A Visualization of Relationships Among Papers Using Citation and Co-citation Information

Yu Nakano, Toshiyuki Shimizu, and Masatoshi Yoshikawa

Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
ynakano@db.soc.i.kyoto-u.ac.jp,{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

Abstract. When we conduct scholarly surveys, we occasionally encounter difficulties in grasping the vast amount of related papers. Because academic papers have relationships, such as citing and cited relationships, we considered utilizing them for supporting scholarly surveys. In this paper, we propose a method for visualizing relationships among papers, and we construct *paper graphs* using two types of relationships, namely, citation and co-citation. Moreover, we quantify the strengths of citations and co-citations based on their frequency and the positions of co-citations, and show both types of relationships together in a graph. We constructed paper graphs using papers in the database field and discussed their usefulness.

Keywords: scholarly survey, co-citation analysis, citation graph

1 Introduction

Researchers examine academic papers related to their research field and acquire knowledge for their own research. This process is called scholarly surveys, and many researchers use academic search engines, such as Google Scholar[1]. Because the number of academic papers has recently been increasing, it is impossible to read all the related papers. Therefore, how efficiently and comprehensively we understand them is one of the problems in scholarly surveys.

A possible approach to overcome this issue is to visualize relationships among papers[1][2]. Understanding relationships among papers makes it easy for researchers to grasp the insistence of each paper and to obtain insights into their research field. Therefore, analyzing and visualizing relationships among papers supports scholarly surveys. In this paper, we show a graph of relationships among papers, aiming to help researchers conduct scholarly surveys more efficiently.

There are many types of research that use citations to analyze relationships among papers[3][4]. This is because by citing their related papers, researchers describe the similarities and differences of their research and make its contributions clear when writing papers; thus, in this paper, we use citations as it was done in previous research.

[1] https://scholar.google.co.jp

In addition to citations, we also focus on co-citations. A co-citation is a situation in which two papers are cited in the same paper. Some research indicates that co-citation provides relationships such as similarities among papers[3][5].

After we extract relationships among papers, we have to consider how understandably we should show the relationships. There are many types of research about visualizing relationships among papers by using citations, which is known as a citation graph[1][2][6]. In this paper, we considered visualizing relationships among papers by using not only citations but also co-citations. We constructed *paper graphs* using the frequency of citations, frequency of co-citations, and positions of co-citations. By obtaining information on citations and co-citations simultaneously, we believe that researchers can understand their research field more effectively.

2 Relationships Among Papers

We visualize the relationships of a given set of papers using a directed graph. We call this directed graph a *paper graph*. We assume that a given set of papers is related to each other, such as the search results of academic search engines. In paper graphs, a node represents a paper, and an edge is a relationship between two papers.

In this paper, we utilize citations and co-citations as relationships among papers. After quantifying the strengths of the two types of relationships, we visualize both of them in the same graph. We can observe citation information and co-citation information from the paper graph simultaneously. From the citation information, we can identify a paper cited from many other papers. Similarly, from the co-citation information, we can grasp how strongly papers are related.

In this section, we describe which papers we should connect in the given paper set considering citations and co-citations and its strength. We also describe how to arrange nodes when visualizing paper graphs.

2.1 Strength of Edge

In this section, we explain which edges we show in paper graphs based on citing and cited relationships in the papers. We assume that two papers are related if they have a citation relationship or co-citation relationship and we describe how to visualize these two types of relationships in paper graphs.

In Case of Citation Two papers that have a citing and cited relationship are related, but one paper typically cites many papers; therefore, if we show all edges of citations in a paper graph, it becomes too complicated to describe the graph. Therefore, we show the edge of citations if a cited paper has a strong relationship to a citing paper. We considered that there is a strong relationship between two papers if a paper cites another paper in the text many times.

We quantify the strength of citing and cited relationships based on the frequency of citations. Let the (i, j)th entry m_{ij}^{cite} in matrix M^{cite} be the strength of a citing and cited relationship between paper p_i and paper p_j; we define m_{ij}^{cite} as follows.

$$m_{ij}^{cite} = \frac{\text{citation frequency of } p_j \text{ in } p_i}{\text{total citation frequency in } p_i} \tag{1}$$

If m_{ij}^{cite} is greater than the threshold $\alpha(i)$, then we connect an edge of citation from p_i to p_j. The threshold α is a function, as follows.

$$\alpha(i) = \text{the value of top } r_\alpha\% \text{ of } i\text{-th row of } M^{cite} \tag{2}$$

The variable r_α is a parameter. When $r_\alpha = 100$, in the paper graph, there are all edges of citations in the given papers. As r_α becomes smaller, there are fewer edges of citations in the paper graph, and there are no edges of citations when $r_\alpha = 0$.

In Case of Co-citation Two papers that have co-citation relationships are related, but showing all co-citations in a paper graph has the same problem as with citations; thus, we show strong relationships even in this case.

