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

Rina Nakazawa[1,2] , Takayuki Itoh[1] and Takafumi Saito [3]

[1]Ochanomizu University, Japan
[2]IBM Research - Tokyo , Japan
[3]Tokyo University of Agriculture and Technology , Japan

Abstract

We usually use a text-based search engine while surveying research papers. Such search systems have difficulties for novice researchers in case they do not know appropriate keywords or do not understand the positions of papers. Many visualization tools of citation networks have been proposed to help this task. These tools demonstrated that not only text information of papers but citation relationships and co-author relationships also are helpful clues for research survey. We propose CoCoa, a linked network visualization of co-citation and co-author relationships for surveying research papers. Our system visualizes both citation and co-author networks at the same time. To make comparison and grasp of correspondence between co-citation and co-author networks easier, the system treats both a paper and an author as bags of words and cluster them into topics applying LDA (Latent Dirichlet Allocation) at the same time. Based on the clustering result, it places the clusters of a citation network by a hybrid force-directed and space-filling algorithm. The position of topic clusters in the networks would have an influence on the correspondence of a particular topic in the networks. Our system extracts the clusters which consist of the common combinations of topics in two networks. Then it reuses the positions of the clusters in a citation network as the initial cluster positions of a co-author network, supposing there are a large number of authors.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** *Visualization; Information visualization;*

1. Introduction

For a long time, many researchers have studied recommendation and visualization techniques for scholarly literature to help a survey of research papers [MRC95] [Sma99]. Federico et al. [FHKM17] reported a large number of visual approaches to scholarly literature and categorized them according to their data type and tasks. In recent years, Huang et al. [HWC*15] and Ebesu et al. [EF17] use neural network models for citation recommendation. A scholarly search engine like Google Scholar [Sch] recommends articles to users. In terms of understanding the detailed recommendations or search results, researchers visualize publication data in many ways [BKW16] [HHKE16] [RHB*17] [LB19]. These studies often 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 [KBV04] [HGEF07]. Nakazawa et al. [NIS18] proposed a visualization technique of a citation network applying topic-based clustering. The technique categorizes papers based on topics of their abstracts and visualizes a citation network applying this categorization to help users to find papers in the same clusters that include similar words.

However, few studies develop combinational visualizations of citation and co-author networks. PivotPaths [DRRD12] 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.

DOI: 10.2312/evs.20191179

Our goal is to help users to understand relationships among research fields, find influential authors in the research fields, and survey research papers casually. Thus, this paper 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.

2. Scenario

We first describe our target users as follows. The users we expect are not familiar with the research field. They are expected to have the following characteristics:

- They do not know all appropriate query keywords in that research field
- They do not know influential researchers in the research field
- 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 [HKPS18] 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.

3. Related Work

In this section, we introduce the existing techniques of visualizing scholarly literature. One of the goals of this study is understanding the trends of research topics. Citation network is one of the popular resources of data to help a survey of scholarly literature as many researchers presented visualization techniques of the network [SCH*13] [MGF12] [SGH12] [RCB16]. In addition to citation relationships, some researchers utilized text data of scholarly literature. Citespace II [Che06] aims to visualize evolution and trend of research topics using hybrid networks of co-cited articles and terms citing the articles. Berger [BMS17] embedded citation relationships into text visualization. Another goal is to help the literature review. Some papers proposed visualization systems to help this task using citation and text data of literature [WLQ*16] [PEM16]. In this paper, we combine text information of papers, citation relationships, and co-author relationships.

4. Proposed System

This section describes an overview of our system CoCoa and its processing flow. Shown in Figure 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 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 [CS07], this interaction presents relationships between two networks.

4.1. Clustering research papers and authors

We employ LDA (Latent Dirichlet Allocation) [BNJ03] to cluster papers and authors. LDA is generally used for topic estimation of documents because it allows a document to include multiple topics and it can avoid over-fitting the data. We extract words from the abstracts of papers to apply the topic model and represent the papers as bags of words. Meanwhile, we define that 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 $S_k = \{w_1, w_2, \cdots, wm\}$. We describe the i-th author a_i, who wrote n papers $p_1, p_2, ..., p_n$, as follows:

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

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.

