Topological Analysis of Interdisciplinary Scientific Journals: Which Journals Will be the Next *Nature* or *Science*?

Yongjun Zhu
College of Computing & Informatics
Drexel University
3141 Chestnut St, Philadelphia, PA
19104, USA
zhu@drexel.edu

Erjia Yan
College of Computing & Informatics
Drexel University
3141 Chestnut St, Philadelphia, PA
19104, USA
ey86@drexel.edu

Il-Yeol Song
College of Computing & Informatics
Drexel University
3141 Chestnut St, Philadelphia, PA
19104, USA
song@drexel.edu

ABSTRACT

Identifying prestigious interdisciplinary journals is very significant for researchers. By publishing research works in prestigious journals, researchers can better propagate their works and get spotlights. Even though the quality of a paper is not represented by the journal that publish the paper, it is a general concern of researchers that how to identify a set of good journals to submit their papers. *Nature* and *Science* are the two journals that have been considering as the two top interdisciplinary journals worldwide. In this paper, we propose a method for identifying journals that have the potential to become the next *Nature* and *Science* through topological analysis of interdisciplinary scientific journals using citation data. By applying three different statistical methods (i.e., Multidimensional scaling, Principal component analysis, Cluster analysis), we identified a set of journals in which *PNAS* has the highest possibility to become the next *Nature* and *Science*. The study showed that citation data is a powerful data to measure similarity among journals.

CCS Concepts

• Mathematics of computing→Probability and statistics→ Statistical paradigms; • Information systems→Information systems applications→ Data mining;

Keywords

Bibliometrics; Data Mining; Multidimensional Scaling; Principle Component Analysis; Cluster Analysis

1. INTRODUCTION

Without doubt, the world's most prestigious interdisciplinary journals are *Nature* and *Science*. While these two journals are prestigious, people, especially scholars could be interested in other journals that resemble these journals or have highest possibilities to become journals that are as prestigious as these two journals. There are many reasons why discovering such journals are meaningful and important. For example, researchers can selectively choose these journals to submit papers. This could be a strategic way because these papers have high potential to become the next *Nature* and *Science* and this could be helpful in many cases such as tenure evaluation. Researchers can also identify a set of such journals and review them to find important research topics because prestigious journals are tend to publish high-impact papers.

Discovering such journals is not easy because there are many factors that contribute to the prestige of *Nature* and *Science*, and we are not able to capture them all. In addition, in terms of impact factor and history, there are other journals that surpass *Nature* and *Science*, but not as prestigious as *Nature* and *Science*. Based on the Journal Citation Reports 2013 of Web of Science, *Nature* has impact factor of 32 and *Science* has impact factor of 31, they rank 5th and 22nd respectively. Obviously, there are other journals that have higher impact factors than these two journals, but not as prestigious. This means we cannot just rely on readily available factors to do the analysis because this method may cause loss of information. A good alternative is to use citation data. By using citation data, we can discover the topology of scientific journals in terms of citation patterns. Through these patterns, we can identify journals that resemble *Nature* and *Science*.

In the area of bibliometrics, citation data have been used to identify author clusters from author co-citation networks [10], journal clusters [4] from journal citation networks, and perform subject area analysis [11] based on citation networks of major areas (e.g., Computer Science, Information Science, etc.). Recent years have witnessed a trend of using network-based bibliometrics indicators to differentiate the weight of citations (e.g., Eigenfactor, Y-Factor, and SCImago Journal Rank Indicator), with the assumption that a journal is prestigious if it is prestigious if it is cited by other prestigious journals [1, 2, 5, 8, and 9]. In these studies, citation data used as an efficient metric to measure similarities among journals, authors, and major areas.

