Visualizing the Scientific World and Its Evolution

I. Samoylenko, T.-C. Chao, W.-C. Liu, and C.-M. Chen
Department of Physics, National Taiwan Normal University, Taipei, Taiwan. E-mail: cchen@phy.ntnu.edu.tw

We propose an approach to visualizing the scientific world and its evolution by constructing minimum spanning trees (MSTs) and a two-dimensional map of scientific journals using the database of the Science Citation Index (SCI) during 1994–2001. The structures of constructed MSTs are consistent with the sorting of SCI categories. The map of science is constructed based on our MST results. Such a map shows the relation among various knowledge clusters and their citation properties. The temporal evolution of the scientific world can also be delineated in the map. In particular, this map clearly shows a linear structure of the scientific world, which contains three major domains including physical sciences, life sciences, and medical sciences. The interaction of various knowledge fields can be clearly seen from this scientific world map. This approach can be applied to various levels of knowledge domains.

Introduction

The scientific knowledge of human beings is a complex and dynamic network. Understanding various knowledge domains is crucial in documenting the history of our knowledge development, and could even reliably predict the future trend of our knowledge evolution. Moreover, such understanding provides objective guidance for allocating resources and efficiently promotes interdisciplinary collaborations. To capture the structure and evolution of mankind's scientific endeavor, two kinds of strategies, including descriptive models (Börner, Maru, & Goldstone, 2004; Boyack, 2004; Chen, 2004; Newman, 2004; Small, 1999) and process models, are commonly adopted. Descriptive models aim to describe the major features of and provide an outlook on a knowledge domain. For example, research in knowledge domain visualization has studied the statistical patterns of citation networks, coauthorship networks, and the identification of research fronts. On the other hand, process models aim to extract the mechanisms and temporal dynamics of a real-world network. The emergence of small-world and scale-free network structures shows two well-known examples of process models in statistical physics.

Previous studies in knowledge domain visualization have unveiled many features or underlying mechanisms of specific knowledge domains. In the attempt to map the structure of science, citation analysis has been shown to play a prominent and productive role. Price has used citation patterns of documents to depict the topography of current scientific literature (Price, 1965). For the analysis of citation patterns, most noteworthy from this field are the methods of cocitation (Small, 1973) and bibliographic coupling (Kessler, 1963). In the cocitation scheme, similarity between two documents i and j is based on the number of documents that cite both i and j. In bibliographic coupling, however, similarity is based on the number of documents cited by both i and j. Because the references of a document do not change after its publication, the disadvantage of bibliographic coupling in structural studies of science is that the structure is in general dynamic over time, whereas bibliographic coupling is a fixed measure. In contrast, cocitation reflects the frequency of being cited, which is a characteristic that is variable over time. Small and Griffith have pioneered the method of mapping the structure of scientific literatures by cocitation analysis of the scientific network (Griffith, Small, Stonehill, & Dey, 1974; Small & Griffith, 1974). On the other hand, Narin works at a more general level by using the citation patterns between journals to define the disciplinary structure of science (Narin, Carpenter, & Berlt, 1972). Journal–journal citations have also been used in scientometric mapping by many authors (Carpenter & Narin, 1973; Doreian & Fararo, 1985; Leydesdorff, 1986; Tijssen, de Leeuw, & van Raan, 1987). More recently, Leydesdorff and Cozzens have also used these journal mappings to indicate change in science (Leydesdorff & Cozzens, 1993). The generalized bibliographic coupling at the journal level (the sum of multiplied citing frequency of all journals by journals i and j) does not have the abovementioned shortcoming of giving a fixed network structure of science over time. In fact, this approach can give a better overall view of human scientific activities by including almost all published documents, and even allow the feasibility of predicting the future trend of the scientific network. A comprehensive review of knowledge domain visualization

Received February 21, 2005; revised August 9, 2005; accepted September 6, 2005

 ● Published online 28 June 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20450

can be found in a recent article by Börner and others (Börner, Chen, & Boyack, 2003).

Recently, because of the development in information technology and the setup of many high-volume and high-quality data sets of scientific publications, the study of scientific network has attracted considerable attention. For example, evolution of coauthorship networks and patterns of scientific collaboration have been studied by using data from various bibliographic databases (Börner, et al., 2004; Newman, 2004). Mapping of the highest-performing papers over the 20-year period within the *Proceedings of the National Academy of Sciences (PNAS)* domain was generated by using citation analysis to study changes and trends in the subjects of highest impact (Boyack, 2004). Progressive visualization of the evolution of a knowledge domain was applied to a cocitation study of the superstring field in theoretical physics (Chen, 2004). Nevertheless, a panorama of mankind's scientific activities is still desired. In particular, a map of science can provide insight into a contemporaneous state of knowledge and help researchers to make new discoveries. A recent study on visualizing science by citation mapping attempts to construct maps using simplified methods for ordination, for a dataset of about 36,000 documents (Small, 1999). Such a map of science could represent the relation among different disciplines, fields, specialties, and individual papers, by their physical proximity as calculated from citation data. To represent the high-dimensional citation data on the two-dimensional surface of either paper or computer screen, several dimensional reduction techniques have been shown to be quite useful. These methods include multidimensional scaling (Kruskal, 1964), eigenvalue decomposition (Davidson, Hendrickson, Johnson, Meyers, & Wylie, 1998), factor analysis (Thurstone, 1931), latent semantic analysis (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), pathfinder network scaling (Schvaneveldt, 1990), and self-organizing maps (Kohonen et al., 2000). Other ordination techniques, such as triangulation (Lee, Slagle, & Blum, 1977) and force-directed placement (Fruchterman & Reingold, 1991), have also been applied to display a large set of documents.