4.2. 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

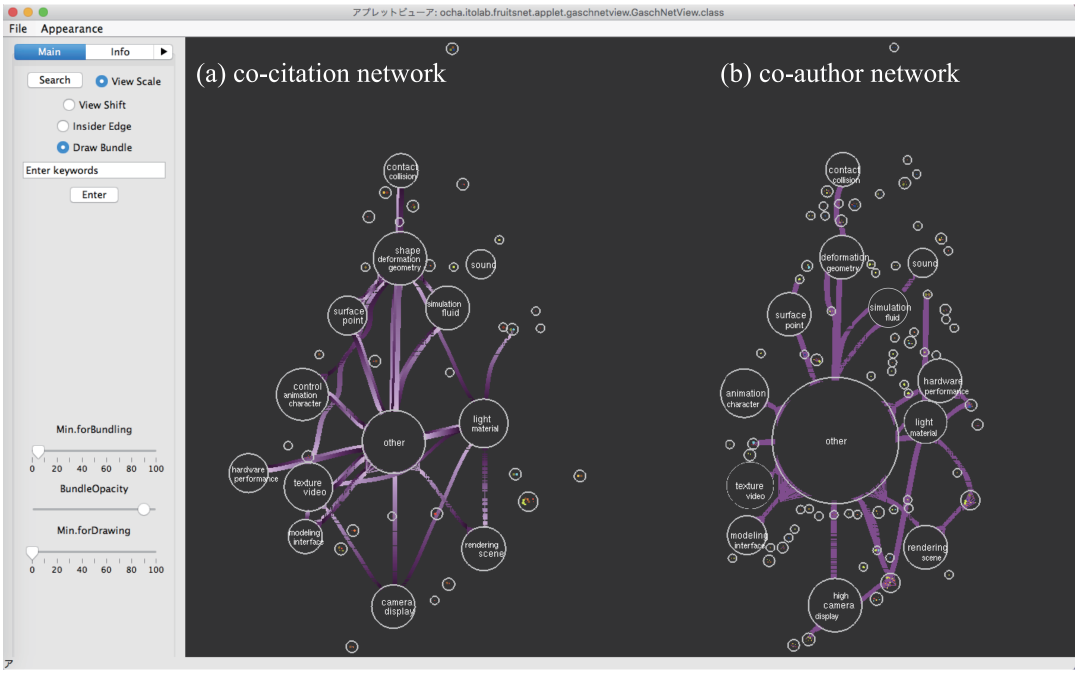

Figure 1: *The overview of the system. The left view (a) shows a co-citation network and the right one (b) shows a co-author network. Each circle represents a cluster of nodes and labels of clusters indicate their topics. When a user zooms in, the labels disappear and the nodes of papers and authors appear.*

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. Another technique is an auto topic labeling technique [MSZ07], which generates labels by combining words with more importance. This technique may generate the same label candidates. Since visualizing appearance frequency of representative words requires larger display space, we visualize topics of clusters using 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.

4.3. Network Layout

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 [NIS18] based on the result of the clustering described in Section 4.1. 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 [ELMS91]. 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 [NIS18]
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 [Hol06].

4.4. 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. Result

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. Note that this dataset does not include papers whose abstract 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 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 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 2(a). The papers are categorized into topics *modeling and interface*, *deformation and geometry*, and *others*. Compared to co-author relationships, there was no publication related to *texture and video* topics. A wide bundle between these two clusters in Figure 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 multiple research fields by both citation and co-author relationships.

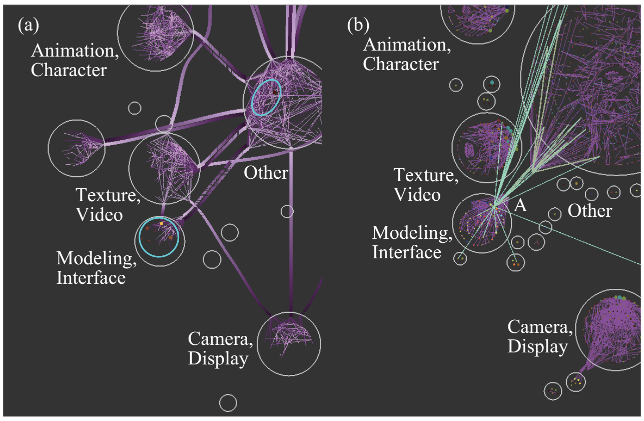

Figure 2: *When selecting modeling and interface topic first, our system shows the clusters including the topic in the center of each view. The green highlighted edges denote the co-author relationships of the clicked node* **A** *in the view of (b). The system filters the nodes in another view (a) whether 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).*

6. Discussion and Conclusion

Both a co-citation network and a co-author network help search for scholarly literature. We 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. We expect such a linked placement allows users to compare them easily. Thus, our system would help novice researchers to understand the relationships among research fields and find desired authors or papers.

As for the 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. We would like to handle much larger dataset including multiple conference papers 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.

