Academic Landscape Based on Network Analysis Considering Analysis of Variation in the Years of Lucubration Publishing

Akira Otsuki and Ayumi Kawakami
Ochanomizu University, Tokyo
Japan

1. Introduction

Recent development in the field of Academic Landscape consists mainly in quotation network analysis. It is possible to sort tens of thousands of academic papers into clusters automatically by making full use of the various techniques that we describe in Section 2. However, it is not yet possible to automatically define the characteristics of each field identified by clustering as well as to automatically extract major papers in each field. Thus, We research the method to automatically identify major papers in each field recognized by clustering. Specifically, We examine the variance of years when a paper was quoted; then we apply the variance to the page rank algorithm to calculate the importance of a paper. We also aim to build up a hiangle map with a temporal axis based on the importance.

Finally, We describe the constitution of this paper. Section 2 surveys the preceding studies of network analysis, clustering and bibliometrics that are element technologies to realize Academic Landscape. Section 3 describes the purpose of this study and proposed method. Furthermore, Section 4 refers to the result of experimental evaluation of the proposed method before the conclusion in Section 5.

2. Related studies

Academic Landscape is kind of network analyses. Network analyses have long history. Network analysis is based on graph theory [1] by Leonhard Euler in the 18th century. There are a various analysis types to Network analysis. For example, "Complex Networks", "Network Optimization", "Small World Phenomenon", „Analysis of the degree distribution", "Clustering". Currently, the most important technology is "Clustering" in the Academic Landscape.

Clustering is the method to divide data into clusters. Clustering can be a simplify the structure of the vast amounts of data by having a common feature of the each clusters. The initial clustering algorithm was focused on the central link, and the common method was disconnect the center link to the first, then disconnect the around links to the second. On the other hand, Girvan and Newman [2-4] proposed a new algorithm that uses the modularity as an Evaluation function. Modularity is focused on links that mediate between the best cluster and disconnect the link from the link-mediated higher.

Then, Bibliometrics is a method developed by Garfield and Price. Bibliometrics is a method to support the resarch activity. It does analysis of academic papers or patents, Then, It can be understand that: "What is hot research topics?", "Which a large number of cited papers?", "Which are important papers?", "Which is related to that area?", "Who is important researcher?", "Which is the important research institute?"
There are three methods of Bibliometrics analysis (Fig. 1.).
The first is the "Direct Citation". It is regarded there is a link between paper A and paper B, If paper A is cited in the Paper C. In this case, there are 2 nodes and 1 link in the network. Papers are regarded as having any link between the papers themselves has been cited, when using a"Direct Citation".
The second is the " Co-Citation". It is the method proposed by Small [5]. Then it is regarded there is a link between paper A and paper B, If paper A and paper B is cited by paper C. In this case, there are 2 nodes and 1 link in the network. Be considered that there are a links between a pairs of those papers that all papers are listed in the bibliographies.
The third is the " Bibliographic Coupling". It is the method proposed by Kessler [6]. Then it i regarded there is a link between paper D and paper E, If paper C is cited in the Paper D and paper E. In this case, there are 2 nodes and 1 link in the network. Be considered that there are links for all pairs of papers that cite.

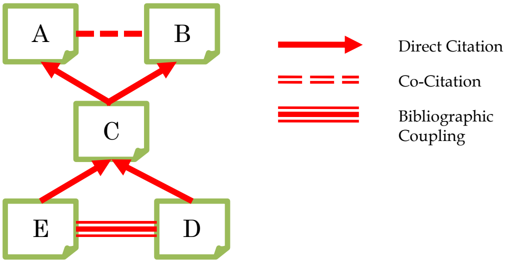


Fig. 1. The kinds of citation analysis

It became possible to analyse the academic landscape by combining these studies. Specifically, Each phase of the "Build of citation networks", "Getting the largest connected components", "Clustering" and "Visualization" will be processed automatically. But each area cannot automatically interpret. It is currently being analysed by the experts (Fig. 2).