Here we describe a simple approach to visualizing the scientific world and its temporal trend using the journal citation data from the Science Citation Index (SCI) during the years from 1994 to 2001. This citation dataset is directly computed from the extraction of the CD version of the SCI database. To find the most relevant features and for simplicity, only research journals of impact factor greater than 1 (about 2,000 journals) are considered. The dataset reported in this article consists of more than 3 million articles and the number of cited papers exceeds 67 million. To begin with, we convert the similarity in journals' citation patterns into distance between journals, such that closely related journals are short distances apart. This distance matrix of journals is then used to cluster journals by a minimum spanning tree (MST) algorithm (Gordon, 1981; Kruskal, 1956). MST is a spanning tree for which the sum of similarities in their citation patterns is the largest. Each cluster of journals is named by the most popular words that appear in the title of its members. The exact mapping of these clusters in a high dimensional space is implemented by classical scaling (CS) according to the accumulated distance matrix of clusters. The eigenvalues calculated from classical scaling (Borg & Groenen, 1997), suggest that a two-dimensional map of these clusters can be obtained from the projection of the high dimensional map onto the two principal axes, with limited distortion. The coordinates of clusters in the two-dimensional map are then optimized by Sammon mapping (SM; Borg & Groenen, 1997). Such an approach produces a sensible map of the scientific world with reasonable coordinates of various knowledge clusters, but the distortion in the distances between clusters is usually comparable to their yearly temporal shift. A final step is to adjust the coordinates locally by Procrustes analysis (Gower & Dijksterhuis, 2004), such that the temporal trend of the map can be studied in detail.

Constructing a Network of Scientific Journals

The scientific knowledge of human beings consists of various knowledge domains, such as physics, chemistry, and neurology. In each knowledge domain, journals of different aims and scope publish numerous papers every year to report the most recent discoveries in a specific research area. These journals are connected to each other through the references in published papers and form a scientific network. The complexity of the scientific network is an interesting research topic and has profound importance. To investigate this problem, we first constructed a network of scientific journals by MST using the SCI database. In bibliometric studies, a common technique for clustering documents is the single-linkage method, which links two clusters together by the "*nearest neighbors*" across clusters. A convenient approach in constructing a single-link partition is the MST, which is usually defined in terms of link lengths (or dissimilarity). The dissimilarity between two documents can be visualized by the differences in their citation patterns. Because the objects in our case are research journals instead of research articles, we find it more adequate to use the generalized bibliographic coupling at the journal level to measure the similarity between two journals. In other words, the citation pattern of each journal is represented by a normalized citation vector and these vectors form a rescaled citation matrix. The similarity between two journals is related to the scalar product of their citation vectors.

To begin with, the citation matrix $\{N_{ij}\}$, number of citations of journal j cited by journal i, is extracted from a dataset Ω. The similarity of two journals i and j in their citation patterns is defined as its cosine measure

$$S_{ij} = \frac{\sum_{k\in\Omega} c_{ik}c_{jk}}{\sqrt{\sum_{k\in\Omega} c_{ik}^2 \cdot \sum_{k\in\Omega} c_{jk}^2}}, \quad (1)$$

DOI: 10.1002/asi

where $c_{ik} \equiv N_{ik}/(\sum_{j\in\Omega} N_{ij})$ is the normalized citation matrix element. Depending on the similarity in citation patterns of journals i and j, the value of s_{ij} ranges from 0 to 1. For mapping or visualization, coefficients of similarity are converted into distances such that closely related journals are short distances apart and remotely related journals are long distances apart. We express this conversion as

$$d_{ij} = \sqrt{\frac{1}{\max(t, s_{ij})} - 1}, \tag{2}$$

where t is a cutoff parameter for this distance conversion and the function max(a,b) chooses a bigger value from a and b. For the case of $t = 0$, the distance between two unrelated journals is infinite. This choice is not practical for science mapping because the projection of journals on the map will be largely distorted. For simplicity, we take $t = 0.01$ in our analysis. A number of algorithms have been developed to construct MST. Here we use the Kruskal algorithm (Kruskal, 1956) to construct MST of the database by successively connecting nearest-neighboring objects from an initially disconnected graph. Decision about whether to connect a pair of objects of the seeding graph is made with the use of the distance array d_{ij}, which is re-sorted in the order of increasing distance. This resorted distance array is scanned from its top and a linkage between two neighboring objects is added to the seeding graph only if no loop appears in that graph. At the end of this procedure, a complete MST of the database Ω has been constructed.

The constructed MST of Ω can be decomposed to cluster journals by breaking adequate links between journals. For a complete MST, we record the distances of all existing links between neighboring journals in a linkage array. Starting from the longest link, a link between journals would be broken if the following two conditions were met: 1. the link belongs to a cluster of size (number of journals) greater than V_{max} and 2. both clusters resulting from breaking that link are of size greater than V_{min}. Here the parameter V_{max} is used to limit excessive chaining and the parameter V_{min} is to prevent too many isolated journals. In general, values of these two parameters will depend on the journal database. The choices of values for V_{min} and V_{max} are not unique for a journal database. Different choices imply clustering journals at different resolutions. However, a slight change in these values will not drastically change the clustering of journals. For each cluster, we define its dominant journal (DJ) as the most cited journal in the cluster. In other words, a DJ has the highest sum of normalized citations in the cluster.

