possible to obtain more detailed information from Figure 2, many other keywords[1] including "data mining" were used between 2006 and 2009. This means that "data mining" was not used on its own. About half of the appearances of the keywords "graph mining" and "feature selection" were in papers that were published during the last three years under consideration; hence, it can be said that these two keywords have recently become more popular.

V. Appearance of Keywords in Papers

It would be possible to guess which topics were popular and to guess which topics were rising and declining in popularity by observing the changes in the appearance frequencies of keywords. However, this would not permit an understanding of the relationships between the topics. Instead, anchored maps were drawn to represent the relationships between keywords and papers. First, the top six keywords were selected from the ranking shown in Figure 1. The selected keywords were drawn as anchors in the anchored maps. There is no inevitability to the number six and the top 10 or top 20 could also have been chosen. However, too many anchors tend to cause unreadable diagrams. The keyword ranking in Figure 1 shows that the top three keywords were significantly frequent. One reasonable approach would have been to choose the top three keywords, but three keywords were considered to be too few to serve as observational targets; thus, six keywords were consequently chosen. Bipartite graphs were extracted that represent the relationships between these six keywords and those papers that contained at least one of these keywords at least once every three years. Figure 3 shows the anchored maps of the four bipartite graphs. The six keywords were drawn as anchors and the papers were drawn as free nodes. The year in which each paper appeared is represented by a color as in Figure 1.

2002-2004: Figure 3(a) reminds us of three major topics: "clustering," "data mining," and "classification." Moreover, a few papers contained pairs of keywords such as {"clustering" and "feature selection"}, {"clustering" and "data mining"}, and {"data mining" and "classification"}. On the other hand, "graph mining" was isolated.

2005-2007: As shown in Figure 3(b), "graph mining" continued to be isolated. Additionally, "classification" is also isolated. The number of papers using "classification" as a keyword was found to have increased, but these papers did not use any of the other top six keywords. The keyword "feature selection" is connected to "clustering" as they appeared together in a paper. Some of the other papers that used "feature selection" did not use any of the other top six keywords.

2008-2010: In Figure 3(c) it can be seen that "classification" is connected to "time series", although "graph mining" continues to be isolated. Additionally, the use of "feature selection" is also isolated.

2011-2013: In Figure 3(d) it is shown that a paper connecting "clustering" and "feature selection" has been revived. The keyword "graph mining" is also connected to another keyword, "clustering", for the first time. The keywords "clustering" and "classification" are also connected to each other for the first time. On the other hand, "data mining" is isolated. The figure seems to indicate a recent upsurge in the use of "feature selection."

Figures 3(a)–(d) all display changes of connectivity between the top six keywords. However, there are many papers containing the top six keywords in which two or more keywords are connected. Therefore, the changes displayed by these bipartite graphs would not provide enough information to allow a discussion of the features of the research field.

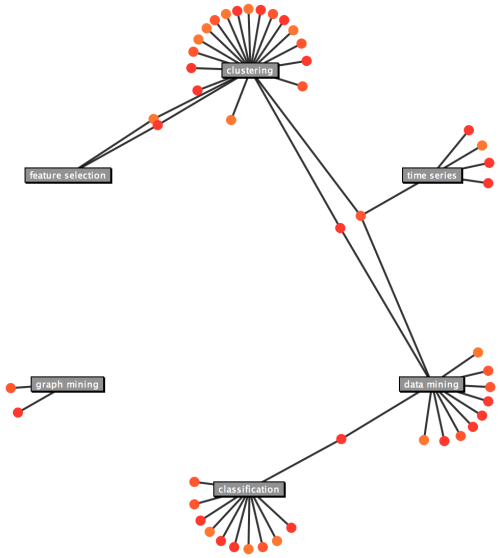
VI. Relationships between Keywords

As there are 2048 different keywords, it is important to also consider those other than the top six. Keyword relationships were studied by using the top six keywords as seeds. First of all, a graph was constructed with the nodes representing keywords and the edges representing keywords that co-occurred in the same paper. Next, a subset of nodes consisting of the top six keywords and all the keywords appearing alongside these keywords was selected. Using this subset of nodes, an induced sub-graph was then drawn in the form of anchored maps (Figure 5) in which the six keywords were drawn as anchors and the remaining keywords as free nodes.