3. Methods

In this study, to consider about automatic extraction of major paper in the each area to solve the issue of previous section. To calculate the severity by the traditional citation analysis is difficult, because shall be interpreted differently even if the same number of citations. For example, the cases that have been cited in many papers during the same period, and, the cases that cited in the long-term.
Therefore, in this study we consider a digraph assuming a paper is a node and a quotation is an edge for each "case" mentioned above. Then, We try to calculate the importance of each node by examining the variance of release years of source nodes whose edges enter into a node after allocating release years to each node. And we aim to establish a hiangle map with

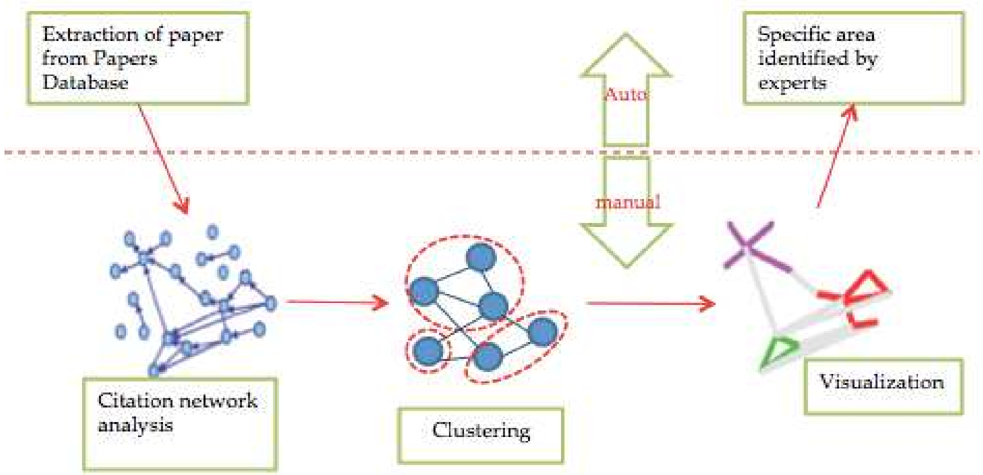

Extraction of paper from Papers Database
Auto
manual
Specific area identified by experts
Citation network analysis
Clustering
Visualization

Fig. 2. Procedure of Academic Landscape based on network analysis

1. To narrow down the number of papers by keyword (query) search in the publication databases.
2. To make a list of quotations (papers) for a target papers.
3. To analyze the variance of release years of the papers in the list above to weight each target papers.
4. To calculate the importance of each target papers by applying the page rank algorithm with the weight above.
5. To visualize the result based on the importance (node & edge)

3.1 Publication databases and search query

This study used SCOPUS for publication databases. As a result of narrowing down the number of papers by "clustering" as query, the number of papers for this study became 87,399.

3.2 Publication year analysis of variance and weight of each cited papers

Follow the steps 1) to 3) below to weight paper:

Step 1. To extract the maximum value in the histogram, the year with the maximum number of quotations is extracted by using the function below and save it in MaxYear.

$$MaxYear=max\{y(x)\,|\,y(x):=\text{Number of times the references in year } Y\} \quad (1)$$

Step 2. To identify the quotation period, find out the maximum year by examining the release years chronologically; the year with the number of quotations exceeding 10% of the maximum year's quotations for the first time should be the start year and it is saved in StartYear. Then, the year with the number of quotations getting below 10% should be the end year and it is saved in LastYear. The period of quotation is achieved by the formula below:

$$Period := (LastYear + 1) - StartYear \tag{2}$$

In addition, if there are two or more peaks in the histogram as shown in Fig. 3., the steps above should be repeated and the periods are saved in Period0,1,2,,,,n.

Step 3. To calculate the variance (standard deviation) of a histogram, it can be defined how long paper has been quoted by examining the variance (standard deviation) of the release years of the papers that referred to the target paper. The following shows a common way to obtain standard deviation; the obtained standard deviation is saved in Variance.

$$Variance = \frac{\sum (x - \overline{x})^2}{n} \tag{3}$$

Again, if the histogram is not normally-distributed (it has two or more peaks) as shown in Fig. 3., the variances (standard deviations) of different periods (Period0,1,,,n) is calculated and the average of them is to be saved in Variance. Then, the value in Variance is used for weighing of the target paper.