Results

As a demonstration, 196 SCI journals of impact factor (IF) greater than 5 were clustered and their MST was constructed as shown in Figure 1, using $V_{min} = 5$ and $V_{max} = 20$ (Supporting information, 2005). Here solid lines represent links between journals within a cluster and dotted lines represent shortest links between two neighboring clusters. DJ of a cluster is represented by a filled circle and each journal is labelled by a number. This dataset of 196 journals is decomposed into 21 clusters of size ranging from 5 to 19 journals. The largest cluster is related to neurosciences and contains 19 journals. Its DJ is *J Neurosci*. All these 19 journals belong to neurosciences or clinical neurology in SCI subject categories. The major part of this cluster contains 12 journals (including the DJ) and the scope of these journals covers a broad range of topics in neurosciences. The remaining part of this cluster (journals 13–19) deals with clinical neurology, which also contains a small branch focusing on blood circulation in the brain (18 and 19). The neighboring clusters of this neuroscientific cluster demonstrate its importance in connecting basic sciences and applications. Two application-related clusters, including psychology (154–159) and psychiatry (172–176), are located at the peripheral region of the MST. The connection between psychological journals and neuroscientific journals is through journals related to cognitive neurosciences and behavior sciences (154 and 155). The connection between psychiatric journals and neuroscientific journals is through molecular psychiatry (172). These brain-science related clusters are connected to biochemistry and molecular biology through physiological journals (59–61). The second largest cluster in this MST contains 18 journals and is related to the SCI subject categories including biochemistry and molecular biology, cell biology, genetics and heredity, and developmental biology. This cluster can be divided into a branch of broader range in biology (20–27) and a branch of more specific topics in genetics (28, 29, and 33–37) and developmental biology (28–32). The DJ of this cluster is *Cell* (25). Research journals in molecular and cell biology, such as *Cell* (25) and *Molecular Cell* (24), provide useful guidance in the research areas of genetics and developmental biology. We note that, by using only journals of IF $>$ 5, many important knowledge fields, such as mathematics, computer science, and materials science, are not included in the MST in Figure 1. Therefore, Figure 1 does not give a complete view of the scientific network. Instead, we use it to demonstrate our methodology of clustering scientific journals. It is also easier to view the detail structure of a small database.

From the MST in Figure 1, it is evident that the number of biological or medical journals is much larger than that of physical and chemical journals for the journal set of IF $>$ 5. These biomedical journals form a big condensed domain on the right hand side of MST, whereas the physical–chemical journals form a linear domain on the left hand side. Although many biomedical instruments are invented based on physical sciences, it seems that physical journals are remotely connected to biomedical journals. The physical cluster contains 14 journals and most of them are in the SCI category of physics. The only exceptions are *Surface Science Report* (84, chemistry) and *Progress in Quantum Electronics* (85, engineering). The DJ of the physical cluster is *Phys Rev Lett* (PRL, 78), which locates at the center of this cluster. Particle and nuclear journals (72–74) are clearly separated from condensed matter journals (79 and 81). Although *Physics Today*

DOI: 10.1002/asi

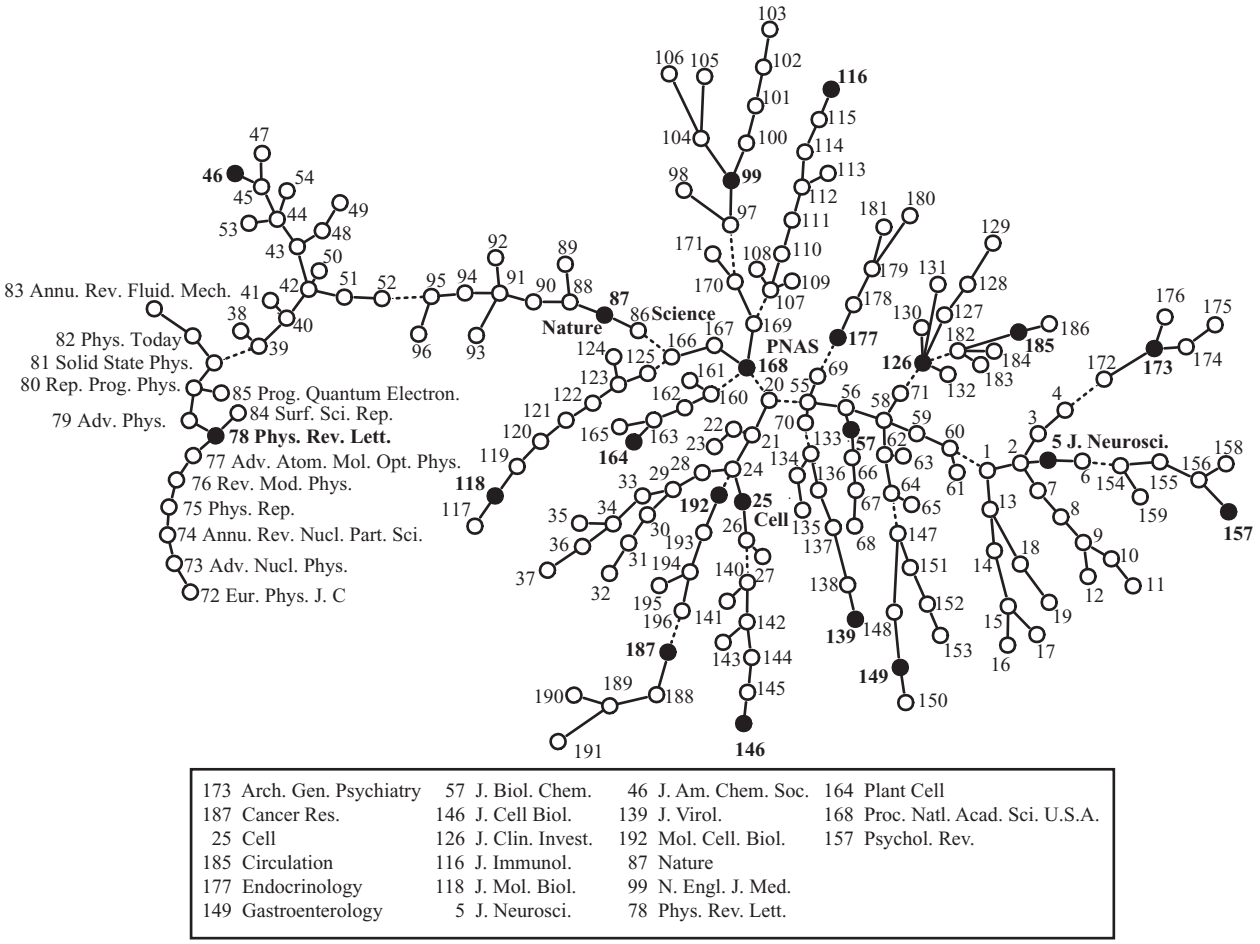
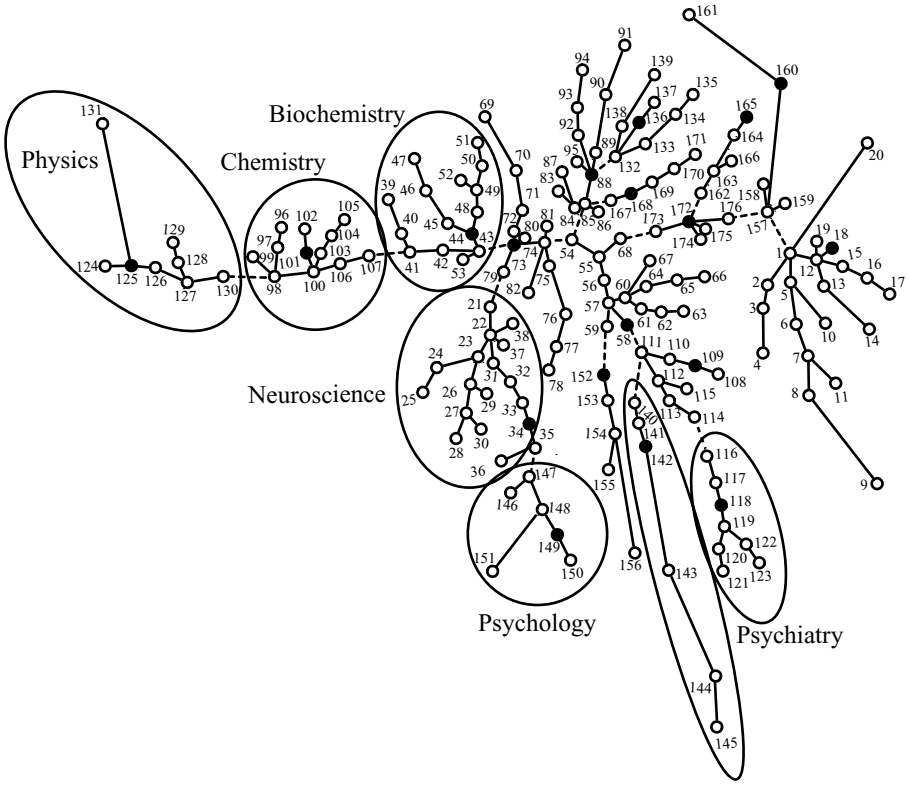


