Jointly published by Elsevier Science Ltd, Oxford and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 38, No. 2 (1997) 275–293

UPDATE ON SCIENCE MAPPING: CREATING LARGE DOCUMENT SPACES

H. SMALL

Institute for Scientific Information, 3501 Market Street, Philadelphia, Pennsylvania 19104 (USA)

(Received October 11, 1996)

Science mapping projects have been revived by the advent of virtual reality software capable of navigating large synthetic three dimensional spaces. Unlike the earlier mapping efforts aimed at creating simple maps at either a global or local level, the focus is now on creating large scale maps displaying many thousands of documents which can be input into the new VR systems. This paper presents a general framework for creating large scale document spaces as well as some new methods which perform some of the individual processing steps. The methods are designed primarily for citation data but could be applied to other types of data, including hypertext links.

Introduction

The mapping of science attempts to find representations of the intellectual connections within the dynamically changing system of scientific knowledge.[1,2] The formal bibliographic citations or footnotes in scientific papers offer us a unique view of these connections. In the world of scholarly and scientific literature, bibliographic citations serve the purpose of pointing to source materials used by the author, and also as acknowledgements of intellectual debt and priority. Hence, citations trace information flows within the scientific community.[3]

Of the many mapping efforts published to date we can distinguish two types: 1) those that aim to map a particular topic, subject domain, or retrieved set of items, and 2) those that aim to map an entire database. The latter, global approach, may provide a new visual paradigm for information retrieval based on the navigation of an information space. Beyond this, however, the rationale for mapping science is closely tied to our interest in understanding and facilitating the discovery process. If Swanson is correct and discovery can be modeled as a recombining of what we already know to find what we do not yet know,[4] then constructing a map of relationships of the contemporaneous branches of knowledge could show us areas that are proximate and

0138–9130/97/US $ 15.00

facilitate the making of new connections between them. At the very least such a map can aid us in tracking how these relationships change as new discoveries are made, and perhaps lead to a more informed management of science and information.

The goal of constructing large scale maps of science has been revived by the advent of virtual reality software and hardware capable of navigating large synthetic landscapes.[5] A large scale mapping of science could be defined as one involving the positioning of over 1,000 documents, but a comprehensive multidisciplinary mapping of science might involve the positioning of hundreds of thousands of documents which constitute the research front of modern science. Two large scale mappings are reported below.

Central to the mapping or visualizing of information is the procedure for fixing or positioning objects in space, commonly known as ordination.[6] Most of the techniques use (dis)similarities between objects as input data and solve a mathematical minimization problem to arrive at coordinates. We suggest an alternative method based on simple geometric calculations.

The Humpty–Dumpty method

Two strategies to achieving a global mapping are 1) to scale up one of the classical ordination methods so that it can deal efficiently with large data sets, and 2) to break the database into smaller chunks by clustering, ordinate each cluster, and then reassemble the pieces into an overall structure. The latter approach might be called the Humpty-Dumpty method.

There are three steps in the Humpty-Dumpty process of creating comprehensive maps: 1) creation of a multi-level hierarchy of clusters or partitions starting with individual documents; 2) ordination of objects within each cluster in the hierarchy, providing a two or three dimensional representation of each group; and 3) the integration of the structures of each cluster into a global structure or common coordinate space. This last step puts the pieces back together and involves expansion, translation, and rotation of clusters at each level.

Earlier clustering work at ISI using co-citations was concerned with building up a nested hierarchy of clusters,[7] the objects in each cluster being plotted by multidimensional scaling.[8] The present effort is an extension of the earlier work in that it fits the various maps together into a common coordinate space in which each document and cluster centroid is assigned an x, y coordinate.

This paper introduces a number of new techniques for creating large scale maps of science: 1) a new citation based measure of document similarity; 2) a simplified method

of ordination termed triangulation; 3) a method for creating a common coordinate space; and 4) a visualization capable of showing hierarchical relationships. The approach we take allows the substitution of different measures of article similarity (e.g. language based versus citation based), and different methods of clustering and ordination than those we have utilized. Hence it represents a general framework for achieving a global mapping.

We will not discuss the many mapping efforts that utilize linguistic data, such as co-word, co-term or co-classification analysis.[9] The most significant example of this approach is the language based visualization work of *Wise* et al.[10]

A new measure of document similarity: Combined linkage

Criticisms of the co-citation methodology have centered on the low "recall" of clusters, i.e., the ability of the method to classify only a fraction of the papers which have citation links.[11] Various proposals have been made to improve "recall" including augmenting individual clusters[12] and using co-words in addition to co-citation.[13] We introduce here a new citation similarity measure which markedly improves the recall of citation based clustering.

From experiments with ISI data, we know that the overall citation network is sparsely linked, however, with localized regions of high linkage density. Within a specific subject area, for example analytical chemistry, as much as 98% of the items can be connected into a single sparsely connected graph, yet only 0.0002% of items which could be connected are connected, ignoring limitations on reference list lengths. The rationale for using indirect linkages, such as co-citation and bibliographic coupling, is that they reinforce regions of dense direct citation and thereby facilitate the breaking up of the network into highly linked chunks by simple thresholding.

