A SCI-MAP CASE STUDY: BUILDING A MAP OF AIDS RESEARCH*

H. SMALL

Institute for Scientific Information, 3501 Market Street, University City Center, Philadelphia, PA, 19104 (USA)

(Received January 18, 1994)

SCI-Map is a new PC based system for mapping the scientific literature. By selecting a seed item, the user can build a network or cluster of nodes interactively, and can view the structure as it is being built. New nodes are selected for addition to the network by the strength of their links to the items already clustered, and the positions of new nodes are determined by a geometric triangulation method. SCI-Map can be used to perform cluster-based retrieval using co-citation or other measures of document association, and enables the user to explore the structure of large document sets. This case study focuses on the AIDS literature and shows how the network is built up topic by topic, the recall of the final cluster, and where AIDS connects to the literature of other fields.

Introduction to SCI-Map

SCI-Map is a PC program running under Microsoft Windows which takes input data on nodes and links, and creates clusters node by node on screen, showing a display of the network as it is being constructed.[1] Nodes can represent any kind of object among which connections can be defined, and links represent the connections between nodes. Links must have numerical similarity values ranging between zero and one, and inter-node distances are computed by a logarithmic transformation of the similarity values. SCI-Map differs from Korfhage's document visualization systems which require the user to specify one or more reference points or dimensions on which the document map is to be constructed.[2,3] In contrast, SCI-Map creates a map of documents based on a geometric triangulation of the strongest inter-document links, in an attempt to represent the intrinsic associations among items.

In the case study presented here, nodes are documents covered by the ISI database which have been cited 10 or more times in the period 1989 through 1991, and links are their normalized co-citation strengths. The data set consists of 23,689

*Paper presented at the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics in Berlin (Germany), September 11-15, 1993.

Elsevier, Amsterdam – Oxford – New York – Tokyo
Akadémiai Kiadó, Budapest

documents connected by 80,871 distinct links (down to the threshold of 0.2 of normalized co-citation). Linkage normalization was performed using the *Salton*, or cosine, formula.[4]

In SCI-Map the user can select among several alternative clustering algorithms, and set a number of linkage parameters including citation and co-citation thresholds.[5] Typically, the user selects a starting or "seed" node and then adds nodes one by one until no further nodes are available in the database that connect to nodes in the cluster. Nodes are displayed on the screen with their links as they are added, and the position of each new node is determined by its two strongest links using a geometrical method similar to the surveyor's triangulation method. If only one link is available or its use would violate planar geometry, only the single strongest link is used. In both the two and one link cases, the geometrical solution is selected which maximizes the new node's distance from the center of the configuration (centroid). Positioning nodes by triangulating their strongest links is an approximate mapping method, because it does not take into account the weaker inter-item associations. Its relation to more commonly used methods of ordination such as multidimensional scaling and correspondence analysis remains to be investigated.[6,7]

Three types of rules govern the selection and order with which nodes are added to the cluster. First, the node and its links must meet the threshold requirements set by the user on citation and co-citation strength (qualifying rule). Second, nodes are added in the order of the sum of their linkage strengths to nodes already in the cluster (ordering rule). For example, the second node added will always be the one most strongly linked to the seed node. A third rule, the inclusion rule, comes into play if an algorithm other than single-linkage is used, for example, an average linkage strength threshold, if the average link algorithm is selected.

In addition, SCI-Map incorporates a number of analytical features useful for constructing and exploring maps, two of which will be extensively used in this study. The first feature enables the user to "cut" a node from the map, thereby preventing it from linking to any new nodes that may be added to the cluster. Another feature called the "collapse" feature allows the user to collapse a set of nodes into a single node, preserving any links the individual nodes may have had to other nodes prior to being collapsed.

Both the geometry of the cluster formed and the order in which nodes are added are sensitive to the starting or seed node. However, in the special case of the single-link clustering algorithm, the final make-up or composition of the cluster is independent of the starting point, since all linked nodes will eventually attach to the cluster as it forms. This is so even if nodes are "cut" along the way, as long as the same nodes are cut each time the cluster is generated using a different seed. One of the questions for this analysis is how much of the structure is preserved as the seed document is varied?