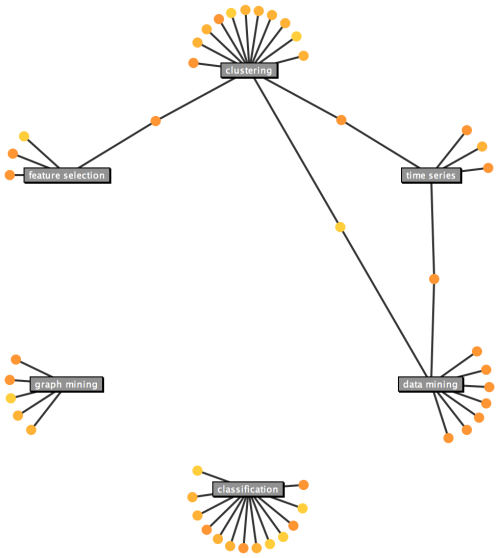
The period during which each keyword was in use is represented by a different color. The first year of each period is represented by a hue of color as in Figure 1, while the last year of the period is represented by saturation. For example, a keyword that first appeared in 2002 was assigned the color red, and if it was still used in 2013, the most vivid red was assigned. If it was not used after 2002, it was assigned a pale red color. The color map that was used is shown in Figure 4. Colors marked with "x" were not used.

2002-2004: Figure 5(a) shows that there were many keywords that were used in conjunction with "clustering." Even in this figure, "feature selection" is connected with "clustering" as shown in Figure 3(a). In Figure 5(a) it can be seen that the keyword "graph mining" is not isolated, for example, there are some paths that connect "graph

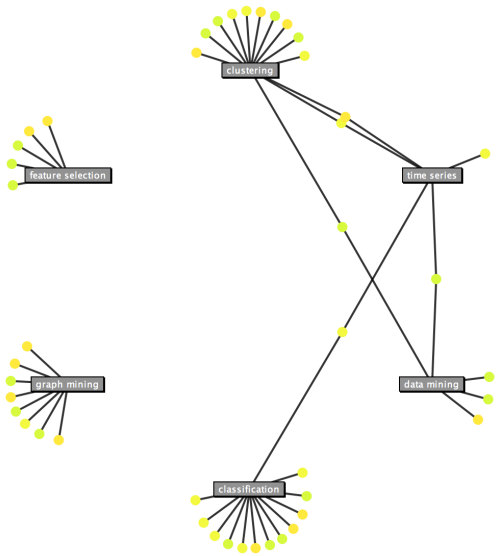
[1]E.g., data mining application, spatio-temporal data mining, high performance data mining, multimedia data mining, relational data mining, distributed data mining, temporal data mining, stream data mining, privacy-preserving data mining, etc.