Taking into account the publication dates, there are three ways two papers can be connected by taking two steps on the citation network: 1) bibliographic coupling,[14] which connects papers by one step back then one step forward; 2) co-citation which connects papers by taking one step forward then one step back, and 3) a third form which connects older and younger papers by taking either two steps forward or two steps backwards. This third form has been called longitudinal coupling,[15] because it is capable of connecting papers across many years.

Since our aim is recover as much structure as possible from a multi-year citation network, it is perhaps best to combine all three into a single combined measure.[16] In addition, empirical work on social networks suggests that more reliable structural information can be obtained by use of direct contacts, rather than indirect ones.[17] This

suggests the idea of combining the direct citation link with all forms of the indirect linkage into a single measure, and to give the direct citation link a weight equivalent to two indirect links, in recognition of its greater significance.

Finally, the new combined measure needs to be caste as a coefficient that varies between zero and one. Such a normalization needs to take into account the total number of links, whether references or citations, incident on each node of the connected pair. The final combined measure is shown in Fig. 1, giving an example of a perfectly linked pair.

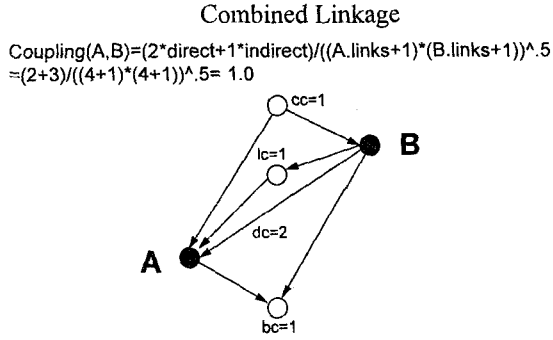

Fig. 1. The shaded circles A and B represent two documents linked by a direct citation (dc) and by three forms of indirect citation linkage: co-citation (cc), longitudinal coupling (lc), and bibliographic coupling (bc). The normalized linkage formula weights the direct citation by two and each indirect citation by one. Since no other links are incident on A or B, the documents have the maximum possible linkage coefficient of 1.0

Experiments using the analytical chemistry dataset, to be reported elsewhere, indicate a doubling in the recall rate of clusters when the combined linkage measure is used compared to the use of co-citation alone (an increase from about 40% to 80% of papers contained in clusters of size 2 or greater).

The citing/cited time window

Working with combined linkage requires a shift in our thinking about how time interacts with our representations. Clearly the role of the different coupling forms will vary with the length of the citing and cited time periods used. For example, co-citation and bibliographic coupling offer cross sectional views given narrow, one year citing periods. But longitudinal coupling becomes effective only when wider periods are used. At the end of a time period documents will be linked through their references to earlier

items, while at the beginning, linking will be through citations received. At mid-period, the items will link through both their references and citations. A cluster created from these data will then be a mix of current and older items, linked through a variety of modes.

When working with combined linkage we define a time window for both citing and cited year. For example, in the physical science dataset discussed later this time window has been set to five years, 1990 to 1994. This means that we consider only those citation links whose citing and cited documents have publication years that fall within the five year time window. In addition, we impose a threshold of 15 on the sum of citations received and references given. For example, a 1994 document making 15 references to documents in the 1990–1994 period and a 1990 document receiving 15 citations from documents in the same period can be selected. These two documents can then be linked by a direct citation or by indirect longitudinal couplings. The advantage of thresholding on the sum of citations plus references is that it samples both current and older items fairly evenly, which can then mix together in the clusters generated. When we use either co- citation or bibliographic coupling alone, this time sampling is skewed to either an older or a more recent set.

The cluster hierarchy

In earlier work, a modification of the single-link clustering method had been developed aimed at limiting the amount of chaining to which this method is prone.[18] This involves setting a maximum cluster size, a starting linkage threshold, and linkage increment. The method finds the lowest threshold possible to create a cluster not greater than a preset size limit. If a cluster is greater than this size, the linkage threshold is incremented and the cluster regenerated. This procedure is equivalent to picking clusters by trimming branches from the full single-link dendogram which are not greater than some maximum value.

The above clustering process is performed iteratively: the clusters of documents formed in the first step are used as objects for the next step which forms clusters of clusters, and so on, until the desired degree of consolidation is reached, usually when most objects fall into a single macro-cluster. Between each iteration, the combined linkage measure is recomputed as if the objects connected to each other were documents that cite each other, in effect collapsing the citation network. Thus at the second level of iteration, we have clusters that link with one another directly or indirectly, just as documents were linked at the first level.

As the cluster hierarchy is built, it is possible for objects to become isolates at any level. Documents may be isolated in the initial clustering and clusters can become isolates by not being incorporated into the next higher level. This isolate formation can be caused by the absence of links or the failure to meet threshold requirements. This need not mean that such objects will be excluded from the final map, since they can be represented as "islands" separate from the mainland.