FIG. 1. An MST of SCI journals of impact factor greater than 5. Each journal is represented by a circle and labeled by a number. Titles of all journals can be found in the supplemental data. Journals in the same cluster are connected by a solid line and clusters are connected by a dash line. The line length is related to the distance between two connected journals, but distortion could occur due to a two-dimensional map of the MST. Filled circles are DJs of these clusters.

(82), *Report on Progress in Physics* (80), *Review of Modern Physics* (76), and *Physics Report* (75) are in the same SCI category of physics & plus; multidisciplinary, these journals focus on very distinct subjects. Journals 75 and 76 emphasize particle and nuclear physics, whereas 80 and 82 are more connected to condensed matter physics. The physical sciences are connected to biomedical sciences through multidisciplinary journals including *Science* (86), Nature (87), and *Proc Natl Acad Sci USA* (*PNAS*, 168). Figure 1 suggests that the cluster of *PNAS* serves as a hub of biomedical sciences, whereas the cluster of *Nature* and *Science* serves as a bridge connecting the two domains.

The journal set for Figure 1 actually consists of many multidisciplinary journals and many review journals. It is also interesting to view the MST structure of the core research journals. Therefore, we construct an MST for 176 core research journals of IF $>$ 4, as shown in Figure 2. The MST in Figure 2 contains 18 clusters and their cluster members can be found in the supporting information (Supporting information, 2005). The clustering of these core research journals seems to be reasonable, except for some small clusters of size V_{min}. For example, the plant and earth science cluster (DJ is *Plant Physiol*) also contains incompatible research journals in climate research (144, 145) and global biogeochemical cycles (143) because the number of journals in these research areas is less than V_{min}, which can be easily found by inspecting the long distances between cluster members. This inconsistency can be alleviated by using a larger data set or by using a variable V_{min} if the distance between cluster members is longer than a threshold. The general features of Figure 2 are similar to that of Figure 1, i.e., physics and chemistry are separated from the rest of biomedical clusters. In the absence of a multidisciplinary bridge, physical sciences are connected to biomedical sciences through the biochemistry cluster. It is found that cell biology also plays an important role in connecting these biomedical clusters. However, the connection of these clusters in the MST of Figure 2 is not robust. Deletion of the connecting journal between two clusters could break their connections. For example, if we remove the journal *Mol Psychiatry* (116) from the data set, the psychiatry cluster will be connected to the neuroscience cluster. To build a robust connection between clusters, we construct the MST of these clusters by calculating the accumulated distance

DOI: 10.1002/asi



118	Arch. Gen. Psychiatry	109	Development	34	J. Neurosci.	125	Phys. Rev. Lett.
44	Biochemistry	152	Genetics	101	J. Am. Chem. Soc.	149	Psychol. Rev.
172	Cancer Res.	73	J. Biol. Chem.	165	J. Virol.	160	Radiology
58	Cell	88	J. Clin. Invest.	18	N. Engl. J. Med.		
136	Circulation	168	J. Immunol.	142	Plant Physiol.		