When quantifying the strength of co-citations, we can utilize the positions of co-citations. Eto[5] calculates similarities between two papers using the positions of co-citations, and he shows that the closer the positions where two papers are cited, the more similar the two papers are.

We attempted to quantify the strength of co-citation relationships based on the frequency and the positions of co-citations. Let the (i, j)th entry m_{ij}^{cocite} in matrix M^{cocite} be the strength of co-citation relationships between p_i and p_j, and let P be a given set of papers that cite both p_i and p_j; then, we define m_{ij}^{cocite} as follows.

$$m_{ij}^{cocite} = year\_coef_{ij} \times \sum_{p_x \in P} cocite\_pos(i, j) \text{ in } p_x \tag{3}$$

In this definition, $cocite\_pos(i, j)$ in p_x is the positions of two papers that have co-citation relationships, and we define the following formula based on Eto[5].

$$cocite\_pos(i, j) = \begin{cases} 1 & \text{(enumeration)} \\ 0.75 & \text{(same sentence)} \\ 0.25 & \text{(same section)} \\ 0 & \text{(across sections)} \end{cases} \tag{4}$$

If there are multiple co-citations in one paper, then we regard the position of them as the closest one and ignore other co-citations for simplicity.

The definition of $year\_coef_{ij}$ is a coefficient, as follows.

$$year\_coef_{ij} = \left(\frac{year_i + year_j - (start - 1) * 2}{interval * 2} \right)^2 \tag{5}$$

Here, $year_i, year_j$ are the publication years of p_i, p_j, respectively; $start$ is the earliest year in the given papers; and $interval$ is the difference between the last year and earliest year in a given paper set. The value of $year\_coef$ becomes larger for newer papers. The intuition of the formula (3) is that if two papers that have co-citation relationships are new, then they are more related than older ones which have the same frequency and the positions of co-citations. The reason why we introduce $year\_coef$ is that older papers tend to have more co-citation relationships because of its nature.

If m_{ij}^{cocite} is greater than the threshold β, then we connect an edge of co-citation between p_i and p_j. The threshold β is as follows.

$$\beta = \text{the value of top } r_\beta\% \text{ of non-zero elements of } M^{cocite} \tag{6}$$

The variable r_β is a parameter. When $r_\beta = 100$, in the paper graph, there are all edges of co-citations in the given papers. As r_β becomes smaller, there are fewer edges of co-citations in the paper graph, and there are no edges of co-citations when $r_\beta = 0$.

2.2 Arrangement of Nodes

When observing paper graphs, it is difficult to obtain useful information if the nodes in the paper graphs are disordered. We arranged papers in paper graphs in chronological order. This helps researchers estimate the history of their research topic. This method is generally used in research of visualizing citations[1][2][6].

3 Preliminary Experiment

3.1 Dataset

We constructed paper graphs using the proposed method described in Section 2. The outline is presented below.

1. define a group of papers as a dataset D
2. retrieve papers D_q by searching with a query q we choose on Google Scholar
3. select target papers D_t which are included in both D and D_q
4. construct a paper graph of the top-k papers of citation count in D_t

The dataset D we used is made of papers published in SIGMOD[2], VLDB[3], and ICDE[4] from 2000 to 2015. The reason why we selected these conferences is because they are top conferences in the database field and papers published there are expected to be strongly related, which is a suitable situation to obtain relationships among papers.

We extracted 201,404 citations and 1,664,014 co-citations from 6,977 papers in the dataset. Citations and co-citations that both of two papers are in the dataset are 47,716 and 100,355, respectively. We used ParsCit[7] to extract citing and cited relationships.

Using this dataset, we constructed paper graphs of three queries, namely, “skyline”, “top-k queries” and “uncertain data”. We set the parameters k, r_α and r_β to various values. Parameter k is the value described in the outline of the method, that is, the top-k papers of citation count in retrieved papers. Parameters r_α and r_β are the values appearing in the formulas (2) and (6), respectively. We used Graphviz[5] to visualize the paper graphs.