Converting similarities to distances (the "Garfield")

The combined linkage coefficient, which varies between zero and one, needs to be converted into a distance in order to create two or three dimensional structures. Three types of distance transformations were tested based on the combined linkage measure: inverse, logarithmic and linear. The linear version simply subtracts the coefficient from one, to form a dissimilarity or distance measure. The log version takes the negative log of the similarity, since the log of a number less than one is negative. The inverse measure uses the reciprocal of the similarity value. In all cases objects with a similarity of one (perfect similarity) have zero distance.

In addition, distances are made relative to the threshold of the cluster in which they originate. This means that the maximum distance, and thus the weakest link, between linked documents is the same across all clusters and equal to one. This insures that the sizes of the clusters in the display will be of a similar scale and roughly proportional to the number of objects they contain, and also provides the space with an absolute distance metric. We will call this underlying unit of distance "the Garfield" in honor of the father of the *Science Citation Index*.

The formulas for the three relative distance measures are:

linear distance = (1- similarity)/(1- threshold)

log distance = (log(similarity))/(log(threshold))

inverse distance = ((1/similarity) -1)/((1/threshold) -1)

On empirical grounds the linear distance measure behaves the best in our ordination by triangulation, because there is less variation in the distribution of distances within a cluster. Wider variations in the distances can lead to more frequent violation of the triangle inequality. This criterion for performance is based on the number of links utilized by the triangulation process, under the assumption that the more distances used, the more accurate the spatial representation.

Ordination by triangulation

An alternative to classical ordination methods such as multidimensional scaling and correspondence analysis is a method based on simple geometric triangulation.[19,20] The triangulation approach focuses on fitting the strongest links and the resulting configurations are order dependent. Nevertheless, by keeping the clusters of objects small (e.g. less than 100), we can take advantage of the speed of this method while avoiding its limitations. This method was first used in a document visualization application in the SCI-Map software package.[21,22]

Ordination of objects in a cluster by triangulation can be done immediately after a cluster is formed. A byproduct of the clustering process is a table of links between the clustered objects. The triangulation process starts with the selection of a seed item and places it at the origin of a two or three dimensional coordinate system. Then the closest object to the seed is found in terms of linkage strength and is placed along the x axis at a distance given by one of the distance formulas above. The third object to be plotted is the one having the maximum sum of strengths to points already plotted. If the third point is linked to both the first and second points then its position is set by triangulating on the first two distances, unless the distances violate the triangle inequality. In this case we revert to the one link case and plot the point at the specified distance but furthest from the center as possible. This causes the cluster to grow outward from its center.

For two dimensional triangulation we can use up to three distances. The third distance can be used to select which of the quadratic solutions of the two distance triangulation is the better of the two. If a third distance is not available, the new point is plotted at the solution that is the furthest away from the centroid of already plotted points. After all the objects in the cluster are plotted, the center of the configuration is translated to the origin.

The algorithm described above was essentially that used in the SCI-Map system. A recent improvement in this methodology is to select the next point to plot based not only on its having the highest cumulative linkage strength, but also on its ability to be plotted utilizing the maximum number of links available, on the assumption that the more links used the better the representation. In addition, if a point fails to triangulate with the two strongest links, another attempt is made using the third strongest link in combination with the strongest, and, if that fails, with the second and third strongest links.

In the three dimensional case we can use up to four distances from the new point to already plotted points: three to position the new point in one of two positions governed

again by the quadratic solution, and a fourth distance to select which of the two quadratic solutions is the best.

Since the triangulation process is order dependent, but fast, each object in the cluster can be tested as a seed and a record kept of the seed which utilizes the greatest number or sum of link strengths. This configuration is selected as the final one. Typically in selecting which seed is the best, the spread from highest to lowest sum of link strengths is about five percent. These differences in the sum of link strengths arise from the different paths which can be taken through the network.

Another measure we can use to ascertain how completely the linkage data is utilized is to compare the number of links actually used in plotting an object versus the number of links available to that object. For triangulation in two dimensions, the reason an object uses fewer than two links to fix its position is either that more are not available or the links provided violate the triangle inequality. In a run of an analytical chemistry dataset, the percentage of objects plotted in which the links used were less than the links available was only 1.5% (85/6014). The triangle inequality was violated in about 3.5% of the cases where two links were available. With single-link clustering the mean links per object for plotting is 1.5, which creates predominantly elongated structures. The links per object can be increased by using, for example, complete-link or average link clustering.[23]

The common coordinate space

Our objective is to merge the various levels of clustering into a common coordinate space such that the hierarchical relationships are reflected in the positions of objects and the relative locations of objects within and between levels is preserved. Suppose we begin with four levels of clusters, where the objects in a given level contain objects within the previous level. For each object corresponding to a cluster we have a configuration of its member objects (either clusters or documents for the first level), with coordinates assigned to each member, centered at the origin.

The strategy is to begin with the most disaggregated level, what we call the first level of the hierarchy, where the objects are document clusters, and expand the coordinate space as we move up the levels in the hierarchy toward the root (to reverse the usual convention). Then we move back down the hierarchy, translating the locations of the member objects within each cluster to the new expanded centers of each cluster.