FIG. 2. An MST of SCI core research journals of impact factor greater than 4. Each journal is represented by a circle and labeled by a number. Titles of all journals can be found in the supplemental data. Journals in the same cluster are connected by a solid line and clusters are connected by a dash line. The line length is related to the distance between two connected journals, but distortion could occur due to a two-dimensional map of the MST. Filled circles are DJs of these clusters. Here we use $V_{min} = 5$ and $V_{max} = 20$.

between clusters. As shown in Figure 3, this MST connection is robust against single journal deletion. It is interesting to note that, although some connections between clusters are modified, the general features observed in Figure 2 still hold in Figure 3.

After demonstrating the applicability of our method in clustering scientific journals, we applied this method to construct an MST of all SCI journals of impact factor greater than 1 using $V_{max} = 50$ and $V_{min} = 5$. Figure 4 shows a collection of physical clusters (Supporting information, 2005). This physical domain is constructed by first locating the largest cluster in physics (the most popular word in the journal titles of cluster members is *phys*). Then, using this cluster as the core, its nearest-neighbor clusters are included into this physics domain. We also include the next-nearest-neighbor clusters into the physics domain if the majority members of a nearest-neighbor cluster contain *phys* in their journal titles. The largest cluster of this domain is related to condensed matter physics. The DJ of this cluster is *PRL*, which suggests that *PRL* is dominated by condensed matter physics, although it is categorized as physics + multidisciplinary. Above the condensed matter physics cluster are the particle physics cluster (DJ is *Phys Rev D*) and the astrophysics cluster (DJ is *Astrophys J*). On the right hand side are statistical physics (DJ is *Phys Rev E*), fluids (DJ is *J Fluid Mech*), mathematical physics (DJ is *Commun Math Phys*), and plasmas (DJ is *Phys Plasmas*). Below the main cluster, we see chemical physics (DJ is *J Chem Phys*), optics (DJ is *Appl Optics*), surface physics (DJ is *Thin Solid Films*), and applied physics (DJ is *IEEE Trans Electron Devices*). In general, an MST of scientific journals can be constructed very efficiently for a large database and details of various knowledge domains can be investigated with the desired resolution.

DOI: 10.1002/asi

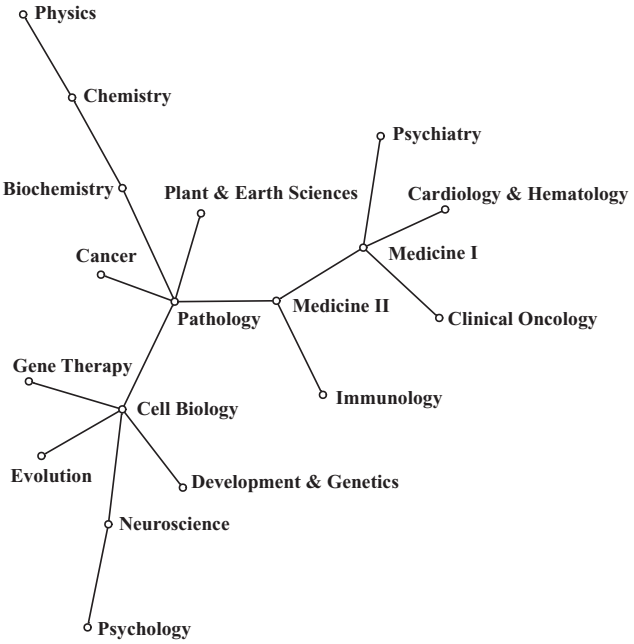


FIG. 3. A cluster MST of SCI core research journals of impact factor greater than 4. Each cluster is represented by a circle and labeled by a number. The line length is related to the distance between two connected clusters, but distortion could occur due to a two-dimensional map of the MST.

Constructing the Map of Science

Although MST is useful in clustering journals, the information about similarity between unconnected journals is missing. Therefore we desired to construct a map of scientific journals, which can be used to visualize the panorama of the scientific world and to study the temporal evolution of science. Using the database of scientific journals of impact factor greater than 1, the largest 20 clusters found in MST were studied and their cumulative citation matrix was calculated. Each cluster of journals is named by the most popular words that appear in the titles of its members. The squared distance matrix $\{d_{ij}^2\}$ can be obtained by Equation (2). The CS method is used to find a set of vectors $\{y^\mu\}$ such that the squared distance matrix between the $\{y^\mu\}$ points matches $\{d_{ij}^2\}$ as closely as possible. The dimension of the data points is chosen to be small, either two or three, so that we can visualize the data. Here a two-dimensional mapping of these scientific clusters is created by projecting our CS results onto the two principal axes. To minimize the distortion of the scientific world map, the coordinates of clusters in the map are further optimized by SM, which minimizes the following cost function:

$$E = \sum_{i<j} \{d_{ij} - d(y^i, y^j)\}^2 \Big(\sum_{i<j} d_{ij}^2 \Big)^{-1}, \qquad (3)$$

where the summation runs over the dataset under investigation, and *d(a,b)* is the distance between points *a* and *b*. Although SM constructs a nearly global optimization for the coordinates of various knowledge clusters, the distortion in the distance between clusters due to the two-dimensional mapping is usually comparable with the yearly temporal change of distance. This distortion incapacitates the attempt to study the temporal evolution of the scientific world through the constructed science map. To resolve this problem, we locally adjust the coordinates in the same domain by minimizing the cost function in Equation (3) for each domain and the temporal evolution of the domain is mapped by Procrustes analysis. Procrustes analysis is a method of comparing two configurations, which matches corresponding points from each of the two configurations by minimizing the sum of squared differences between the two configurations. Therefore, the yearly shift of each cluster simply results from the yearly change in the accumulated distance matrix.