3.3 Calculation of the severity of each cited papers

The PageRank algorithm [7] is a technique to calculate quantitatively which page is most important if there are cross reference relations among pages such as the hyperlink structure. In this study, We calculate the importance of papers by utilizing this algorithm. The calculation follows the steps below:

(1)Each Paper have a unique scores. and each cited have a unique score,too.

(2) Assuming there is paper (X);

- The score of X is P,
- When X is quoted by other papers, X obtains scores, Variance_1,,,,Variance_n,
- When X quoted other papers, X provides scores to others,O_1,,,,O_m.

Where, the statement below is assumed to be true:

$$\boldsymbol{Variance_1 + \ldots + Variance_n = P}$$

$$O_1 = \cdots = O_m = \frac{p}{m}\left(= \frac{\sum_{i=1}^{m} \text{Variance}_i}{m}\right)$$

In other words, assuming that the total score of "out going" quotation for paper should be equal to the total score of "in coming", the total score should be considered to be the base of the paper. Then, paper should be considered to be more important as the score becomes higher. Thus, We intend to identify major papers in each field by applying the value in Variance to calculation of the "in coming" score for paper. In the conventional algorithm, if there are more than one "in coming" quotations, the score of each quotation was thought to be equal. On the contrary, in this research a quotation with a larger value in Variance should be thought to have a larger score. In this algorithm, the importance of paper can be calculated with the factor of quoted years reflected.

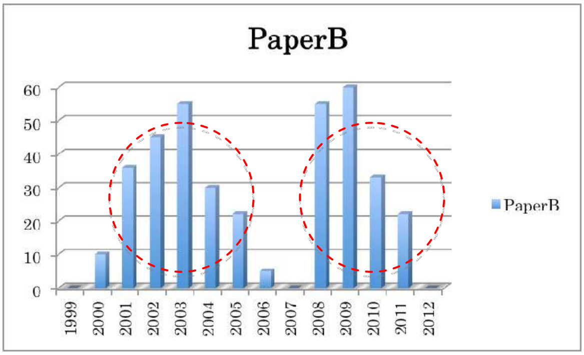


Fig. 3. Outside the normal distribution case that the shape of the histogram

3.4 Visualization based on the severity

Fig. 4. shows a visualization of the quotation network based on the importance obtained in the previous section. Each nodes has the title of paper displayed. As described in the previous section, a node with a larger importance is displayed as a larger node. In addition, this tool is called SciHi (**Sci**ence **Hi**ghangle).