For example, suppose we have two clusters of documents which are contained in the same cluster at level 2. Clearly, we need to expand the size of the coordinate space for the level 2 cluster in order to accommodate the document configurations of the two level 1 clusters. Once this space has been expanded by multiplying the coordinate

values of the two level 1 objects by some expansion factor, we can then translate the positions of all documents in the clusters by adding the coordinate values of the respective level 2 objects to each of the document coordinates. Then the position of the higher level object becomes the centroid of the lower level objects.

To determine how much to expand each higher level object, we find the radius of each object contained within it. This radius defines the circle (or sphere in 3D) which is just large enough to contain all the objects within it. Since we know the distance between objects before expansion, we can compute how much to expand the coordinates so that, in the worst case, the largest adjacent circles or spheres would be exactly tangent.

For sake of efficiency, all expansions are performed first moving up the hierarchy. Then each object need be translated only once to its expanded centroid. When we reach the level just prior to the most aggregated level of the hierarchy (the root), the coordinate system of each branch of the tree has been expanded, preserving the relative locations of objects within it. The centroids of the different coordinate systems are still however at the origin, so now we must return down the hierarchy, and translate the coordinates of objects within each cluster to their new expanded locations. By the time we return to level 1, the location of a specific document will reflect a cumulative series of expansions, one for each of the levels, moving the documents for any given cluster, possibly, quite far from the origin. Only the coordinates of objects at the highest level of aggregation remain centered at the origin.

Table 1
Radii of objects in expanded coordinate system

Dataset	Level	# Objects	Mean radius (Garfields)
Anal. Chem.	doc	4,797	na
	1	718	1.4
	2	106	9.2
	3	15	59.3
	4	1	4,652.2
Phys. Sci.	doc	27,547	na
	1	3,486	1.5
	2	460	18.3
	3	50	431.5
	4	1	107,746.7

Table 1 shows examples of the degree of expansion obtained by two test datasets: the dataset consisting of about 4,700 papers from the journal *Analytical Chemistry* covering the years 1981–1993, and a physical science dataset consisting of about 27,500 papers covering the five year period 1990–1994. The table shows the level of clustering, the number of objects for levels 1 through 4, and the average radius of objects at that level in Garfields. The radius at the fourth level for physical science indicates that one cluster has been moved over 100,000 Garfields from the origin.

The changes in the radii from levels 1 through 4 give an idea of the degree of expansion the coordinate system has undergone. For analytical chemistry this amounted to a total expansion of about 3,000 fold, and a 73,000 fold expansion for the larger physical science dataset. In addition, the document density for the analytical chemistry space is about a factor of 10 higher, namely, 1.7×10^{-4} documents per square Garfields, than for the physical science space.

Cluster orientation

Since we plot objects within clusters centered at the location of the cluster, we are free to rotate a cluster about its centroid. This offers another way to reduce the distance between closely related objects, namely, by orienting clusters to one another through rotation. The problem is difficult because there are multiple clusters each containing many objects with many relationships. We implement an approximate solution, namely, use a minimal spanning tree pathway through the cluster, and orient each successive cluster to its predecessor in the path. Since we have generated a minimal spanning tree for each cluster as part of the triangulation process, we can use this sequence for orienting clusters to one another.

Successive clusters along that path can be oriented to one another using the following method: Compute the strength of linkage from each member of the cluster to be oriented relative to the fixed cluster, and compute a center of "gravity" of the cluster with respect to the fixed cluster. This is the weighted average location of objects that have some linkage value to the fixed cluster. This establishes a point within the cluster to be rotated that should be as close as possible to the fixed cluster. We then rotate the cluster until this center of gravity is as close as possible to the fixed cluster. Tests with the analytical chemistry dataset indicate that the mean angle of rotation is about 75 degrees, and 88 percent of the clusters undergo rotation of some amount.

The volvox visualization: Examples from astrophysics

We are now ready to plot the configuration of documents and clusters. Since our goal is to have the final mapping in 2D or 3D reflect the hierarchical structure of documents and clusters, in two dimensions we can think of this as a series of non-overlapping circles inside of other circles, where an outer circle represents a cluster and the inner circles the objects it contains. In three dimensions, of course, the circles become spheres. This could be termed the volvox representation in recognition of its similarity to the microorganism of the same name.[24] Circles were a natural choice because a radius for each cluster was calculated in the process of expanding the coordinate space. In addition to circles, links are drawn between objects and represent the strong linkages which were used in the ordination process. Links are useful in clarifying relationships in complex maps.

Our results are drawn in two dimensions using a PC interface implemented in Visual Basic. Work is currently underway with Sandia National Laboratory to transform this static display into a virtual reality application, allowing a navigation through the document space.

The dataset on the physical sciences covers the five year period 1990 to 1994 and focuses on papers having a sum of citations received plus references given of 15 or more. There are about 27,500 documents on the resulting map. Figure 2 shows the overall map at level 4 and gives topic labels for the main level 3 groups. Physics is on the lower left and chemistry is at the top right with materials science in between. Areas of physics closest to the bottom are fundamental areas such as particle physics and astrophysics. Just above these are areas of condensed matter physics including superconductivity. Materials science areas are above physics and include topics such as C60, catalysis and cluster compounds. Organic and inorganic chemistry are closest to the top. Thus, the main feature of this global map is a physics-chemistry axis.