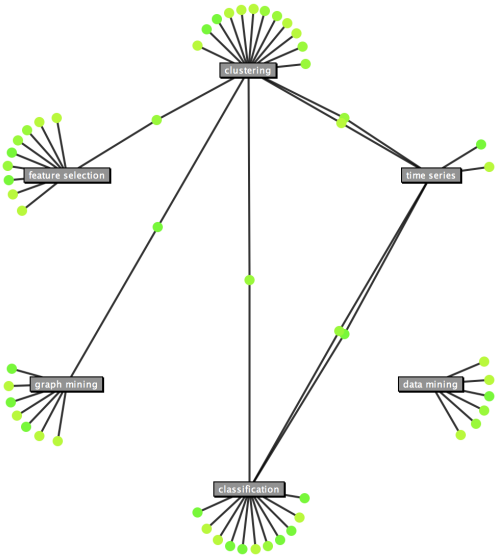
(a) 2002–2004



(b) 2005–2007

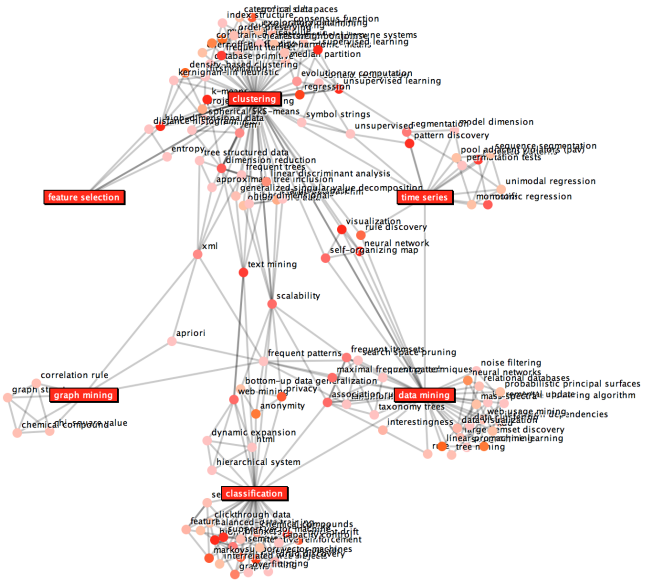


(c) 2008–2010

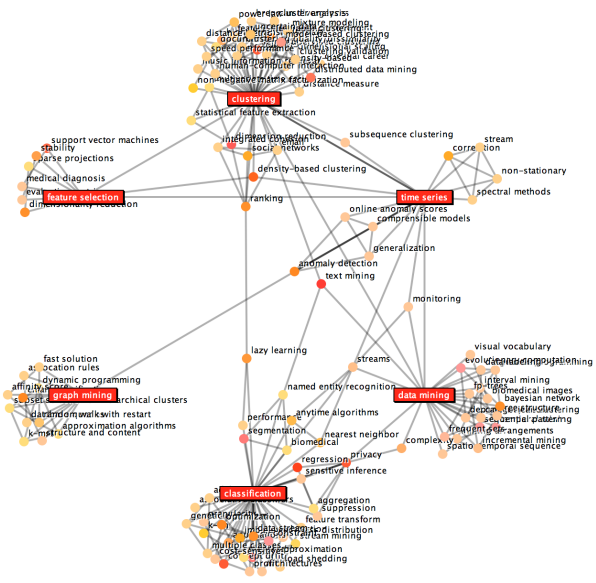


(d) 2011–2013

Figure 3. Anchored maps representing relationships between keywords and papers.

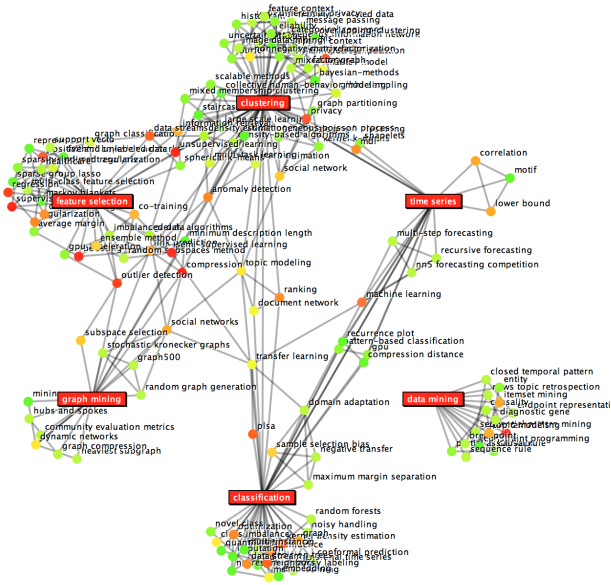


(a) 2002–2004



(b) 2005–2007

(c) 2008–2010



(d) 2011–2013

Figure 5. Anchored maps representing relationships between keywords.

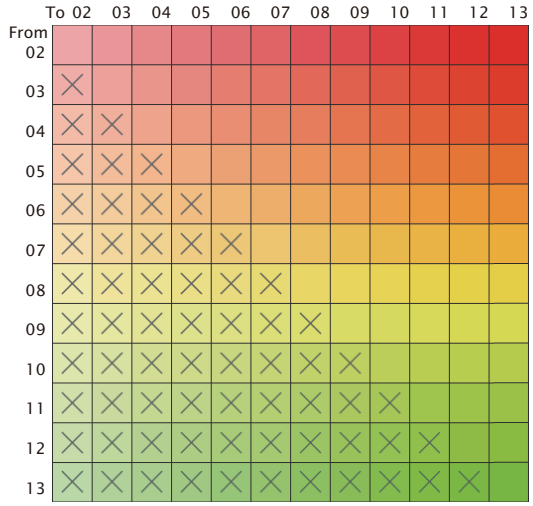


Figure 4. Color map for representing the terms during which keywords were used.

mining" with "clustering." Some keywords appear to be hubs, for example, "scalability," "text mining," and "xml", although the low saturation of the "xms" means it is no longer used. Other keywords such as "visualization," "neural network," and "self-organizing map" connect the keywords "clustering" and "data mining."

2005-2007: In Figure 5(b) it is shown that "feature selection" is connected not only with "clustering" but also with "time series." Moreover, the number of keywords that were only connected with "feature selection" increased. These keywords did not exist in 2002-2004 (in Figure 5(a)). This might mean that "feature selection" had grown as an independent topic. A hub-like keyword "text mining" exists in the figure. However, both "xml" and "scalability" seem to have disappeared.

2008-2010: In Figure 5(c), the keywords "feature selection" and "graph mining" are isolated. Keywords connecting "data mining" and "classification" have also disappeared. The keyword "bayesian" seems to be a pivot connecting "clustering" and "classification."

2011-2013: In Figure 5(d) drastic changes can be seen in the structure over this period. Keywords around "clustering" have been greatly revived. In addition, those around "feature selection" have also been reactivated. New keywords have also appeared around "graph mining" and have made connections with other keywords. The keyword "social networks" appears to be a pivot of some connections. Keywords such as "anomaly detection," "topic modeling," and "transfer learning" seem to be functioning as hubs. On the other hand, "data mining" appears to be becoming isolated, with some new keywords around it, but with connections to other keywords becoming thin.

