PERGAMON

Information Processing and Management 38 (2002) 491–508

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Mining a Web Citation Database for author co-citation analysis

Yulan He *, Siu Cheung Hui

School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, 639798 Singapore, Singapore

Received 14 December 2000; accepted 6 August 2001

Abstract

Author co-citation analysis (ACA) has been widely used in bibliometrics as an analytical method in analyzing the intellectual structure of science studies. It can be used to identify authors from the same or similar research fields. However, such analysis method relies heavily on statistical tools to perform the analysis and requires human interpretation. Web Citation Database is a data warehouse used for storing citation indices of Web publications. In this paper, we propose a mining process to automate the ACA based on the Web Citation Database. The mining process uses agglomerative hierarchical clustering (AHC) as the mining technique for author clustering and multidimensional scaling (MDS) for displaying author cluster maps. The clustering results and author cluster map have been incorporated into a citation-based retrieval system known as PubSearch to support author retrieval of Web publications.

Keywords: Author co-citation analysis; Data mining; Web Citation Database; Intelligent information retrieval

1. Introduction

In published journal articles, there are always some papers or books that are cited as references for the concepts or ideas presented in them. These cited references are used to refer the reader to the relevant papers for further reading on the concepts and ideas that are introduced in the source paper. They reveal how the source paper is linked to the prior relevant research on the assumption that citing and cited references have a strong link through semantics. They provide a valuable source of information and directives for researchers in the exchange of ideas, the current trends

* Corresponding author. Tel.: +65-790-4318, 4930; fax: +65-792-6559.
E-mail addresses: asylhe@ntu.edu.sg (Y. He), asschui@ntu.edu.sg (S. Cheung Hui).

0306-4573/02/$ - see front matter
PII: S0306-4573(01)00046-2

and the future development in their respective fields. Therefore, citation indices can be used to facilitate the searching and retrieval of information.

We have developed a citation-based retrieval system known as PubSearch (He, 2001), which generates a Web Citation Database from online scientific publications that are made available over the Internet, and supports the retrieval of the publications based on the Web Citation Database. The Web Citation Database is a data warehouse used for storing citation indices, which contain the references that the publications cite. The Web Citation Database can be generated using an autonomous citation indexing agent by searching through the Web sites on the Internet, which downloads the scientific publications, extracts the citations, generates the citation indices and stores the information in the Web Citation Database. This technique has also been demonstrated in another system known as CiteSeer (Bollacker, Lawrence, & Giles, 1998, 2000).

As most researchers are interested in scientific publications from certain research areas, identifying authors from the same research area becomes one of the most important knowledge for most researchers. To achieve this, author co-citation analysis (ACA) (White & Griffith, 1981) can be applied to identify inter-relationships between authors. It is an analytical method that has been traditionally used to trace the intellectual structure in science studies. ACA assumes that two authors are correlated if the frequency that they are cited together by later works is high. So if the frequency of two authors cited together by the same publication is very high, these two authors will belong to the same or similar research field.

The Web Citation Database contains rich information that can be mined to help document retrieval. In this paper, we focus on mining (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Mitchell, 1999) the Web Citation Database using the ACA for author clustering. Agglomerative hierarchical clustering (AHC) (Everitt, 1986) is used as the mining technique for author clustering and multidimensional scaling (MDS) (Green, Carmone, & Smith, 1989) is adopted for displaying author cluster maps. The mining process has been incorporated into the PubSearch system for author retrieval. The rest of the paper is organized as follows. First, a brief introduction of the ACA process is presented followed by the related work on the application of ACA method. Then, the structure of the Web Citation Database is described. The author co-citation mining process is discussed. The performance of the author clustering technique is evaluated. Finally, a conclusion is given.

2. Author co-citation analysis

ACA (White & Griffith, 1981) is carried out as follows. It first selects the targeted authors from a huge data store. Then, the number of co-occurrences of any two different authors is calculated as the author co-citation frequency, which will then be used to form a two-dimensional raw data matrix with identically ordered authors' names for rows and columns. The raw data matrix is converted into a correlation matrix of proximity values that indicate the degree of similarity between author pairs. Pearson's correlation coefficient (Johnson, 1988) is commonly used for this purpose. Finally, multivariate analysis is applied to the correlation matrix to reveal the inter-relationships among authors. Three approaches can be used for multivariate analysis (McCain, 1990): cluster analysis, MDS and factor analysis.