To illustrate the detail structure we zoom in on the region labeled astrophysics, at the bottom left of the global display. Figure 3 shows this single level 3 cluster containing 23 level 2 objects. The largest cluster here deals with galaxies and the red shift. We then zoom in on the region around the level 2 cluster near the center labeled quantum gravity (Fig. 4). This region also includes clusters on topics such as black holes, quantum cosmology and string field theory.

Zooming in on the level 2 cluster for "black holes" at the upper left we get Fig. 5. Some of the level 1 clusters contained within it have been labeled. Of these we select the "2D black holes" cluster to view in detail.

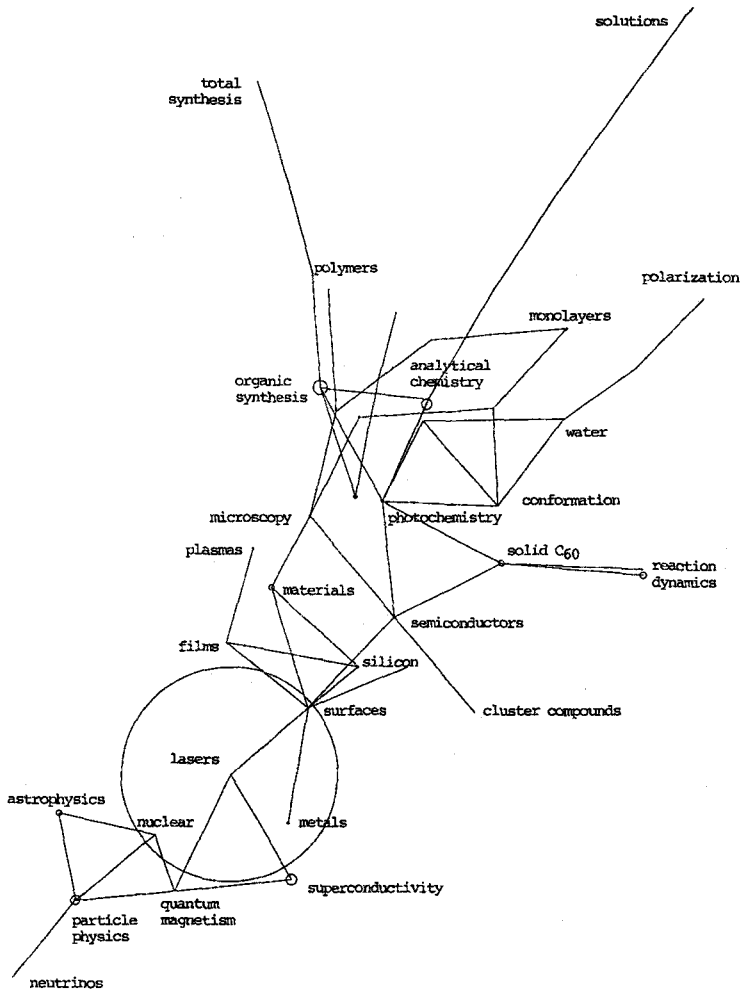

Fig. 2. The map is a two-dimensional representation of 27,547 documents in the physical sciences from the years 1990 - 1994, selected by having at least 15 references plus citations. The vertices or circles represent 50 third-level clusters which are joined together in the fourth level of the hierarchy. Lines are strong links among the 50 superclusters. The size of the circle indicates the spatial spread of lower level objects. The subject matter ranges from fundamental physics at the lower left to organic and inorganic chemistry at the upper right