The experiment

The plan of the experiment is to select a topic defined by a set of key title words, pick an arbitrary starting node (document) on the topic, and generate the cluster around it by adding nodes one at a time. In addition, any node not having the appropriate key title words will be cut as soon as it is added. After the cluster is completed (i.e. when the documents in the data set that can be added are exhausted) the "recall" of the cluster will be computed, i.e., the proportion of the total population of documents on the topic which were retrieved by the cluster (# documents in the cluster minus cut-nodes divided by # on topic in the file).

Beyond this retrieval issue, I am also interested in how the structure of the nodes and the order in which they were added reflect the sub-areas within the topic and the nature of the cut-nodes themselves. If cut-nodes represent topics which are related to the starting topic but do not directly bear on it, then we can pose Don Swanson's[8,9] question: Can knowledge from these linked areas be brought to bear directly on the primary target area? Are these areas from which new discoveries may emerge?

The topic I have selected is AIDS research. The consensus seems to be that we are a long way from understanding this affliction let alone finding a cure. Since the disease has resisted a direct assault, perhaps what is needed is a flanking movement. One of my goals is to locate trails that lead from AIDS to other research areas, in the hope that information from these cognate fields can be brought to bear on the problem. As an operational definition of AIDS I used as title word search terms the acronyms AIDS, HIV, the full versions of these acronyms, and animal analogs of the disease such as SIV (simian immunodeficiency virus).

Results

The first document found in the SCI-Map data set having one or more of the AIDS terms (internally documents are ordered alphabetically by first author) was used as the "seed" for the cluster. This document had the title word "AIDS", and was a 1989 paper in the *Annals of Internal Medicine* having 23 citations. The paper was on the use of oral dextran sulfate in treating AIDS. The next paper added to the cluster was also on dextran sulfate , but because its title did not contain any of the AIDS terms, the second node was cut. However, the next five documents had the AIDS terms in their titles, and were also concerned with sulfate drug treatments. This pattern of papers adding to the cluster in topical groups was repeated many times as the cluster was built up.

The next topic to add was treatment with the drug DDI, which in turn was followed by a set of AZT papers, to which DDI is chemically related. AZT added about a dozen documents, when another topic began to appear, namely other reverse-transcriptase inhibitors. The phrase "anti-retroviral therapy" fits this region well. After 26 documents had been added, only two nodes had been cut.

Novel anti-HIV agents then appeared, as well as tests of combined agents. The use of interferon-alpha and beta-interferon then linked. After 46 nodes had been added, there were only four cut-nodes. Figure 1 shows the cluster at this stage in its formation. The topic specific regions are labeled and the order of their entry is given by a number in parentheses below the label.

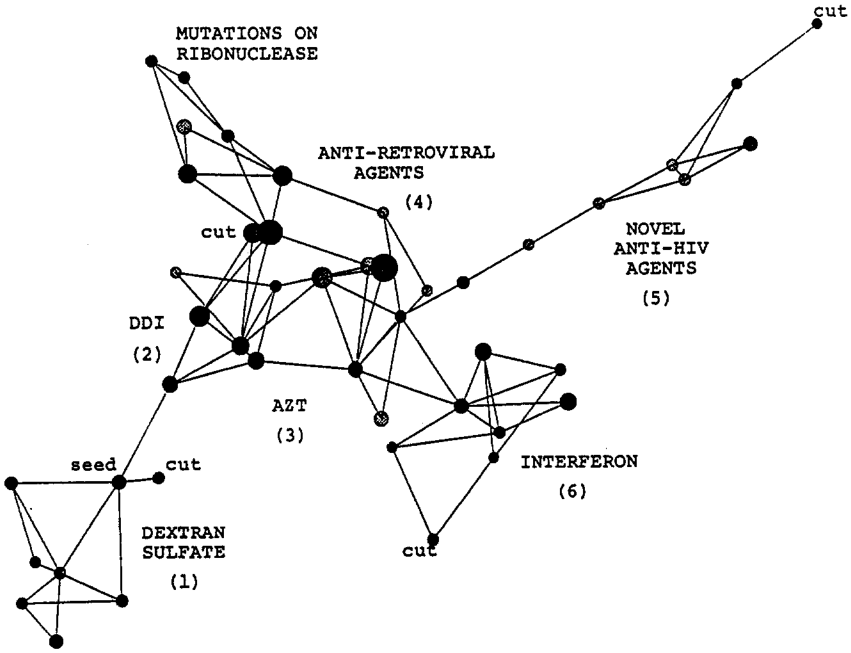

Fig. 1. AIDS/HIV cluster (46 nodes)