For cluster analysis, the two most popular techniques are AHC (Everitt, 1986) and iterative partitioning algorithms (Jain & Dubes, 1988). Both techniques involve a tree-like building process. In AHC algorithm, clusters are built from the bottom with individuals or groups of individuals gradually joining to form clusters, while in iterative partitioning algorithm, a collection of all individuals are split from top to bottom and the process iterates until the desired number of clusters is reached. The AHC algorithm is commonly used in ACA research work.

MDS (Green et al., 1989) is used to create visual displays or maps so that the underlying structure within a set of objects can be studied. Heavily co-cited authors appear close to each other in the multidimensional space. Authors with many links to others tend to be in central positions while weakly linked authors will be placed in the periphery. In this way, central and peripheral research specialization can easily be shown (Kruskal, 1977).

Factor analysis (Gorsuch, 1983) attempts to "explain" the inter-relationships observed among the original variables through the creation of a much smaller number of "derived" variables or factors. In ACA, a factor is interpreted by the subset of authors loading on it, i.e. making substantial contributions to its construction. Every author loads on (contributes to) every factor, and the interpretation or definition of each new factor is based on those authors with high loadings. The strength of inter-correlation among the factors reveals the subject-related linkage between authors. The advantage of factor analysis is its ability to demonstrate the breadth of contributions by authors who load substantially on more than one factor.

The computation complexity for factor analysis is $O(N^f)$, where N is the number of authors being analyzed and f is the number of factors finally extracted. The AHC method has the computation complexity $O(N^2)$, where N is the number of clusters generated. It is much faster than the factor analysis and the algorithm is simpler to implement. On the other hand, MDS can be used to create a visual display of points, which form the author cluster maps.

3. Related work

Traditionally, researchers in information studies field rely on some statistical tools, such as SPSS (2000) which supports cluster analysis, MDS and factor analysis, to perform ACA (Larson, 1996; Perry & Rice, 1998; White & McCain, 1998; White, 1990). To do this, they need to get the author co-citation raw matrix either manually or use other software tools before feeding it into the statistical tools. As the returned results are not well-clustered author groups, researchers still need to group authors manually according to the positions of the authors in the output display. So the whole process depends very much on human interpretation and interaction.

In recent years, information visualization techniques have been researched to display author co-citation information. Lin (1997, 2000) has applied the Kohonen's self-organizing map (KSOM) (Kohonen, 1995) algorithm to the co-citation matrices to generate a display map that clusters information scientists into several subject areas. Users can click on an author's name from the display map. The selected author name is then passed to the Alta Vista search engine, and Web pages containing the selected author are returned. Chen and Carr (Chen & Carr, 1999a,b; Chen, 1999; Chen, Chennawasin, & Yu, 2000) have used Latent Semantic Indexing (Deerwester, Dumais, Furnas, & Landauer, 1990) and Pathfinder Network Scaling (Schvaneveldt, Durso, & Dearholt, 1989) to extract the semantic structures and citation patterns from document

collections. Factor analysis supported by SPSS has been used to identify the major research areas from the author co-citation patterns. The author co-citation map can then be visualized in three-dimensional and color-codable form. Users can select the author name on the map to retrieve the corresponding documents under the author. The major research areas can also be displayed on the map. In (White, Buzydlowski, & Lin, 2000), it has also proposed to use Pathfinder Network Scaling to generate the author co-citation maps, which can then be served as interfaces for document retrieval. However, only design issues have been discussed in the paper.

As can be seen, most recent research works on ACA focus on investigating visualization techniques for displaying author co-citation information. However, before it can be visualized, the users still need to spend much time on collecting and converting the author data into the right format before feeding them into the statistical tools or visualization tools for display. This is tedious and time-consuming whenever a user wants to generate author co-citation information. In this research, we focus on automating the author co-citation process by mining the author information directly from the Web Citation Database. Such automated process can then be incorporated into a retrieval interface to generate different forms of author information.