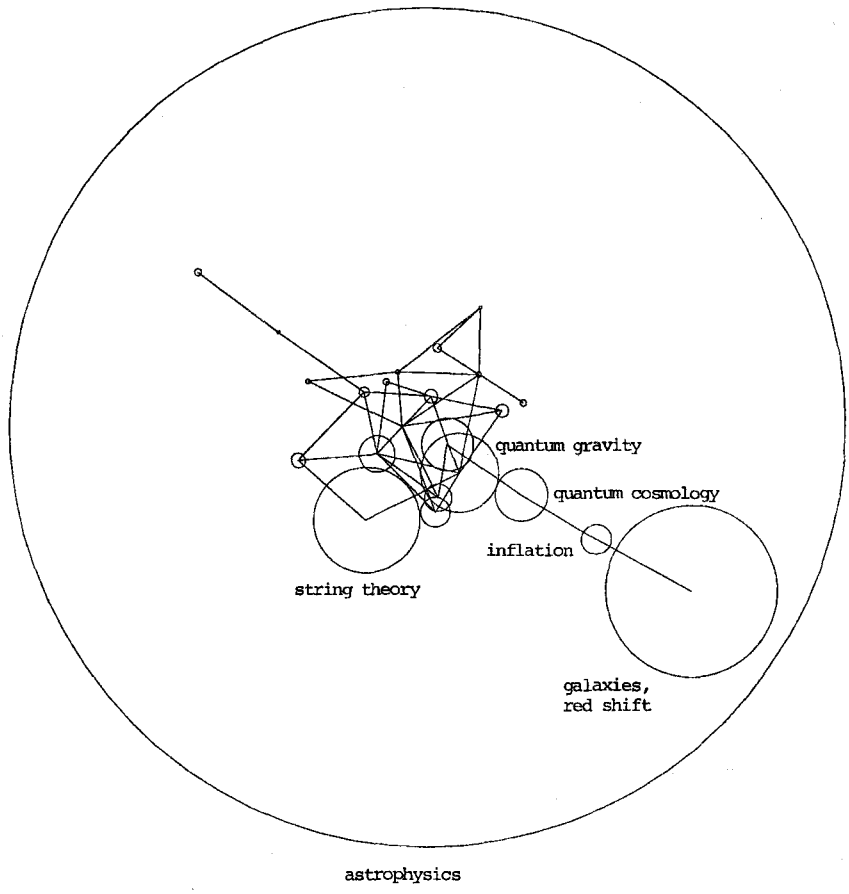

Fig. 3. A detailed view of the astrophysics cluster at the lower left of Figure 2.
This third-level cluster contains 23 second-level clusters whose links with one another are shown

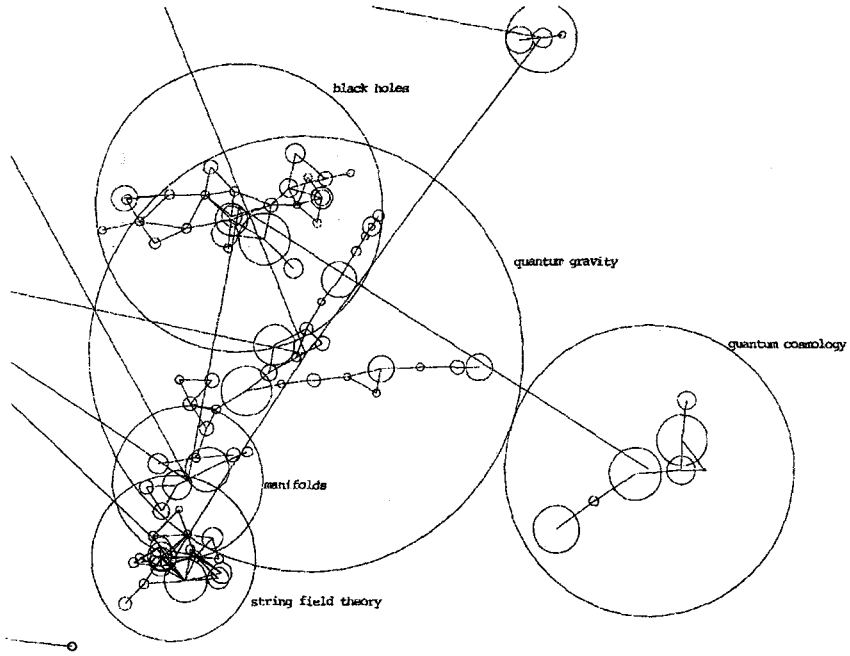

Fig. 4. A zoom in on the central region of the astrophysics cluster (Figure 3) showing five second-level clusters on quantum gravity, black holes, quantum cosmology, string field theory, and manifolds. The first-level clusters within each second-level object are shown, as are the links for both levels

Finally we arrive at the individual documents within the 2D black hole area, labeling the documents with their year of publication and indicating the titles of the oldest and youngest papers (Fig. 6): one paper from 1990 and one from 1994, the beginning and end points of the dataset. The largest number of papers is from 1992.

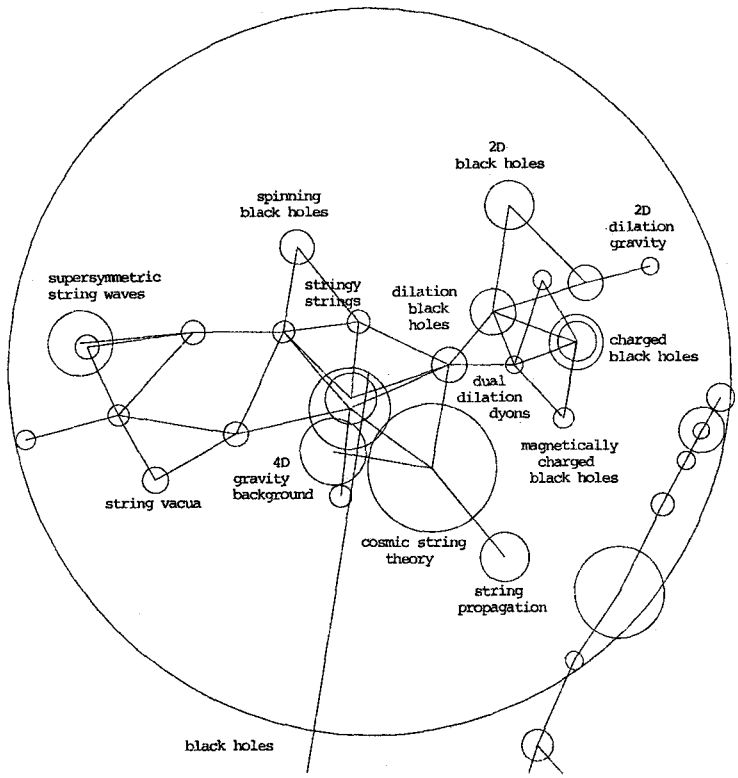

Fig. 5. A detailed view of the second-level cluster on black holes shown in the upper left of Figure 4. Some of the 26 first-level clusters are labeled by topic