As shown in Figure 5(a), the constructed map of science for year 2001 shows a linear structure of the scientific world, which contains three major domains including medical sciences, life sciences, and physical sciences (Supporting information, 2005). This result is consistent with the science mapping of Small (Small, 1999), in which biology is found to locate in between physical sciences and medical sciences. The linear structure of our science map is a direct result of the distance conversion in Equation (2), because the distance between two unrelated clusters is $\sqrt{t^{-1} - 1}$ (infinite for $t = 0$). The structure of the science map could change if a different distance conversion formula were adopted. For example, the constructed science map has a ring structure if we use $d_{ij} = 1 - s_{ij}$. In this case, the distances between all pairs are less than unity, which leads to a ring structure of the constructed science map. Such a ring structure of science has also been proposed by Boyack and others (Boyack, Klavans, & Börner, 2005, in press). Nevertheless, we consider this ring structure as an artifact due to an improper distance conversion. Although our science map only contains the 20 largest clusters, we expect that this linear structure remains stable after adding smaller clusters, because these smaller clusters will not affect the distance between two unrelated clusters. For cell biology in year 2001, the amount of incoming citations is 295,734 and that of outgoing citations is 129,228. The incoming citations of physics, cell biology, and general medicine are many more than their outgoing citations, which indicates their importance in their knowledge domains. The gap between physical sciences and life sciences is quite obvious, whereas the gap between life sciences and medical sciences is small. This is due to the intensive interaction between cell biology and medical sciences. The average distance between knowledge clusters in the domain of medical sciences is much shorter than that in the other two domains. This is a clear indication for the high correlation among knowledge clusters in medical sciences. Moreover, the property of each knowledge cluster can be judged from this map. For example, the focus of neurology is clearly different from that of neuroscience. Neurology is close to general medicine, while neuroscience is closely related to cell biology. The MST structure of these clusters as shown by the solid lines is consistent with our two-dimensional mapping, which confirms our global minimization of Equation (3). An enlargement of

DOI: 10.1002/asi

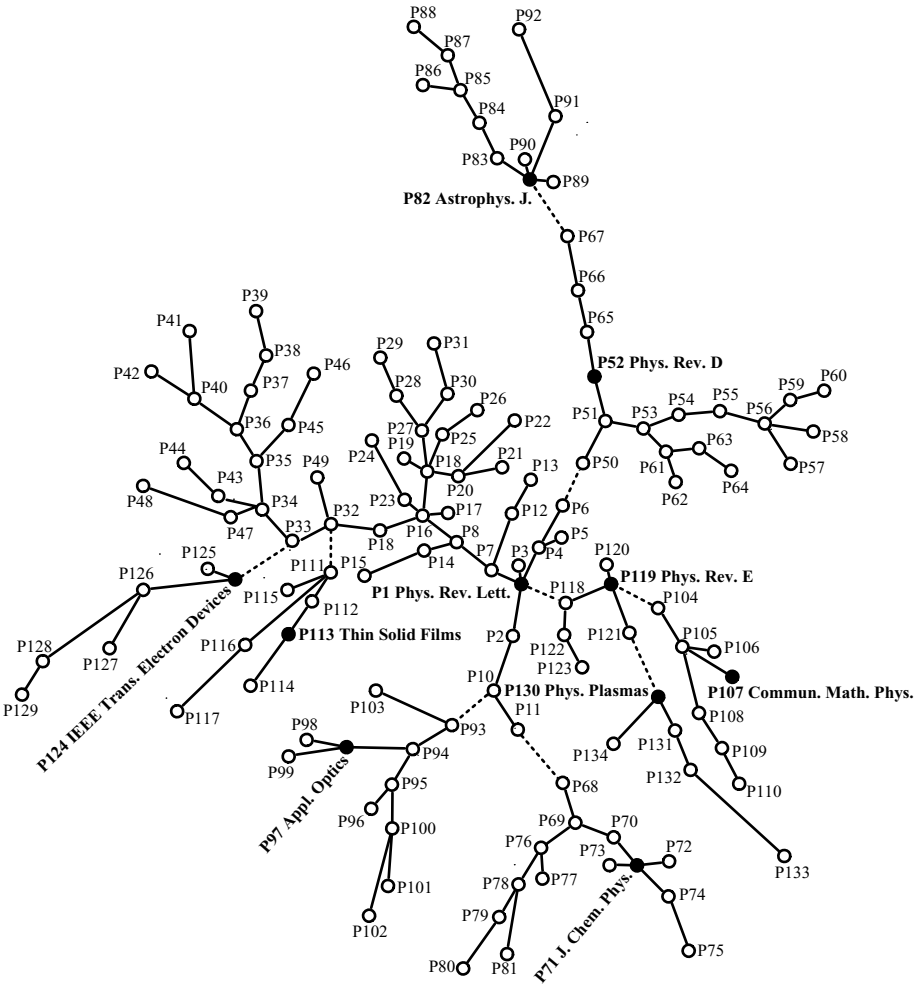


FIG. 4. An MST of physical journals in SCI database of impact factor greater than 1. Each journal is represented by a circle and labeled by a number. Titles of all journals can be found in the supplemental data. Journals in the same cluster are connected by a solid line and clusters are connected by a dash line. The line length is related to the distance between two connected journals, but distortion could occur due to a two-dimensional map of the MST. Filled circles are DJs of these clusters.