Throughout Figure 5(a)–(d), new keywords can be seen to be born. Over the period of 12 years under consideration the global structure of the keyword network was found to have changed and changes in last three years of this period are more prominent than in previous years.

VII. Discussion

The process described above involves an exploratory analysis of data from academic papers. No hypotheses were prepared, but some trends were found in the data. Bar charts are useful to present the ranking of frequencies. As shown in Figure 1, a stacked bar chart was used to understand changes in frequencies. However, it is difficult to read changes in rankings from this chart. An attempt was made to understand relations by using network structures that were extracted from the data, because a series of anchored maps representing changes in the network structures obviously provides more information than bar charts.

The figures in this paper seem to provide an overview of a series of academic conferences; however, these figures do not reveal everything, as it is not possible to estimate information excluded from the figures. Figure 3 only includes papers containing one of the top six keywords, while other papers were excluded. This means that more than a few papers were excluded and that the feature changes caused by these papers cannot be presupposed. Future work will attempt to increase the number of seed keywords, i.e., the number of anchors with the aim of comparing the features that can be read from anchored maps. Figure 5 only focused on seed keywords and the keywords adjacent to them. In the future, we will focus on increasing the number of seeds and/or enlarging the set of target keywords by finding other relationship landscapes among keywords.

VIII. Conclusion

This paper reports an exploratory analysis of the research field of data mining. Keywords that were provided by authors of regular papers for a series of international conferences, the ICDM, were used as target data. First, the main research areas were identified by representing the appearance frequencies of keywords in the form of a bar chart. In this way three major topics were found: "clustering," "data mining," and "classification." Next, the changes in the frequency of each keyword were investigated, for example, the keywords "graph mining" and "feature selection" were found to have increased in popularity during the last years forming part of the study.

Moreover, graphs were extracted by using relationships between papers and keywords. This enabled us to develop an understanding of the features of the relationships among keywords by observing anchored maps of the extracted graphs. Our analysis only included the top six keywords; therefore, we would have to study other variations as well. Future work will include testing the validity of the features that were obtained from the anchored maps.

ACKNOWLEDGMENT

We thank Professor Shusaku Tsumoto, Shimane University, for providing the data from the ICDM papers and the opportunity to analyze them.

REFERENCES

[1] Xia Lin, Visualization for the Document Space, *in Proceedings of IEEE Conference on Visualization (Visualization '92)*, pp. 274–281, 1992.

[2] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow, Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents, *in Proceedings on Information Visualization (Infovis '95)*, pp. 51–58, 1995.

[3] Ko Fujimura, Shigeru Fujimura, Tatsushi Matsubayashi, Takeshi Yamada, and Hidenori Okuda, Topigraphy: Visualization for Large-scale Tag Clouds, *WWW 2008 / Poster Paper*, pp. 1087–1088, 2008.

[4] Daniel Fried and Stephen G. Kobourov, Maps of Computer Science, *in Proceedings of 2014 IEEE Pacific Visualization Symposium*, pp. 113–120, 2014.

[5] Chaomei Chen and Ray J. Paul, Visualizing a Knowledge Domain's Intellectual Structure, *IEEE Computer*, Vol. 34, No. 3, pp. 65–71, 2001.

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

[7] Antonio Perianes-Rodríguez, Carlos Olmeda-Gómez, and Félix Moya-Anegón, Detecting, identifying and visualizing research groups in co-authorship networks, *Scientometrics*, Vol. 82, Issue 2, pp. 307–319, 2010.

[8] Kazuo Misue, Visual Analysis Tool for Bipartite Networks, *in Proceedings of 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2008)*, *LNAI 5178*, pp. 871–878, 2008.

[9] Antoine Naud, Shiro Usui, Naonori Ueda, and Tatsuki Taniguchi, Visualization of documents and concepts in neuroinformatics with the 3D-SE viewer, *Frontiers in Neuroinformatics*, Vol. 1, Article 7, 2007.

[10] Takao Ito, Kazuo Misue, and Jiro Tanaka, Sphere Anchored Map: A Visualization Technique for Bipartite Graphs in 3D, *in Proceedings of 13th International Conference on Human-Computer Interaction (HCI International 2009)*, *LNCS 5611*, pp. 811–820, 2009.

[11] Kazuo Misue, Anchored Map: Graph Drawing Technique to Support Network Mining, *IEICE Transactions on Information and Systems*, E91-D, No. 11, pp. 2599–2606, 2008.

[12] Peter Eades, A Heuristic for Graph Drawing, *Congressus Numerantium*, Vol. 42, No. 11, pp. 149–160, 1984.