The series of maps illustrates how the hierarchical structure of objects can be graphically depicted. The degree to which structures do not overlap is an indicator of the success of the ordination. Avoiding overlap could be enhanced by going to 3D space.

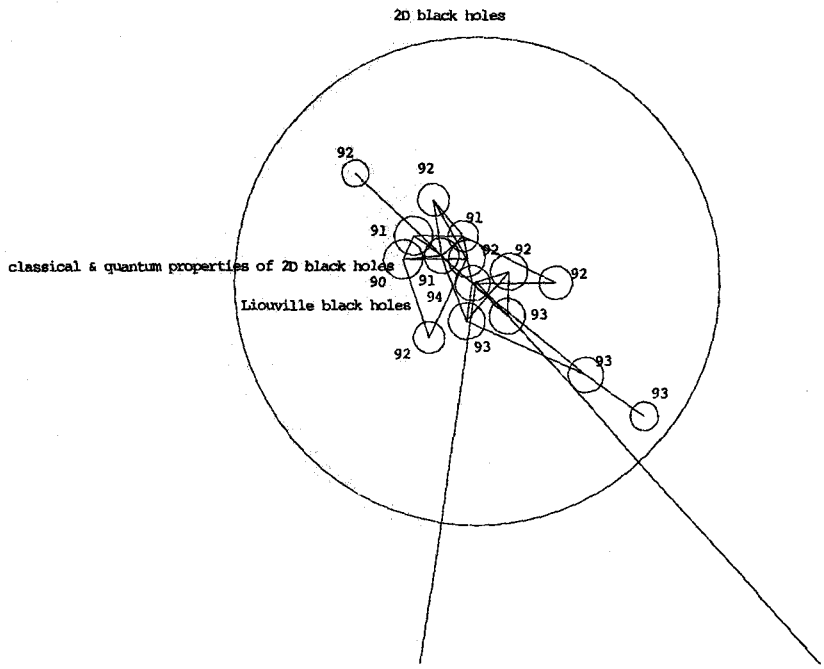

Fig. 6. A detailed view of the first-level cluster on "2D black holes" at the upper right of Figure 5 showing the 15 documents within it. The documents are labeled by publication year, and the titles of the oldest (1990) and the youngest (1994) documents within the dataset limits are shown

Discussion

No attempt at a general assessment or validation of this mapping has yet been attempted. The evaluation of such large scale maps needs to address both the reasonableness of macro- and micro-structures. In the absence of specialized expert knowledge of these fields, we can only rely on gross indicators of topic consistency. The main physics - chemistry axis is certainly plausible and consistent with earlier ISI mappings using multidimensional scaling of co-citations.[25]

At the micro-level we have delved into only one region, astrophysics, and found a high degree of consistency of topic. The kind of question which could be raised here is the location of string theory, which is applicable to fundamental physics generally and not specific to astrophysics. It turns out there is another cluster of string theory papers in the particle and high energy physics region and so string theory in this mapping is split between two locations.

Conclusions

The data presented illustrate a general framework for creating large scale maps of science, while presenting some new techniques for implementing individual steps. First we have focused on citation links as reliable indicators of intellectual connections within science. A new linkage measure was proposed which improves the "recall" of the mapping methodology over previous citation-based measures. However, other linkage mechanisms such as shared vocabulary or index terms could also be applied.

Second, we have chosen to first break up the large dataset by hierarchical clustering so that we can apply an ordination technique to each piece. An alternative would have been to scale up one of the ordination methods, but this would have imposed a much larger computational burden. Even though we have used a modified version of single-link clustering to break up the dataset, other clustering methods could have been applied which create less elongated clusters.

Third, we have chosen to apply a simple ordination procedure which we call triangulation, instead of one of the standard methods. Triangulation preserves short distances exactly, rather than attempting to fit all distances. Only further testing will determine whether the resulting configurations can approximate standard techniques and whether they are acceptable to users. If triangulation proves inadequate, we can revert to methods such as multidimensional scaling.

The procedure for reassembling the pieces into the hierarchical structure, involves successive expansions and translations of the coordinate systems of each cluster. One of the issues is how much to expand each cluster to avoid overlap. Our approach was to search for the largest overlap of adjacent objects. Also for orienting the clusters to one another by rotation, we have used a center of gravity approach again following a minimal spanning tree. These admittedly pragmatic solutions gave reasonable results and ran quickly, but they are not optimal.