Starting with document 45 a new research line began to emerge having to do with mutations and genetic insertions on the ribonuclease that controls the reverse transcription process for the HIV virus replication. Then the focus shifted to the reverse transcriptase itself and its structure. This is the basic molecular biology behind the search for drugs which can inhibit HIV replication.

The topic of HIV proteases then emerged. Documents were added on the "rational design" of HIV protease inhibitors. Proteases added many more documents than the previous anti-HIV drug agent region. The cluster now contains 111 documents. The new regions add on to the top of the cluster, attached to the anti-HIV agents region below.

The next major topic to add was the assembly of HIV like particles. Certain mutations in RNA caused changes in the "packaging" of the virus particles which resulted in a non-infectious form of HIV. The proteins which controlled this packaging, such as vpu or vpr, were studied with the hope of finding the mutations which would make HIV non-infectious. The virus "packaging" papers added to the upper end of the protease structure.

Protein gene products of HIV were next to appear. These were studied for their role in regulation of transcription. Nef protein was considered a transcription inhibitor. Other proteins regarded as possibly useful for inhibition of HIV transcription were tat, rev,and tev. Some proteins interfered with the messenger RNAs. The rev protein, sometimes called the rev transactivator, was considered a potential regulator of HIV gene expression. Its structure and the exact site of its interaction with RNA was studied. Papers on these various proteins added in succession to the upper left end of the cluster.

The next major topic to emerge was the idea of transactivation regions (tars) which can enhance RNAs. This marked a shift from what might inhibit HIV transcription to what might enhance it. The tat protein was supposed to bind to a tar. Tat was considered a transactivator. Perhaps it was hoped that mutations in the transactivator region could stop HIV replication. At this point the cluster contained 203 nodes, and is shown in Fig. 2.

The next topic to emerge was simian AIDS and more generally animal models of the disease. These documents attached to the gene regulation region rather than the transactivation region. Branching off to the right was a small region on the development of vaccines for animal AIDS-like diseases, and immunization tests for them. The cluster now contained 273 documents.