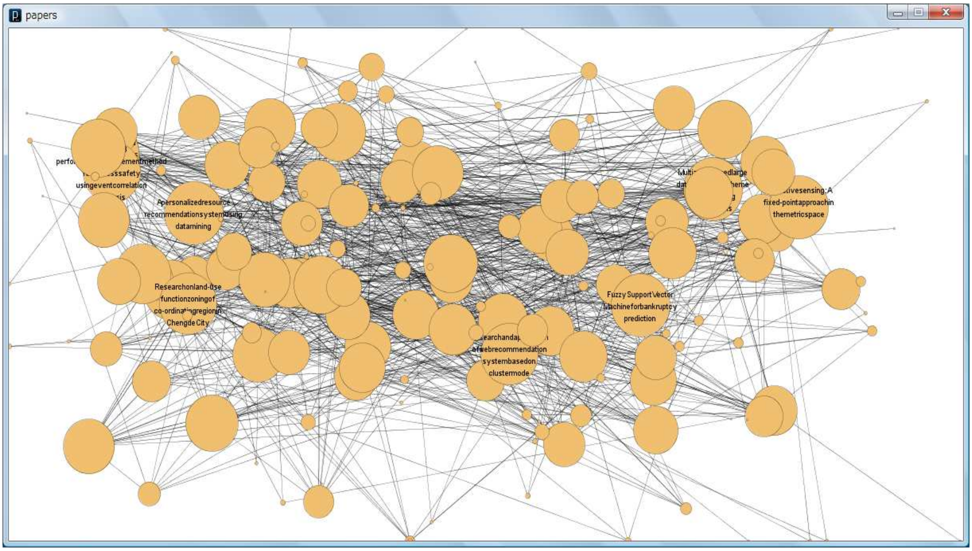


Fig. 4. Example of Academic Landscape based on weight of each cited papers

4. Evaluation experiment

In the survey report on study trend published in 2004, Tatsubori et al. used the publication databases of IBM to survey the research trend of software architecture after 1999. They manually extracted 51 major papers [8]. In this study, we examine how well the automated method can extract the major papers that were manually extracted by experts.

Tatsubori et al. used GoogleScholar to obtain the number of quotations for each year and extracted top 40 papers in a year. In addition to them, they extracted 11 more papers about software architecture by identifying international conferences that they thought particularly important, resulting in 51 major papers. Furthermore, they adopted a unique classification method in order to evaluate study trend quantitatively. Specifically, they classified 51 papers by defining each paper focused on which one of the following five roles that software architecture played in software development process. Fig.5. shows the classification result.

R: Requirements for the system of various stakeholders are reflected in the architecture.
M: Architect designs the architecture based on the meta-model.
A: The architecture is actually build up/modified.
C: If all the requirements cannot be met, the architecture is changed to meet the requirements.
S: If the architecture is abstract, it must be modified so that it can be implemented in a working system. Contrary, if the system has some changes, the changes should be reflected to the architecture.

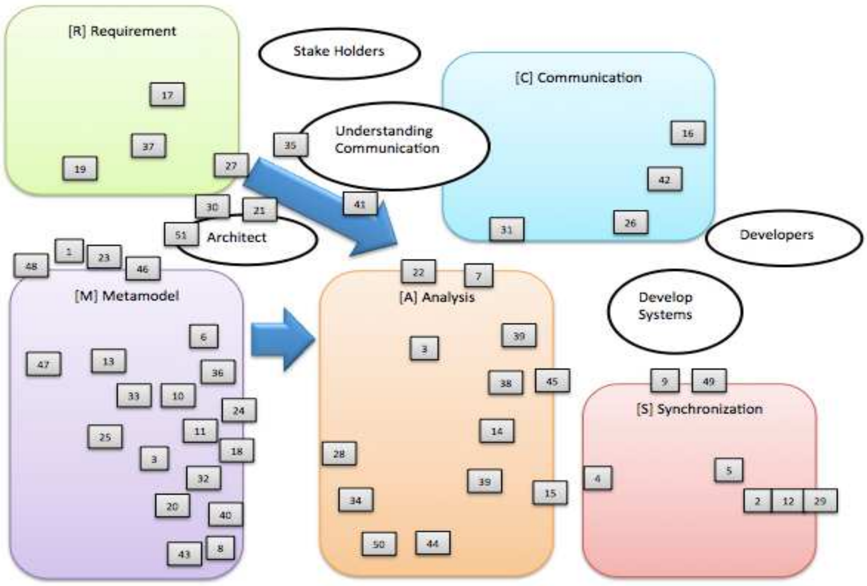


Fig. 5. An image of classification of papers by five roles of software architecture

Fig.6. shows the major papers extracted by SciHi in a similar manner to that of Tachibori et al. Fig.5. and Fig.6. both indicate the paper numbers.

The following describes the way of extraction by SciHi: We used SCOPUS as publication databases. Then, we selected “Software Architecture” for query and “1999 to 2004” for the period for extraction of papers. Among 51 papers, 28 papers were not included because they did not exist in SCOPUS (see the papers numbers in Fig.5. : 1, 5, 12~14, 16~19, 21, 24, 27, 29, 30, 32~39, 41~43 and 49~51). Regarding five roles in Fig.5., we manually made minute adjustment for SciHi (see Fig.6.) because they were an original classification set by Tatsubori et al. for quantitative evaluation of research trend.