To the best of our knowledge, there was no study that tried to identify journals that resemble *Nature* or *Science* using citation data. Thus, in this study, we investigate a set of journals that have potential to become the next *Nature* and *Science* by exploring a unique citation dataset. Specifically, we use journal-to-journal

RACS'15, October 9–12, 2015, Prague, Czech Republic.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
© 2015 ACM. ISBN 978-1-4503-3738-0/15/10 …$15.00.
DOI: http://dx.doi.org/10.1145/2811411.2811475

citation data to measure similarities between interdisciplinary journals and *Nature* and *Science* by applying three different statistical methods including multidimensional scaling, principal component analysis, and cluster analysis.

2. DATA

Granted citation dataset from Elsevier is used. This dataset contains journal-level citation instances among all indexed fields during the past 15 years (1997-2011) with a two-year citation window. The dataset contains 4,287,565 citation records, 128,625 journal and proceedings, and 324 subject area descriptions.

In terms of identifying journals that are similar to *Nature* or *Science*, we can use various criteria such as history, impact factor, etc. But the limitation of this approach is that we cannot enumerate out all the criteria that affect the prestige of journals when finding the next *Nature* or *Science*. This means this approach causes information loss. Instead, citation data are a complete dataset that contains all the citation patterns of indexed journals, which is a meaningful and valuable dataset for analyzing relationships (e.g., similarity) among journals.

Journal-to-journal citation data is used to monitor knowledge flow. A citation from journal A to journal B means knowledge flow from journal B to journal A because journal A imports knowledge from journal B by citing journal B (Figure 1). Thus, by monitoring citations, we can detect patterns of knowledge flow among journals. In turn, we can identify journals that have similar patterns with *Nature* or *Science* by investigating patterns of knowledge flow.

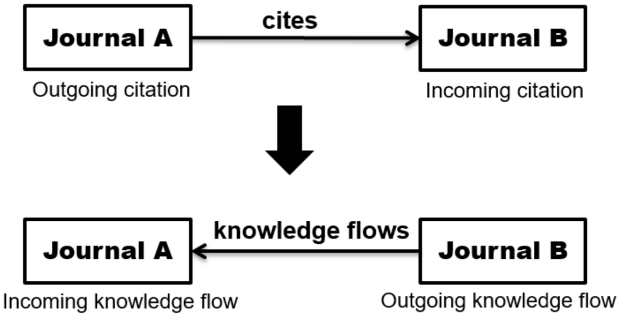


Figure 1. Citation and knowledge flow

As shown in Figure 1, two kinds of citations exist: incoming and outgoing citations. Thus, we can construct two kinds of citation matrices. Incoming citation is used to capture the status of cited journals, i.e., how cited journals are viewed by citing journals. Outgoing citation is used to capture knowledge composition of citing journals (Figure 2). Thus, we can capture two different patterns from these matrices.

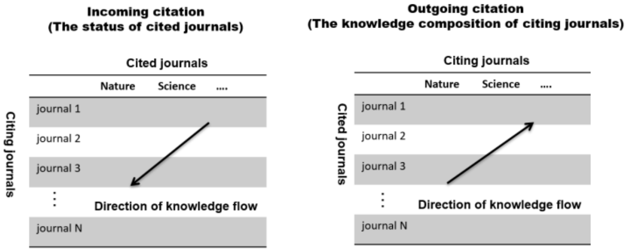


Figure 2. Incoming and outgoing citation matrices

The number of columns in two matrices is the same as the number of interdisciplinary journals. In the dataset used in this study, all journals have ASJS code, which is used for classifying journals based on their topics. In our study, this code should be "1000", because *Nature* and *Science* are interdisciplinary journals and "1000" represents interdisciplinary journals. Total 76 interdisciplinary journals indexed by Elsevier were included in the analysis. The number of rows of these matrices is the same as the number of journals indexed by Elsevier, i.e., 128,625 journals were used. The study also focuses on temporal evolution of the topology of scientific journals by dividing the dataset into five groups (i.e., 1997/1999, 2000/2002, 2003/2005, 2006/2008, and 2009/2011) and using 10 matrices (i.e. five incoming citation matrices and five outgoing citation matrices).