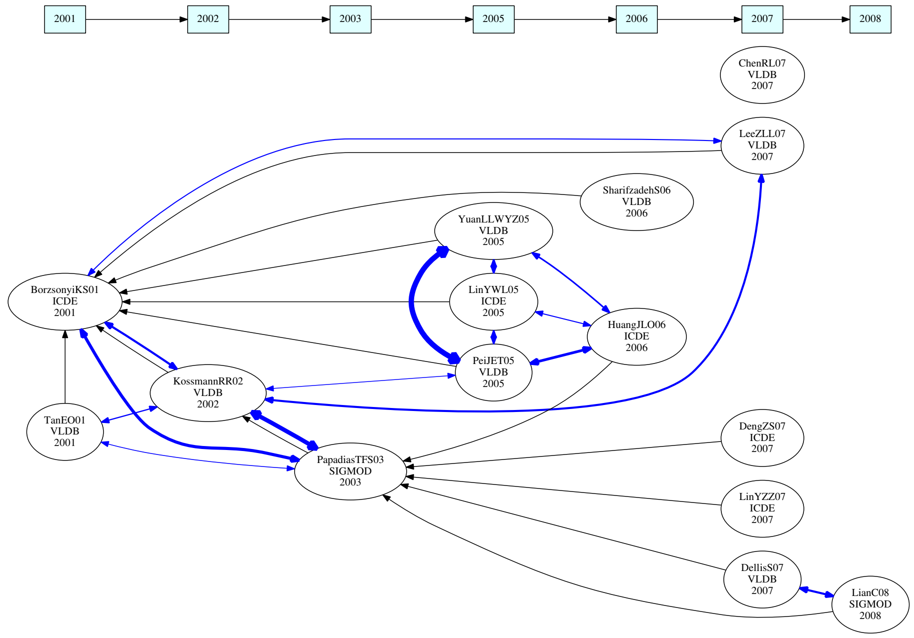

Fig. 1. $q =$“skyline”, $k = 15, r_\alpha = 5, r_\beta = 20$

3.2 Results and Discussion

Figure 1 is a paper graph of a query “skyline” with parameters $(k, r_\alpha, r_\beta) = (15, 5, 20)$. In this figure, one node is one of the papers in the dataset and it has information such as its ID, published conference and published year. While black edges represent citation relationships, blue edges represent co-citation relationships. In other words, a black edge from node A to node B means that paper A cites paper B, and a blue edge between two papers means that the two papers are cited together in another paper. Moreover, the width of edges indicates the strength of relationships.

From this figure, for example, we can understand the fact that because papers BorzsonyiKS01 and PapadiasTFS03 are frequently cited by other papers, the two papers strongly affect other papers. When looking at the edges of co-citations, we can estimate that because the four papers YuanLLWYZ05, LinYWL05, PeiJET05, and HuangJLO06 or the three papers BorzsonyiKS01, KossmannRR02, and PapadiasTFS03 are connected to each other by blue edges, they form one cluster of similar topics. We can observe these two results by focusing only on either of the two relationships, but from Figure 1, we can find out which pa-

[2] http://www.sigmod.org/
[3] http://www.vldb.org/
[4] http://www.icde.org/
[5] http://www.graphviz.org/

pers affect a cluster and how two clusters influence each other. We observe this advantage in paper graphs of the other two queries as well.

As r_α/r_β increase, the number of edges of citations/co-citations increase. While the increase of the edges of citations allows us to examine relationships among papers in more detail because more edges are connected to one node, the increase of the edges of co-citations allows us to observe paper graphs in a larger scale because the cluster size becomes larger. However, the increase of the edges makes a paper graph complicated; thus we need to adjust the parameters according to how much detail we want to observe in the paper graph.

4 Conclusion and Future Work

In this paper, to help researchers understand relationships among papers and support efficient scholarly surveys, we proposed the method of constructing *paper graphs* considering both citations and co-citations. For this purpose, we described some information that we can use to construct paper graphs, such as citation frequency, co-citation frequency, and the positions of co-citations. Additionaly, we attempted to quantify the strength of the two relationships. Moreover, we actually applied our method to papers published in the database field, and we discussed the advantages and disadvantages of the proposed visualization, that is, visualizing both citations and co-citations.

There are some directions for future work, that is, evaluations of the proposed method such as the strength of relationships that we quantified; the use of more papers published in other conferences; improvement of visualizing paper graphs; and so forth. Although we used citation frequency, co-citation frequency, and the positions of co-citations, it is worth considering other information, such as the position of citation, citation contexts, and citation functions[8]. To fully understand relationships among papers, visualization of summaries of citation contexts will be another improvement of paper graphs.

References

1. Shogen, S., Shimizu, T., Yoshikawa, M.: Enrichment of academic search engine results pages by citation-based graphs. In: AIRS. (2015) 56–67
2. Nanba, H., Abekawa, T., Okumura, M., Saito, S.: Bilingual presri-integration of multiple research paper databases. In: RIAO. (2004) 195–211
3. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science **24**(4) (1973) 265–269
4. Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: IJCAI. (1999) 926–931
5. Eto, M.: Evaluations of context-based co-citation searching. Scientometrics **94**(2) (2013) 651–673
6. Shahaf, D., Guestrin, C., Horvitz, E., Leskovec, J.: Information cartography. Commun. ACM **58**(11) (2015) 62–73
7. Councill, I.G., Giles, C.L., Kan, M.: Parscit: an open-source CRF reference string parsing package. In: LREC. (2008)
8. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: EMNLP. (2006) 103–110