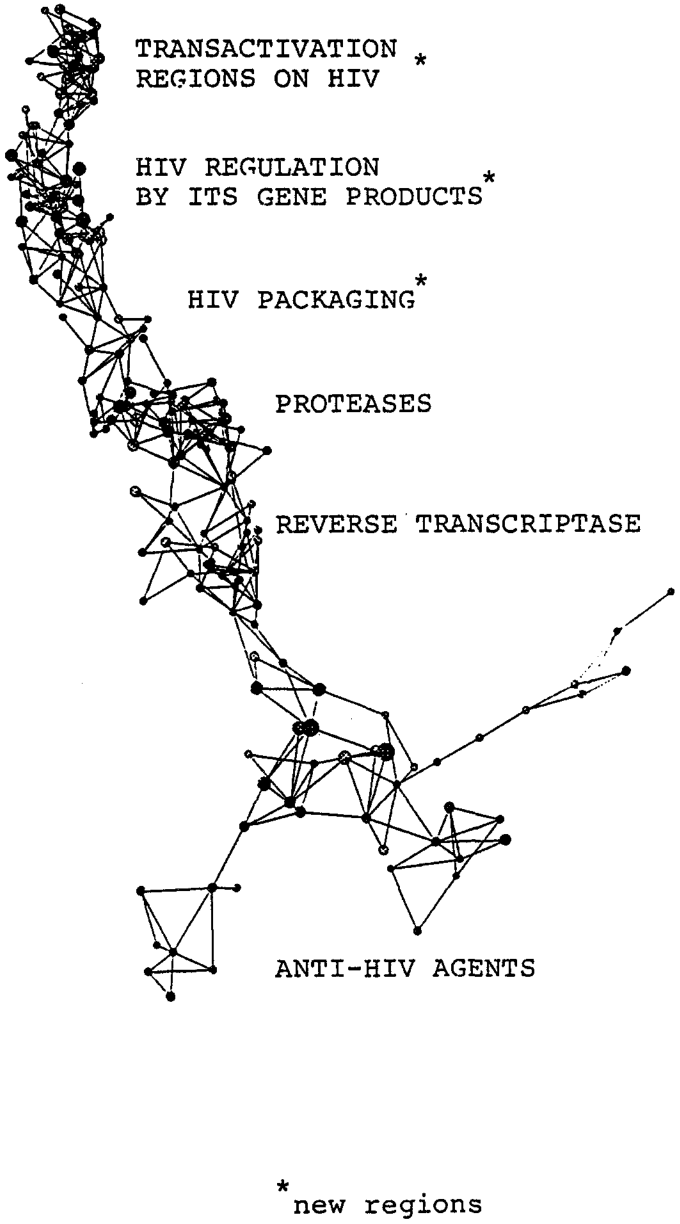

Fig. 2. AIDS/HIV cluster (203 nodes)

The previous topic led to the first immunological topic encountered so far in the cluster. The envelope glycoprotein (gp120) was found to play a role in eliciting antibodies that neutralize the HIV virus. After the addition of a cytotoxic lymphocyte region, the cluster contained 315 items.

Two branches grew out from the gp120 region: on the left a branch concerned with the structure and sequencing of the HIV envelope glycoprotein, and to the right a region on the binding of the CD4 receptor on lymphocytes to the gp120 envelope protein on HIV. It was hoped that by cloning the CD4 receptor and introducing it in the blood stream, it would be possible to bind all the HIV viruses. Therapies were designed and tested on the use of cloned soluble CD4. At this point in the cluster formation (429 documents) a CNS infection region began to emerge at the top of the cluster, and a perinatal transmission (mother to child) region formed just above the gp120 epitopes region.

As the cluster build-up process neared completion, few larger regions added in regular fashion to the top of the figure, and more individual documents added in a more scattered fashion over the entire structure. For example, from 429 to 494 documents, only two new topical regions were added: one having to do with the activation of HIV by other viruses, mainly human herpes virus and human cytomegalo virus, which attached midway up the figure and to the right of the existing HIV regulation region; and the other having to do with the induction of HIV expression by tumor necrosis factor (TNF), which attached at the bottom of the cluster.

At 630 documents no new major areas had added but several mini-regions emerged. These included (starting at the bottom of the figure and moving clockwise around it): micoplasma infection and AIDS; human mannose binding protein; transfusion AIDS; chemical markers during course of infection with HIV; HIV in the central nervous system; antisense oligodeoxyribonucleotides inhibition of HIV; and murine AIDS (MAIDS).

The final step in building the cluster was from 630 to 757 documents, when all linking documents in the data set were exhausted. The majority of documents added in this final phase were cut-nodes and only two new minor themes emerged: double stranded protein kinase (attached to transactivation), and macrophage-HIV interaction, which grew in a long tail from the simian AIDS region. The full cluster of 757 nodes is shown in Fig. 3.