the domain of physical sciences is drawn in Figure 5(b), which shows the temporal evolution of physical sciences during 1994–2001. The average distortion for the two-dimensional mapping is about 10% of the maximal temporal shift in the distance between two knowledge clusters. During these eight years, it seems that analytical chemistry is running away from the rest of domain members and constantly drifts toward the domain of life sciences. Note that here we are not attempting to give a complete view of the complex scientific network. Putting all the information together will simply confuse the readers. Instead, we demonstrate a feasible approach to understanding the scientific network. The details of the scientific network can be investigated by choosing a knowledge area with the desired resolution. We further note that, although only a portion of journals is selected in our analysis, these selected journals are quite representative. Garfield had argued that the significant scientific literature appears in a small core of journals (Garfield, 1996). In fact, in the 1987 SCI, 500 journals account for half of what is published and more than 70 percent of what is cited, and 2,000 journals published about 85 percent of all SCI-indexed articles that year and 95 percent of cited articles. Our 20 clusters actually cover 763 journals (more than one third of journals of IF > 1) and contain almost all the most-cited and most-productive journals. The limitation of choosing these dominant journals actually helps us in analyzing the scientific database and understanding our scientific world. Nevertheless, we admit that some knowledge fields with low impact factors, such as mathematics and computer sciences, are not considered in this analysis.

The statistical properties of cluster citations are depicted in Figure 6 for year 2001. Statistically, the mean self-citation of physical sciences is 0.75 with a standard deviation of 0.09. The mean self-citation of life sciences is 0.60 with a standard deviation of 0.06. The mean self-citation of medical sciences is 0.45 with a standard deviation of 0.08. The standard deviations of these three domains are of the same order, whereas their mean values vary a lot. We thus conclude that the citation culture of journals is consistent within the same domain but differs a lot for different domains.

DOI: 10.1002/asi

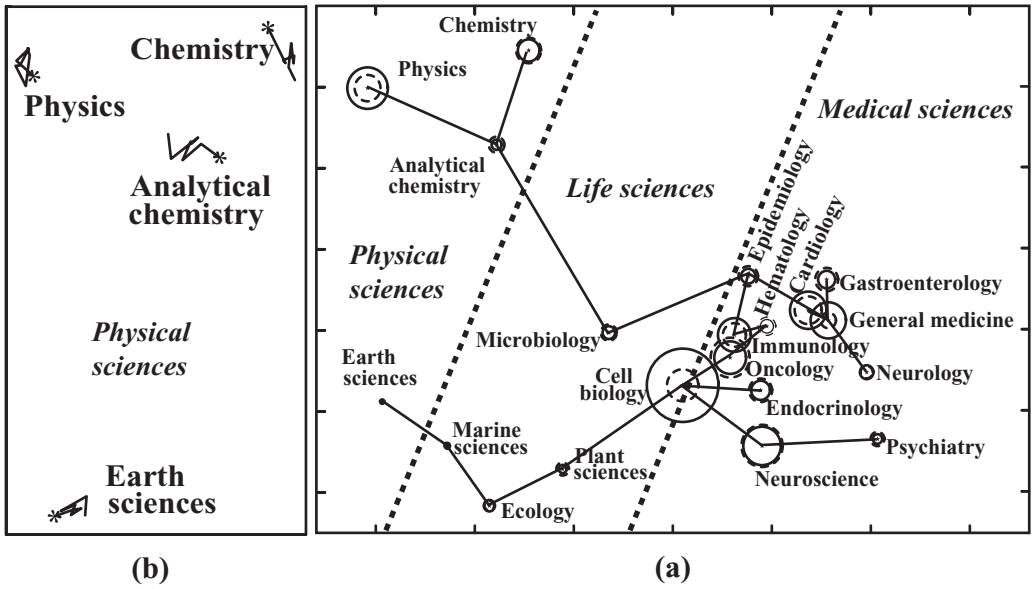


FIG. 5. A two-dimensional map of the scientific world (a) and its temporal evolution in the domain of physical sciences (b). The distance (in arbitrary units) between two knowledge clusters is related to the similarity in their citation patterns. In (a), solid and dash circles indicate the number of incoming and outgoing citations, respectively, for the 20 clusters shown in the map. Solid lines show the MST structure of these 20 clusters. Dash lines are deliberately sketched to indicate domain boundaries. In (b), solid lines show the temporal evolution of each cluster in physical sciences during 1994–2001. Stars represent their positions on the map in 2001.