3. METHODS

Three types of statistical methods were used: Principal component analysis (PCA) [7], Multidimensional scaling [3], and Cluster analysis (Hierarchical clustering) [6].

Principal component analysis was used to discover journals that share the same component with *Nature* and *Science*. PCA is generally used to explain as much variance as possible by using limited number of variables. Variables included in the same component tend to explain the same aspect of the original data, and we assume journals that share the same component have similarities. We first calculated correlations among journals based on citation matrices, and then applied PCA to the obtained correlation matrices.

Multidimensional scaling was used to discover journals that are located the closest to *Nature* and *Science*. We used Euclidian distance as similarity measure and drew all the journals in a 2D map, and investigated the distance between these journals and *Nature* or *Science*. Journals that located near *Nature* or *Science* are the ones that are the most similar to *Nature* or *Science.*

Clustering analysis was used to discover journals that share the same cluster with *Nature* and *Science*. We used hierarchical clustering to manually investigate journals that are clustered with *Nature* and *Science*. While K-means clustering requires users to pre-define the number of clusters, hierarchical clustering shows the whole process of clustering, and is more appropriate in our case. Complete linkage clustering was used as the linkage criteria.

By synthesizing the results of these three analyses, we can get more meaningful results. We did not explicitly assign weights for each method, but synthetically considered all the results.

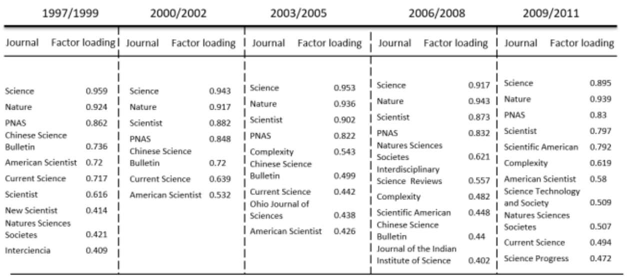
4. RESULTS

As incoming and outgoing citations signify different characteristics, we report the result of each analysis separately. Incoming citation analysis is used to identify journals that resemble *Nature* or *Science* from the perspective of citing journals; whereas outgoing citation analysis is for identifying journals that resemble *Nature* or *Science* in terms of knowledge composition because they tend to cite journals that also cited by *Nature* or *Science.*

4.1 Outgoing Citation Analysis

4.1.1 Principal Component Analysis

Figure 3 shows the result of PCA across five time periods.



1997/1999		2000/2002		2003/2005		2006/2008		2009/2011	
Journal	Factor loading	Journal	Factor loading	Journal	Factor loading	Journal	Factor loading	Journal	Factor loading
Science	0.959	Science	0.943	Science	0.953	Science	0.917	Science	0.895
Nature	0.924	Nature	0.917	Nature	0.936	Nature	0.943	Nature	0.939
PNAS	0.862	Scientist	0.882	Scientist	0.902	Scientist	0.873	PNAS	0.83
Chinese Science Bulletin	0.736	PNAS	0.848	PNAS	0.822	PNAS	0.832	Scientist	0.797
American Scientist	0.72	Chinese Science Bulletin	0.72	Complexity	0.543	Natures Sciences Societes	0.621	Scientific American	0.792
Current Science	0.717	Current Science	0.639	Chinese Science Bulletin	0.499	Interdisciplinary Science Reviews	0.557	Complexity	0.619
Scientist	0.616	American Scientist	0.532	Current Science	0.442	Complexity	0.482	American Scientist	0.58
New Scientist	0.414			Ohio Journal of Sciences	0.438	Scientific American	0.448	Science Technology and Society	0.509
Natures Sciences Societes	0.421			American Scientist	0.426	Chinese Science Bulletin	0.44	Natures Sciences Societes	0.507
Interciencia	0.409					Journal of the Indian Institute of Science	0.402	Current Science	0.494
								Science Progress	0.472

Figure 3. PCA result of outgoing citation analysis

In each of the five time periods, the component in which *Nature* and *Science* have the highest factor loading was selected and other journals with factor loadings greater than 0.4 in the same component were selected. The result showed that *PNAS* and *Scientist* are the two journals that appeared all five times, and journals such as *American Scientist*, *Current Science*, and *Chinese Science Bulletin* appeared four times. This means *PNAS* and *Scientist* are the two journals that most resemble *Nature* and *Science*.