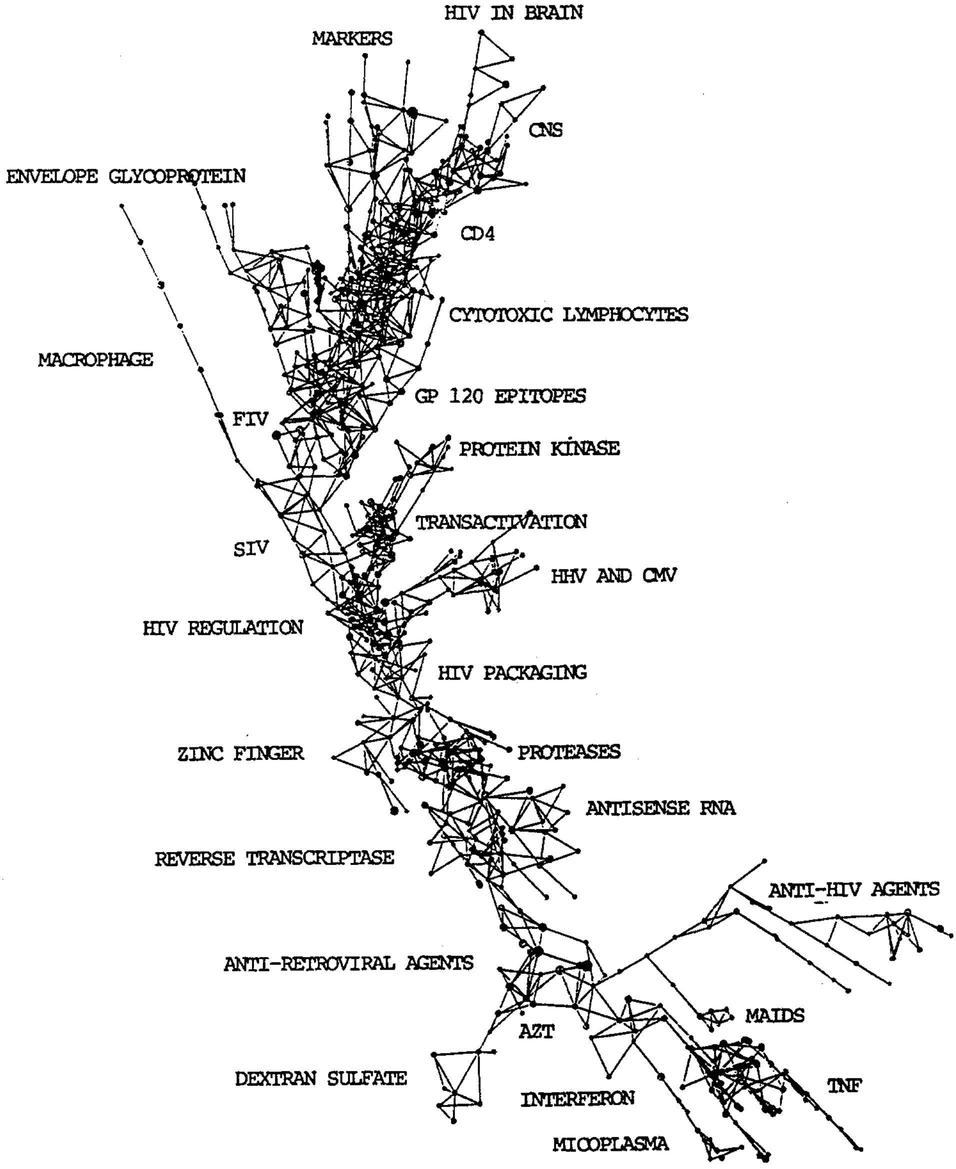

Fig. 3. AIDS/HIV cluster (757 nodes)

Overall structure of the cluster

The final map shows numerous dense regions and pockets of activity, connected together in almost a linear manner with relatively few side branches. There are a few regions of lower density which appear as appendages to the main body. The starting node proves to be part of a region of lower density compared to some of the biochemical and immunological regions above it. Also evident are a number of tails or loose ends which point in a direction away from the centroid, as we would expect. The loose ends are mostly late entering nodes which are weakly linked to the main body. However, it is possible that some of these nodes would have triangulated on their second strongest links and thereby be doubly linked, if the geometry had allowed it. The number of such non-triangulating nodes remains a topic for further research.

The most marked feature of the overall map is the consistency of topic focus within the various regions. Generally, all the papers on a topic added in sequence and formed regions of high density. A summary of the overall structure of the cluster was generated by use of the "collapse" feature of SCI-Map. Documents in the same region were collapsed and the collapsed regions were labeled as shown in Fig. 4.

Recall of the AIDS cluster

SCI-Map enables us to test how well the clustering strategy was able to bring together documents on the topic of AIDS by comparing it to the document set retrieved by searching on the same title words. It has been asserted that key word and citation searches often retrieve only slightly overlapping document sets.[10] Hence, our preliminary hypothesis is that documents retrieved by a key word search would not likely be retrieved by clustering using a citation linkage measure such as co-citation. A perfect cluster would include all items retrieved by the title word search.