To visualize the hierarchical organization, we have opted for what we call the volvox representation in two dimensions which is easily implemented with circles and lines. This can be extended to three dimensions by substituting spheres for circles, provided of course a three dimensional ordination is available. An advantage of hierarchical organization is that users can elect to view only large scale structures, thereby suppressing unwanted detail and speeding up the display.

This paper has not dealt with user interface issues, but clearly these will be critical to the successful utilization of large scale document ordinations. Other unresolved technical issues are how the structures will be updated and their stability over time.

*

Support from Sandia National Laboratory contract #AR–8321 is gratefully acknowledged.

References

1. H. SMALL, E. GARFIELD, The geography of science: Disciplinary and national mappings, *Journal of Information Science*, 11 (1985) 147–159.
2. H. SMALL, B. C. GRIFFITH, The structure of scientific literature I: Identifying and graphing specialties, *Science Studies*, 4 (January 1974) 17–40.
3. H. SMALL, Navigating the citation network, *Proc. 58th Annual Meeting Amer. Soc. Infor. Sci.*, 32 (1995) 118–126.
4. D. R. SWANSON, Two medical literatures that are logically but not bibliographically connected, *J. Am. Soc. Info. Sci.*, 38(4) (1987) 228–233.
5. P. HIGGS, Labs' 'virtual reality' shell get down to business, *Sandia Lab News* (Dec. 1, 1995).
6. P. H. SNEATH, R. R. SOKAL, *Numerical Taxonomy*, San Francisco: W. H. Freeman and Co., 1973, p. 245.
7. H. SMALL, E. SWEENEY, E. GREENLEE, Clustering the Science Citation Index using co-citations. II. Mapping science, *Scientometrics*, 8 (1985) 321–340.
8. B. C. GRIFFITH, H. SMALL, The structure of scientific literatures II: The macro and micro-structure of science, *Science Studies*, 4 (October 1974) 339–365.
9. M. CALLON, J. LAW, A. RIP, Qualitative scientometrics. *Mapping the Dynamics of Science and Technology*, M. CALLON, J. LAW, A. RIP (Eds), London: The MacMillan Press Ltd., 1986, pp. 103–123.
10. J. A. WISE, J. J. THOMAS, K. PENNOCK, D. LANTRIP, M. POTTIER, A. SCHUR, V. CROW, Visualizing the non-visual: Spatial analysis and interaction with information from text documents, *Proc. of IEEE Symposium on Information Visualization '95*, N. GERSHON, S. G. EICK (Eds), Los Alamos: IEEE Computer Society Press, 1995, pp. 51–58.
11. P. HEALEY, H. ROTHMAN, P. K. HOCH, An experiment in science mapping for research planning, *Research Policy*, 15 (1986) 233–251.
12. M. ZITT, E. BASSECOULARD, Recall rates of co-citation techniques: Bibliometric constraints and improvement in micro studies, *Proceedings of the 5th Biennial Conference of the Int. Soc. for Scientometrics and Informetrics*, M. E. D. KOENIG, A. BOOKSTEIN (Eds), Medford, NJ: Learned Information Inc., 1995, pp. 659–668.
13. R. R. BRAAM, H. F. MOED, A. F. J. VAN RAAN, Mapping of science by combined co-citation and word analysis, I. Structural aspects, *J. Amer. Soc. Info. Sci.*, 42 (1991) 233–251.
14. M. M. KESSLER, Bibliographic coupling between scientific papers, *American Documentation*, 14 (1963) 10–25.
15. H. SMALL, *Proc. of the 58th ASIS Annual Meeting*, op. cit.
16. R. A. AMSLER, Applications of citation-based automatic classification, unpublished report, University of Texas, 1972.
17. P. DOREIAN, V. BATAGELJ, A. FERLIGOJ, Partitioning networks based on generalized concepts of equivalence, *Journal of Mathematical Sociology*, 19(1) (1994) 1–27.
18. H. SMALL, E. SWEENEY, E. GREENLEE, *Scientometrics*, op. cit.

19. A. JAIN, R. C. DUBES, *Algorithms for Clustering Data*, Englewood Cliffs, N.J.: Prentice Hall, 1988, p. 41.
20. R. C. T. LEE, J. R. SLAGLE, H. BLUM, A triangulation method for the sequential mapping of points for N-space to two-space, *IEEE Transactions on Computers*, C26 (1977) 288–292.
21. H. SMALL, A. SCI-map case study: Building a map of AIDS research, *Scientometrics*, 30 (1994) 229–241.
22. H. SMALL, H. ROTHMAN, Investigations into the structure of science and social science using the SCI-map system, in: *Identifying Innovation in Social Science: Some Bibliometric Approaches*, SPSG Review Paper No. 8, Oct. 1994.
23. R. BURGIN, The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity, *J. Am. Soc. Info. Sci.*, 46(8) (1995) 562–572.
24. K. MCCAIN, personal communication.
25. H. SMALL, E. GARFIELD, *Journal of Information Science*, op. cit.