4.1.2 Multidimensional Scaling

Multidimensional scaling analysis was performed for each time period, and journals that close to *Nature* and *Science* in terms of dimensional distance were selected. Figure 4 shows selected journals for the time period of 2009/2011. They were also connected with *Nature* or *Science* with red lines to show the approximate distance.

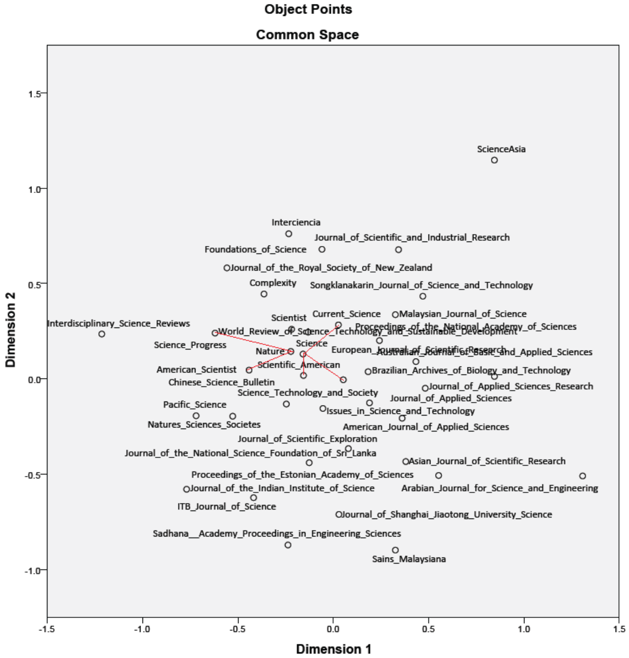


Figure 4. MDS result of outgoing citation for 2009/2011

Based on five MDS maps, we can easily identify journals that are closely located with *Nature* or *Science*.

Table 1. MDS result of outgoing citation

1997/1999	2000/2002	2003/2005	2006/2008	2009/2011
PNAS, Scientist, Scientific American, Natures Sciences Sociétés American Scientist, Chinese Science Bulletin, Current Science	*PNAS, Journal of the Royal Society of New Zealand, Natures Sciences Sociétés, Discovery and Innovation, Scientist, American Scientist, Chinese Science Bulletin, Current Science*	*PNAS, Journal of the Royal Society of New Zealand, Journal of Scientific Exploration, Current Science, Chinese Science Bulletin, Scientist*	*PNAS, Scientist, Scientific American, Natures Sciences Sociétés, Current Science*	*PNAS, Scientific American, Science Technology and Society, American Scientist*

We can divide them into two groups by reviewing five MDS maps (Table 1). Group A consists of two journals: *PNAS* and *Scientist* that are highly close to *Nature* or *Science* with the smallest distance. Group B consists of four journals: *American Scientist*, *Current Science*, *Scientific American*, and *Chinese Science Bulletin* that are close to *Nature* or *Science*.

4.1.3 Cluster Analysis

Hierarchical clustering was performed for each time period. Journals that belong to the same cluster with *Nature* and *Science* were identified. For hierarchical clustering, we need to manually cut the dendrogram to identify cluster. Thus, we selected journals that joined the cluster of *Nature* and *Science* at the early stage by reviewing the structure of dendrograms. Figure 5 shows the result of hierarchical clustering for the last time period 2009/2011. As we can see from Figure 5, journals such as *PNAS*, *Scientist*, and *Scientific American* were clustered in the same group with *Nature* and *Science*. When cutting the dendrogram, we tried to limit the number of journals (i.e., less than five) that are included in the same cluster with *Nature* and *Science* in order to select only highly similar journals. We can cut the dendrogram in many ways, but by applying our heuristic (i.e., less than five journals excluding *Nature* and *Science* in the cluster), we can cut the dendrogram as shown in Figure 5.