To determine the recall of the cluster a count was made of all documents in the dataset having the various AIDS terms in their titles. A total of 783 were found, 3.3% of the 23,689 documents in the file. Since the cluster retrieved 757 documents of which 151 were cut points and did not contain the AIDS terms, this gives 606 AIDS documents retrieved by the cluster, for a recall of 77.3% (606/783). This high recall for a single cluster contradicts the usual expectation of low recall associated with citation based retrieval.

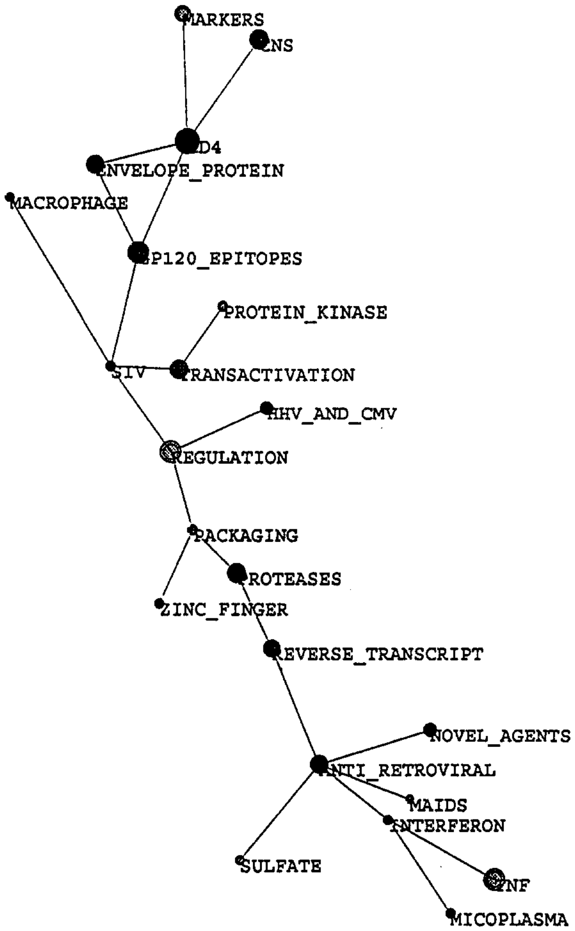

Fig. 4. AIDS/HIV cluster (overall structure)

Why did the remaining 177 documents not link? It is likely that if these remaining documents were used as seeds we would find additional clusters, smaller in size, devoted to other aspects of AIDS, such as public health and other social or medical issues. The large grouping we have retrieved dealt mainly with basic research issues in immunology, molecular biology, virology and pharmacology. Only further analysis of these remaining AIDS items would show what clusters might form.

Examination of cut-nodes

Of the 757 documents in the AIDS clusters at closure, 151 or about 20% of these were cut-nodes, that is, did not have any of the AIDS terms in their titles. Cut-nodes tend to cluster at certain points on the map. For example, a total of 17 documents were cut at a single point of attachment, to a paper on the binding of peptides to the major histocompatibility complex (MHC). The clustering of cut-nodes suggest that further structural regularities exist at a disciplinary or field level which could only be explored by a higher level mapping exercise.[11,12]

As expected, the number of cut-nodes goes up rapidly as the cluster approaches completion. Table 1 gives the number of cut-nodes for each 50 nodes added to the cluster. The table suggests that as the cluster is built up the supply of AIDS documents becomes exhausted. This implies that there is a tendency for documents having AIDS terms in their titles to be added into the cluster earlier than documents not having such terms because the former items are more strongly linked. As the AIDS nodes are added, only documents without AIDS terms remain, and these are generally less strongly linked. This is another way of saying that AIDS constitutes a coherent research field in which documents are strongly connected by citation patterns. Some of the more prominent themes among the 151 cut-nodes are given in Table 2.