4. Web Citation Database

Fig. 1 shows the relationships of the two major tables created in the Web Citation Database. They are the SOURCE and CITATION tables. The SOURCE table stores the information of source papers while the CITATION table stores all the citations extracted from the source papers. Most attributes of these two tables have the same data definitions such as the paper title, author names, journal name, journal volume, journal issue, pages and the year of publication. URL_link is the Web URL address of the corresponding document. With this field, full-text access is possible. "Paper_ID" of the SOURCE table and "Citation_ID" of the CITATION table are the primary keys in these two tables, respectively. "No_of_citation" of the SOURCE table is the number of references contained in the source paper. "Source_ID" of the CITATION table links to the "Paper_ID" of the SOURCE table to identify the source paper that cites the particular publication stored in the CITATION table. As can be seen, most fields in the CITATION table are similar to those in the SOURCE table. It should also be noted that for all the papers, only the first three authors are stored in the Web Citation Database. This is based on the assumption that the fourth author onwards contribute little to the paper.

An example of records stored in the SOURCE and CITATION tables is illustrated in Fig. 2. Records in the SOURCE and the CITATION tables have many-to-many relationships. That is, one source paper from the SOURCE table may cite multiple papers in the CITATION table. While one record in the CITATION table may be cited by more than one source paper in the SOURCE table. The example shows that both source papers with Paper_ID 1068 and 1124 cite the same paper entitled "A simple blueprint for automatic Boolean query processing" written by Salton. On the other hand, the source paper 1068 cites papers by Salton and by Harter at the same time.

For experimental purpose, we have set up a test Web Citation Database by downloading the publications from 1987 to 1997 in information retrieval (IR) field of Social Science Citation Index from the Institute for Scientific Information (ISI, 2000) Web site, which includes all the journals on Library and Information Science. A total of 1466 IR related papers were selected from 367

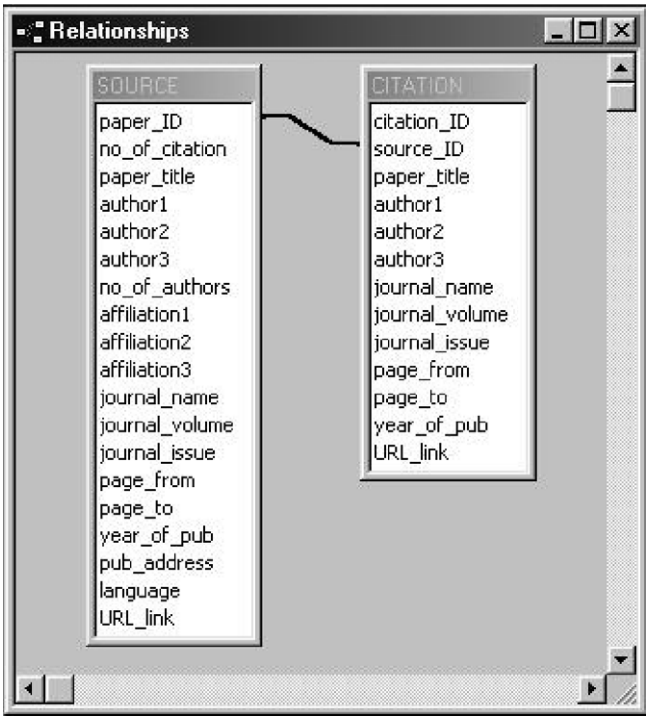

Fig. 1. Database structure of the Web Citation Database.