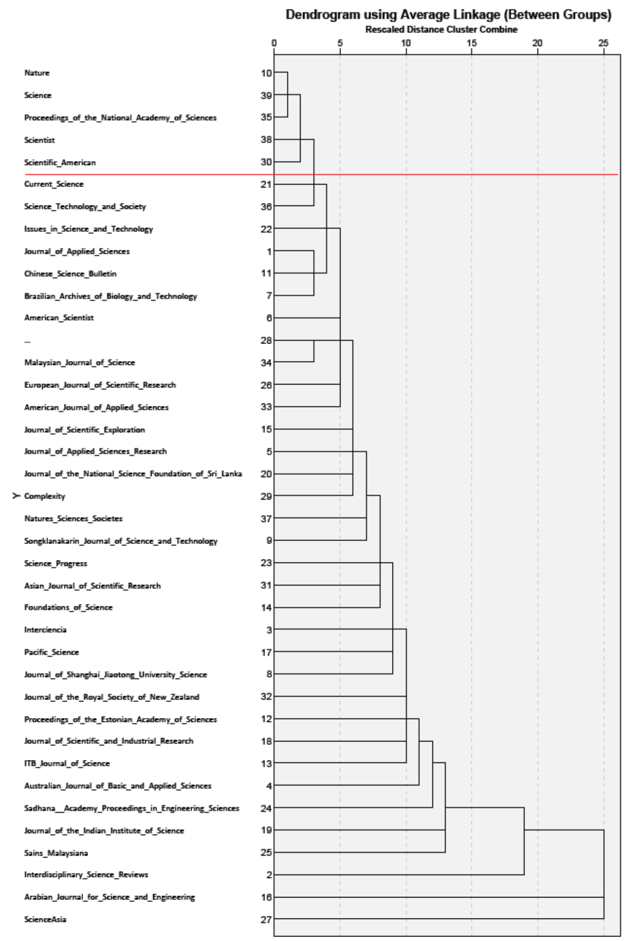


Figure 5. HCA result of outgoing citation for 2009/2011

Based on the result of hierarchical clustering, we can also divide them into two groups (Table 2).

Table 2. HCA result of outgoing citation

1997/1999	2000/2002	2003/2005	2006/2008	2009/2011
PNAS, American Scientist, Chinese Science Bulletin, Scientist, Natures Sciences Sociétés	*Scientist, PNAS, Chinese Science Bulletin, Current Science*	*Scientist, PNAS, Current Science, Chinese Science Bulletin*	*Scientist, PNAS, Natures Sciences Sociétés*	*PNAS, Scientist, Scientific American*

Group A consists of two journals: *PNAS* and *Scientist* that are clustered with *Nature* and *Science* at the very early stage, and Group B consists of four journals: *American Scientist*, *Current Science*, *Chinese Science Bulletin*, and *Natures Sciences Sociétés* that are clustered with *Nature* and *Science* at early stage.

In summary, based on the results of three different analyses, five journals that are closest to *Nature* and *Science* were detected, and they were divided into two groups based on similarity (i.e., group A has a greater similarity). *PNAS* and *Scientist* are in group A whereas *American Scientist*, *Current Science*, and *Chinese Science Bulletin* are in groups B (Figure 6).