As shown in Fig.6., visualization result by the page rank algorithm (hereafter, it is referred as this algorithm) with the variance value of a target paper taken into consideration as weight showed all the 23 papers were extracted as major papers. Especially, the papers (44~58 in Fig.6.) were extracted manually by Tatsubori et al. and all of them were extracted as major papers.

Thus, this algorithm can automatically extract major papers that are the same papers manually extracted by experts. However, SciHi extracted major papers that were not extracted by Tatsubori et al. such as a large (green) node in the cluster of [R]. As a future task we will examine these papers to see what meaning they have.

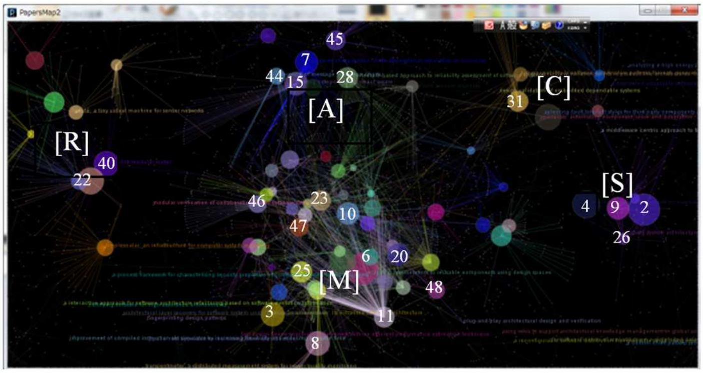


Fig. 6. An image of extraction of papers by SciHi

5. Conclusion

In this study on the network analysis of quoted academic papers, we tried automated extraction of major papers in each one of the fields identified by clustering. Specifically, we analyzed the variance of the release years of papers quoting the target paper; applied the result to the page rank algorithm to calculate the importance of the target paper. Then, we examined how well the automated method can extract the major papers that were extracted by experts by comparing the result with the report on research trend in the software architecture field that was published by Tatsubori et al. in 2004. As a result, this algorithm could extract all the target papers as major papers. In other words, this algorithm could automatically extract major papers with the same result as what experts manually obtained.

Last of all, in our experiment for evaluation, SciHi also extracted major papers that experts did not manually. As a future task we will examine what these papers mean.

6. References

[1] R•J•Wilson .(2001). *Introduction to Graph Theory*, Kindai Kagaku Sha Co.,Ltd.

[2] M. E. J. Newman and M. Girvan.(2004). *Finding and evaluating community structure in networks*, Physical Review E, Vol. 69.

[3] M. E. J. Newman. (2004).*Fast algorithm for detecting community structure in networks*, PHYSICAL REVIEW E 69, 066133, pp1-5, 2004．

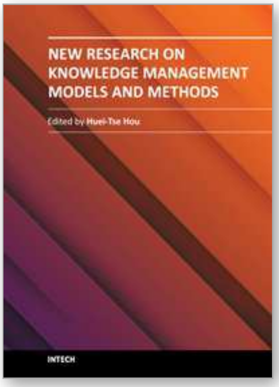
[4] M. E. J. Newman. (2005). *A measure of betweenness centrality based on random walks, Social Networks*, Vol. 27, No.1, pp. 39-54.

[5] H. Small. (1973). *Co-citation in the scientific literature: a new measure on the relationship between two documents*, Journal of the American Society for Information Science, Vol. 24, pp. 28-31.

[6] M. Kessler. (1963). *Bibliographic coupling between scientific papers*, American Documentation Volume 14, Issue 1, pages 10–25.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. (1998). *The PageRank Citation Ranking*. Bringing Order to the Web'.

[8] M.Tachibori,H.Maruyama,M.Kobayashi,D.Yellin,N.Yoshida,N.Kawai .(2005). *A Survey Report Digest on Research Trends in Software Architecture*，IPSJ SIG Technical Report, pp45-52.



New Research on Knowledge Management Models and Methods
Edited by Prof. Huei Tse Hou

ISBN 978-953-51-0190-1
Hard cover, 426 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821