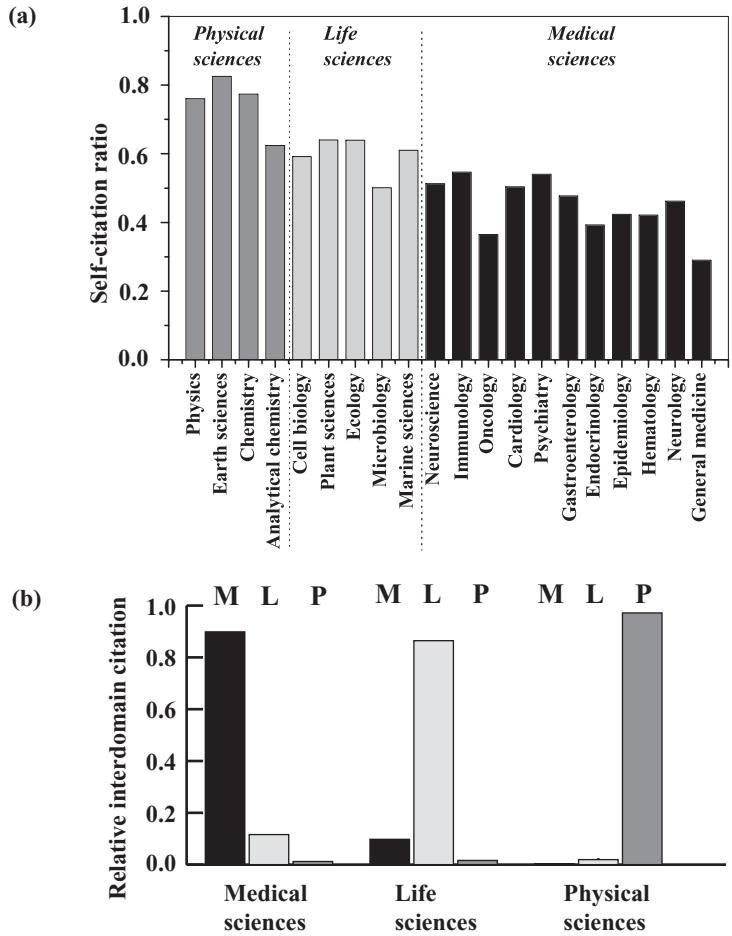


FIG. 6. Statistical citation properties of clusters (a) and domains (b). In (a), self-citation ratio of each cluster is shown for three domains. In (b), relative interdomain citations are shown for three domains. Here *M* represents medical sciences, *L* represents life sciences, and *P* represents physical sciences.

DOI: 10.1002/asi

As shown in Figure 6(b), the relative interdomain citations for the domain of physical sciences are negligible. Based on our limited data set, it seems that research results in medical sciences have little effect on the discoveries made in physical sciences. The domain of physical sciences seems to separate from other domains of science. On the contrary, interdomain citations between medical and life sciences are visible and nearly identical. This indicates that research results in these two domains might promote new discoveries made in one another.

Conclusion

In this paper, we have presented an approach to constructing an MST and a two-dimensional map of scientific journals. This approach can be applied to various levels of knowledge domains. As a demonstration, we tested this approach using 2,000 SCI journals of impact factors greater than 1. The structures of constructed MSTs at various knowledge levels are consistent with the sorting of SCI categories. Therefore, this approach could be very useful in automatically clustering scientific journals. The connections between related knowledge clusters can be obtained in a glance at these MSTs. Particularly, we find that the cluster of *PNAS* serves as a hub of biomedical sciences, and the cluster of *Nature* and *Science* serves as a bridge connecting the two domains. In addition, we also constructed a two-dimensional map of the scientific world based on our MST results. Such a map shows the relation of various knowledge clusters and their citation properties. Extending our present work by using a more complete data set might give a better view of the scientific network. The temporal evolution of the scientific world map can also be delineated. Our results suggest that a two-dimensional map of science can be constructed to help comprehend various knowledge domains and to capture the structure and evolution of mankind's scientific endeavor at a glance.

Acknowledgments

The authors are grateful to the Science and Technology Center of National Science Council of Taiwan for providing the SCI database. This work is supported, in part, by the National Science Council of Taiwan under grant no. NSC 93-2112-M-003-004.

References

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

Börner, K., Chen, C., & Boyack, K.W. (2003). Annual Review of Information Science & Technology, 37, Chapter 5, 179–255.

Börner, K., Maru, J.T. & Goldstone, R.L. (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences of the United States of America, 101, 5266–5273.

Boyack, K.W. (2004). Mapping knowledge domains: Characterizing PNAS. Proceedings of the National Academy of Sciences of the United States of America, 101, 5192–5199.

Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64, 351–374.

Carpenter, M.P. & Narin, F. (1973). Clustering of scientific journals. Journal of the American Society for Information Science, 24, 425–436.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America, 101, 5303–5310.

Davidson, G.S., Hendrickson, B., Johnson, D.K., Meyers, C.E., & Wylie, B.N. (1998). Knowledge mining with VxInsight: Discovery through interaction. Journal of Intelligent Information Systems, 11, 259–285.

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R.A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391–407.

Doreian, P. & Fararo, T.J. (1985). Structural equivalence in a journal network. Journal of the American Society for Information Science, 35, 28–37.

Fruchterman, T.M.J. & Reingold, E.M. (1991). Graph drawing by force-directed placement. Software-Practice & Experience, 21, 1129–1164.

Garfield, E. (1996). The significant scientific literature appears in a small core of journals. The Scientist, 10, 13–16.

Gordon, A.D. (1981). Classification: Methods for the exploratory analysis of multivariate data. London: Chapman and Hall.

Gower, J.C. & Dijksterhuis, G.B. (2004). Procrustes problems. Oxford: Oxford University Press.

Griffith, B.C., Small, H.G., Stonehill, J.A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure of science. Science Studies, 4, 339–365.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14, 10–25.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., et al. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11, 574–585.

Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. Proceedings of the American Mathematical Society, 7, 48–50.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115–129.

Lee, R.C.T., Slagle, J.R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from *N*-space to two-space. IEEE Transactions on Computers, 26, 288–292.

Leydesdorff, L. (1986). The development of frames of references. Scientometrics, 9, 103–125.

Leydesdorff, L. & Cozzens, S.E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the *SCI*. Scientometrics, 26, 135–156.

Narin, F., Carpenter, M., & Berlt, N.C. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science, 23, 323–331.

Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, 101, 5200–5205.

Price, D.J.D. (1965). Networks of scientific papers. Science, 149, 510–515.

Schvaneveldt, R.W. (1990). Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex Publishing.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265–269.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50, 799–813.

Small, H.G. & Griffith, B.C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. Science Studies, 4, 17–40.

Supporting information can be found at the Web page http://www.phy.ntnu.edu.tw/~cchen/paper/map.htm.

Thurstone, L.L. (1931). Multiple factor analysis. Psychological Review, 38, 406–427.

Tijssen, R., de Leeuw, J., & van Raan, A.F.J. (1987). Quasi-correspondence analysis on square scientometric transaction matrices. Scientometrics, 11, 347–361.

DOI: 10.1002/asi