Figure 6. Journals that are similar to *Nature* and *Science* (Outgoing Citation)

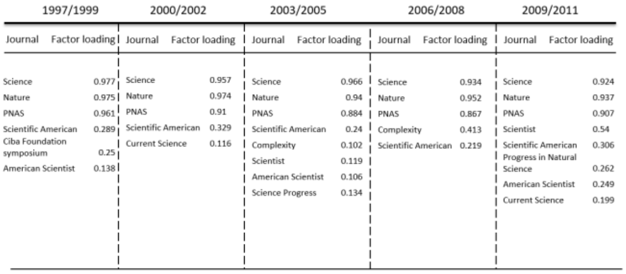
As someone may notice, *PNAS* is the official journal of the United States National Academy of Sciences. Interestingly, the other two journals appeared in Figure 6: *Current Science* and *Chinese Science Bulletin* are published by Indian Academy of Sciences and Chinese Academy of Sciences, respectively. This result shows that outgoing citations can meaningfully capture similarity among journals in terms of knowledge composition.

4.2 Incoming Citation Analysis

For incoming citations, we also used three statistical methods to identify journals that resemble *Nature* and *Science*. We assume that *Nature* and *Science* receives many citations because they are prestigious, and the volume of citations indirectly represents the status of these journals. Thus, by reviewing patterns of incoming citations, we can identify journals that have similar status to *Nature* and *Science*.

4.2.1 Principal Component Analysis

Figure 7 shows the result of PCA across five time periods based on incoming citation data (i.e., outgoing knowledge flow). Journals that are included in the same component with *Nature* and *Science* were identified.



1997/1999		2000/2002		2003/2005		2006/2008		2009/2011	
Journal	Factor loading	Journal	Factor loading	Journal	Factor loading	Journal	Factor loading	Journal	Factor loading
Science	0.977	Science	0.957	Science	0.966	Science	0.934	Science	0.924
Nature	0.975	Nature	0.974	Nature	0.94	Nature	0.952	Nature	0.937
PNAS	0.961	PNAS	0.91	PNAS	0.884	PNAS	0.867	PNAS	0.907
Scientific American	0.289	Scientific American	0.329	Scientific American	0.24	Complexity	0.413	Scientist	0.54
Ciba Foundation symposium	0.25	Current Science	0.116	Complexity	0.102	Scientific American	0.219	Scientific American	0.306
American Scientist	0.138			Scientist	0.119			Progress in Natural Science	0.262
				American Scientist	0.106			American Scientist	0.249
				Science Progress	0.134			Current Science	0.199

Figure 7. PCA result of incoming citation analysis

As shown in Figure 7, *PNAS* and *Scientific American* appeared five times and *American Scientist* appeared three times. The result is quite different from that of outgoing citation.

4.2.2 Multidimensional Scaling

The MDS result of incoming citation for 2009/2011 is shown in Figure 8. Journals that are closely related to *Nature* and *Science* were identified.

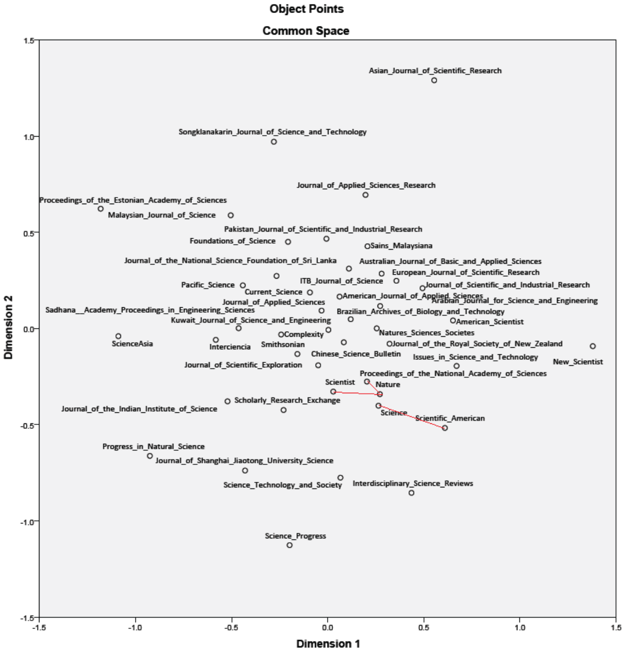


Figure 8. MDS result of incoming citation for 2009/2011

In Figure 8, *PNAS*, *Scientist*, and *Scientific American* are closely collocated with *Nature* and *Science*.