Table 1
Cut nodes per 50 documents added to cluster

Document number	# cut nodes
1-50	5
51-100	8
101-150	4
151-200	0
201-250	11
251-300	1
301-350	5
351-400	0
401-450	7
451-500	5
501-550	10
551-600	17
601-650	18
651-700	30
701-750	27
751-757	3

Table 2
The clustering of cut-nodes:

cut node topic	# nodes cut
binding of peptides to HLA or MHC	17
relation of HIV to HHV and CMV	12
tumor necrosis factor and HIV	10
retroviral proteases	7
NF-kappa-B	7
RNAse-H	5
HTLV-I	5
myristoylation	4
cytotoxic T-cells	4
interleukin-6	4
multiple sclerosis	4
DDI and AZT	4

The importance of the starting point

It is assumed that the order of cluster build up and cluster geometry can be greatly affected by the choice of starting node. A prior, though less systematic, attempt to create an AIDS cluster using the same data file started from a different seed node, and gave both a different geometry and a different ordering of topics. Most of the large blocks of papers on specific topics remained intact, but their sequence of addition to the cluster and their points of attachment to one another changed. This suggests that strong subject groupings are preserved in the structure independent of starting point, although the juxtaposition of these groups can be variable. How robust the overall structures are with respect to varying the seed node will require further experimentation.

Conclusion

The question remains whether a breakthrough in the fight against AIDS will come from within the field itself or from some seemingly unrelated field? Should we attempt to explore the territory on the fringes of AIDS research on the assumption that innovations will come from outside the field? This paper has not attempted to address this question, although it remains a motivating force behind this research. This type of exploration of information trails requires a highly interactive and flexible interface which SCI-Map offers. In my view, the sign posts to follow are the cut-points which lead to research areas outside AIDS proper. How to assess their potential for new insights into AIDS is still an open question.

Many methodological issues also remain to be explored, such as the stability of the macro-structures formed by SCI-Map and how well the triangulation method approximates more mathematically based mapping methods. The ultimate goal of the SCI-Map system is to provide the user with a highly flexible system for discovering, exploring and displaying information structures. While offering a new approach to cluster-based information retrieval, SCI-Map also provides an interactive tool for understanding the natural structure of information and the organization of scientific knowledge.

*

The author wishes to acknowledge the creative collaboration of Susan *Ramee*, formerly of ISI, who developed the SCI-Map software under Microsoft Windows and now does Windows programming in the state of Washington, USA.

References

1 H. SMALL, H. ROTHMAN, Investigations into the Structure of Science and Social Science using the SCI-Map System, Final Report on ESRC contract, 1993.

2 K. A. OLSEN, R. R. KORFHAGE, K. M. SOCHATS, M.B. SPRING, J. G. WILLIAMS, Visualization of a document collection with implicit and explicit links, *Scandinavian Journal of Information Systems*, 5, (1993) 79-95.

3 K. A. OLSEN, R. R. KORFHAGE, K. M. SOCHATS, M. B. SPRING, J. G. WILLIAMS, Visualization of a document collection: The vibe system, *Information Processing and Management*, 29(1), (1993) 69-81.

4 G. SALTON, D. BERGMARK, A citation study of computer science literature, *IEEE Transactions on Professional Communication*, PC-22, (1979) 146-158.

5 H. SMALL, S. RAMEE, SCI-Map Operating Instructions, Institute for Scientific Information, July, 1992.

6 J. B. KRUSKAL, Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis, *Psychometrica*, 29, (1964) 1-27.

7 M. J. GREENACRE, *Theory and Applications of Correspondence Analysis*, New York, John Wiley, 1984.

8 D. R. SWANSON, Two medical literatures that are logically but not bibliographically connected, *Journal of the American Society for Information Science*, 38(4), (1987) 228-233.

9 H. SMALL, E. GARFIELD, letter to the editor, Verification of results that logically related noninteractive literatures are potential sources of new knowledge, *Journal of the American Society for Information Science*, 40(3), (1989) 152.

10 K. W. MCCAIN, Descriptor and citation retrieval in medical behavioral sciences literature: Retrieval overlaps and novelty distribution, *Journal of the American Society for Information Science*, 40, (1989) 110-114.

11 H. SMALL, Macro-level changes in the structure of co-citation clusters: 1983-1989, *Scientometrics*, 26, (1993) 5-20.

12 H. SMALL, E. GREENLEE, A Co-citation study of AIDS research, In: *Scholarly Communication and Bibliometrics*, C. BORGMAN, (Ed.) London, Sage Publications, 1990, pp. 166-193.