References

[BKW16] Beck F., Koch S., Weiskopf D.: Visual analysis and dissemination of scientific literature collections with survis. *IEEE Trans-*

actions on Visualization and Computer Graphics 22 (2016), 180 – 189. 1

[BMS17] BERGER M., MCDONOUGH K., SEVERSKY L.: cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 691–700. 2

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), pp. 993–1022. 2

[Che06] CHEN C.: Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology 57*, 3 (2006), 359–377. 2

[CS07] COLLINS C., SHEELAGH C.: Vislink: Revealing relationships amongst visualizations. *IEEE Transactions on Visualization and Computer Graphics 6* (2007), 1192–1199. 2

[DRRD12] DÖRK M., RICHE N. H., RAMOS G., DUMAIS S.: Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2709–2718. 1

[EF17] EBESU T., FANG Y.: Neural citation network for context-aware citation recommendation. In *In Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval* (2017), pp. 1093–1096. 1

[ELMS91] EADES P., LAI W., MISUE K., SUGIYAMA K.: *Preserving the mental map of a diagram*. Tech. rep., Technical Report IIAS-RR-91-16E, Fujitsu Laboratories, 1991. 3

[FHKM17] FEDERICO P., HEIMERL F., KOCH S., MIKSCH S.: A survey on visual approaches for analyzing scientific literature and patents. *IEEE transactions on visualization and computer graphics 23*, 9 (2017), 2179–2198. 1

[HGEF07] HENRY N., GOODELL H., ELMQVIST N., FEKETE J. D.: 20 years of four hci conferences: A visual exploration. *International Journal of Human-Computer Interaction 23*, 3 (2007), 239–285. 1

[HHKE16] HEIMERL F., HAN Q., KOCH S., ERTL T.: Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics 1* (2016), 190–199. 1

[HKPS18] HOSSAIN M. I., KOBOUROV S., PURCHASE H., SURDEANU M.: Rematch: Research expert matching system. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)* (2018). 2

[Hol06] HOLTEN D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions On Visualization And Computer Graphics 12*, 5 (2006), 741–748. 3

[HWC*15] HUANG W., WU Z., CHEN L., MITRA P., GILES C. L.: A neural probabilistic model for context based citation recommendation. In *AAAI* (2015), pp. 2404–2410. 1

[KBV04] KE W., BORNER K., VISWANATH L.: Major information visualization authors, papers and topics in the acm library. In *IEEE Symposium on Information Visualization* (2004). 1

[LB19] LATIF S., BECK F.: Vis author profiles: Interactive descriptions of publication records combining text and visualization. *IEEE transactions on visualization and computer graphics 25*, 1 (2019), 152–161. 1

[MGF12] MATEJKA J., GROSSMAN T., FITZMAURICE G.: Citeology: visualizing paper genealogy. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (2012), pp. 181–190. 2

[MRC95] MACKINLAY J. D., RAO R., CARD S. K.: An organic user interface for searching citation links. In *the SIGCHI conference on Human factors in computing systems* (1995), pp. 66–73. 1

[MSZ07] MEI Q., SHEN X., ZHAI C.: Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007). 3

[NIS18] NAKAZAWA R., ITOH T., SAITO T.: Analytics and visualization of citation network applying topic-based clustering. *Journal of Visualization 21*, 4 (2018), 681–693. 1, 3

[PEM16] PONSARD A., ESCALONA F., MUNZNER T.: Paperquest: a visualization tool to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2016), pp. 2264–2271. 2

[RCB16] RENOUST B., CLAVER V., BAFFIER J.-F.: Flows of knowledge in citation networks. In *International Workshop on Complex Networks and their Applications* (2016), pp. 159–170. 2

[RHB*17] RIND A., HABERSON A., BLUMENSTEIN K., NIEDERER C., WAGNER M., AIGNER W.: Pubviz: Lightweight visual presentation of publication data. In *Proc. Eurographics Conf. Visualization (EuroVis)–Short Paper* (2017). 1

[Sch] SCHOLAR G.:. In http: //scholar.google.co.jp/ . 1

[SCH*13] STASKO J., CHOO J., HAN Y., HU M., PILEGGI H., SADANA R., STOLPER C. D.: Citevis: Exploring conference paper citation data visually. In *IEEE Information Visualization Conference (Poster Session)* (2013). 2

[SGH12] SHAHAF D., GUESTRIN C., HORVITZ E.: Metro maps of science. In *the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), pp. 1122–1130. 2

[Sma99] SMALL H.: Visualizing science by citation mapping. *Journal of the American society for Information Science 50*, 9 (1999), 799–813. 1

[WLQ*16] WANG Y., LIU D., QU H., LUO Q., MA X.: A guided tour of literature review: Facilitating academic paper reading with narrative visualization. In *In Proceedings of the 9th International Symposium on Visual Information Communication and Interaction* (2016), pp. 17–24. 2