By combining five MDS maps (Table 3), we also identified two groups based on distance. Group A includes *PNAS* while Group B includes *Scientist* and *Scientific American*. The result of incoming citations is largely different from that of outgoing citations. This means while there are journals similar to *Nature* and *Science* in terms of knowledge composition, not many journals are actually considered as counterparts of *Nature* and *Science* by other journals.

Table 3. MDS result of incoming citation

1997/1999	2000/2002	2003/2005	2006/2008	2009/2011
PNAS	*PNAS*	*N/A*	*PNAS*	*PNAS, Scientist, Scientific American*

4.2.3 *Cluster Analysis*

Result of the hierarchical clustering analysis for incoming citations for 2009/2011 is shown in Figure 9. We also applied the same heuristics as outgoing citations and manually cut the dendrogram.

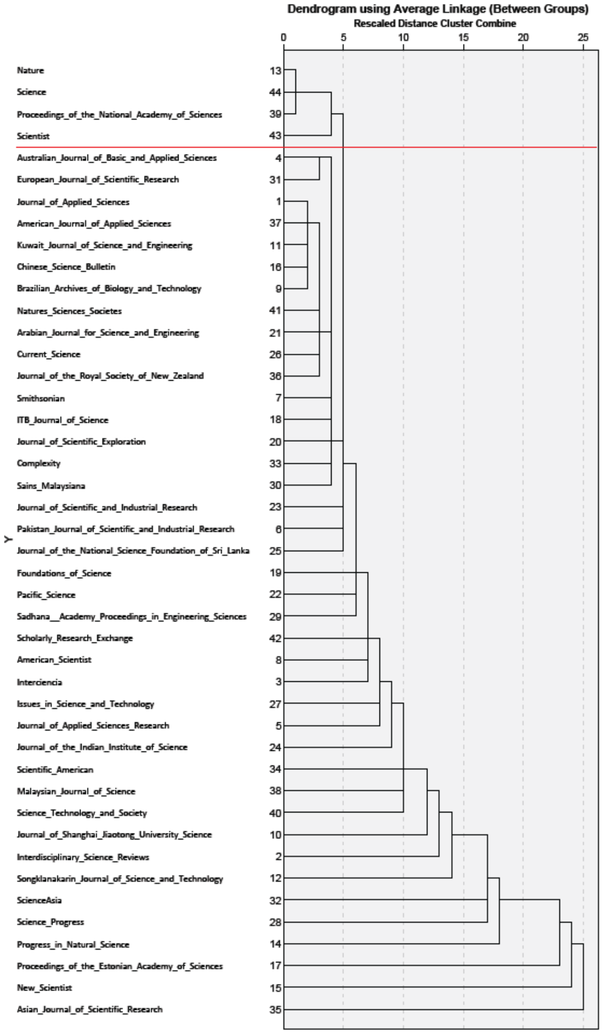


Figure 9. HCA result of incoming citation for 2009/2011

In Figure 9, *PNAS* and *Scientist* belong to the same cluster with *Nature* and *Science*.

Table 4 shows the HCA result of incoming citation, which is very similar to the MDS result.

Table 4. HCA result of incoming citation

1997/1999	2000/2002	2003/2005	2006/2008	2009/2011
PNAS	*PNAS*	*PNAS*	*PNAS*	*PNAS, Scientist*

Based on the results of three different analyses, three journals that are the closest to *Nature* and *Science* were detected, and they were also divided into two groups based on similarity. Group A includes *PNAS* and Group B includes *Scientist* and *Scientific American* (Figure 10).

PNAS
Policing synchronous reproduction
TheScientist
UNWINDING RNA
SCIENTIFIC AMERICAN
DID HUMANS STOP AN ICE AGE?
8,000 YEARS OF GLOBAL WARMING

Figure 10. Journals that are similar to *Nature* and *Science* (Incoming Citation)

By combing both incoming and outgoing citations, *PNAS* and *Scientist* are the two journals that are the closest to *Nature* and *Science*. They appeared in the results of both incoming and outgoing citation analyses. Because *PNAS* was included in group A in both analyses, it is the single closest journal to *Nature* and *Science* followed by *Scientist*. Even though there are journals that have higher impact factor and longer history than *PNAS*, *PNAS* was chosen as the next *Nature* and *Science* in this study. This means that the prestige of a journal is not decided by some apparent factors. Instead, by reviewing citation patterns, we can get more meaningful and well-rounded results.

5. CONCLUSIONS

In this study, we aimed to identify journals that resemble *Nature* and *Science* under the research question of "Which journals will be the next *Nature* or *Science*" by exploring journal-to-journal citation data from 1997 to 2011. Three different analyses (i.e., principal component analysis, multidimensional scaling, and cluster analysis) were used to answer the research question. Both incoming citation and outgoing citation data were analyzed separately in order to draw a more meaningful conclusion.

The results showed that *Proceedings of the National Academy of Sciences* (*PNAS*) is the closest journal to *Nature* and *Science*, and we can conclude that among the 76 journals included in the analysis, *PNAS* has the highest possibility to become the next *Nature* and *Science*. Other journals such as *Scientist*, *American Scientist*, *Current Science*, *Chinese Science Bulletin*, and *Scientific American* are similar to *Nature* and *Science* in some degree, but we still need more concrete evidence. The results also showed that topological analysis of journals is a reasonable and applicable way to capture similarity of journals. Especially, one can focus on the temporal evolution to find out the changing landscape of topology.

Results also showed that citation data is a powerful asset to measure similarity among journals. Patterns captured in incoming and outgoing citation data can be directly used as similarity measure and complement the limitations of multiple criterion-based method.

While the methods and results of this study help draw solid conclusions, there are also some limitations. The study focused on 76 interdisciplinary scientific journals indexed by Elsevier, and omitted journals that are not indexed by Elsevier or not assigned to the same subject area as those 76 journals. Furthermore, the results are purely based on citation patterns, other editorial, managerial, and latent socio-technical factors were not captured. As a feature work, we plan to expand our dataset and construct an integrated metric that can rank the possibility to become the next *Nature* and *Science* in a more automated way.

6. REFERENCES

[1] Bergstrom, C. T., & West, J. D. (2008). Assessing citations with the Eigenfactor™ Metrics. *Neurology, 71*(23), 1850-1851.

[2] Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics, 69*(3), 669-687.

[3] Borg, I., & Groenen, P. J. F. (1997). Modern multidimensional scaling: Theory and applications. New York: Springer.

[4] Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, 24(6), 425-436.

[5] Dellavalle, R. P., Schilling, L. M., Rodriguez, M. A., Van de Sompel, H., & Bollen, J. (2007). Refining dermatology journal impact factors using PageRank. *Journal of the American Academy of Dermatology, 57*(1), 116-119.

[6] Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

[7] Jolliffe, I. T., & NetLibrary, I. (2002). Principal component analysis (Secondition.; 2nd ed.). New York: Springer.

[8] Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology, 60*(7), 1327-1336.

[9] SCImago (2007). SJR: SCImago Journal & Country Rank. Retrieved August 31, 2009 from http://www.scimagojr.com

[10] White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American society for information science*, 49(4), 327-355.

[11] Zhu, Y., & Yan, E. (2015). Dynamic subfield analysis of disciplines: An examination of the trading impact and knowledge diffusion patterns of computer science. *Scientometrics*, 104(1), 335-359.