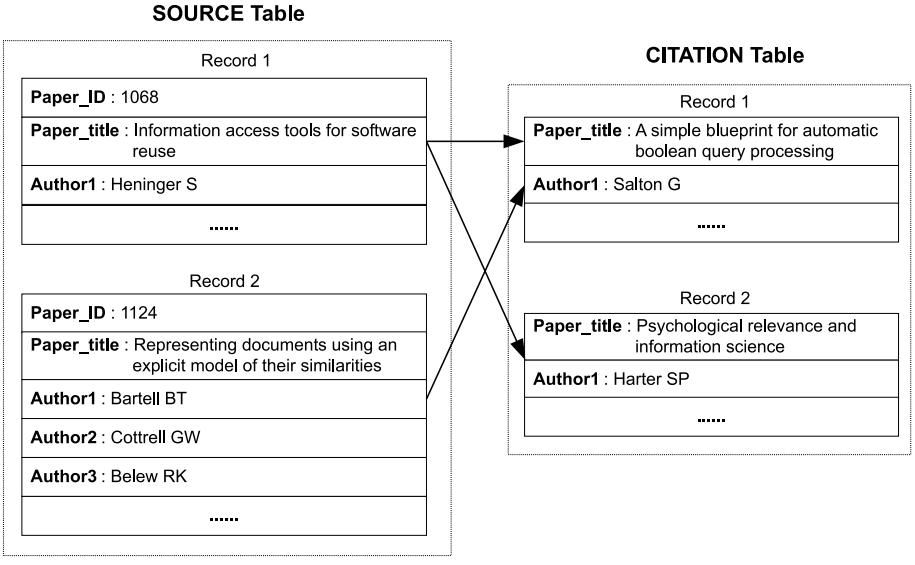

Fig. 2. Example of records stored in the SOURCE and CITATION tables.

journals with 44,836 citations. The two tables, SOURCE and CITATION, were created based on these IR papers.

5. Author co-citation mining process

Fig. 3 shows the data mining process for ACA. The author co-citation pairs are first created from the Web Citation Database. The co-citation frequency and the co-citation link strength of each author pair are calculated. The author pairs with the co-citation link strength below a certain threshold are filtered out and the rest will form the raw co-citation matrix. The raw co-citation matrix is then converted into the correlation matrix by substituting the author co-citation frequency with Pearson's correlation coefficient. The AHC algorithm is applied to the correlation matrix to generate the author clusters. MDS is modified to display author cluster maps. The author cluster information is incorporated into the PubSearch retrieval system for author retrieval based on user queries.

5.1. Create author co-citation pairs

The input to this step is the CITATION table in the Web Citation Database. One of the attributes of the CITATION table is "source_ID", which specifies the source paper of the current record. The records with the same "source_ID" are citations from the same source paper. Therefore, the author co-citation pairs can be created by grouping two distinct authors together with the same "source_ID". The count of co-citation of each author pair is calculated as the co-citation frequency.

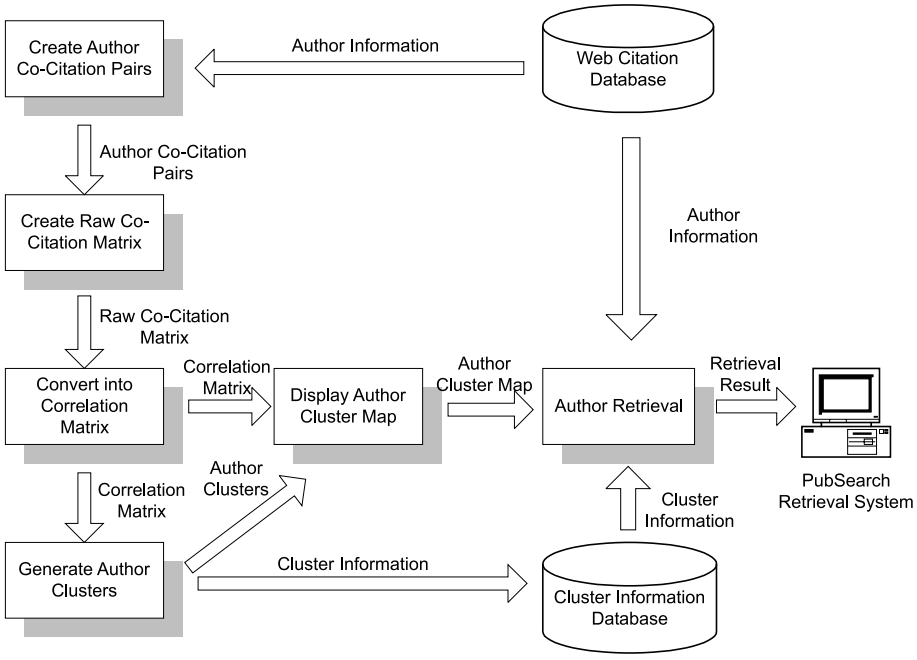

Fig. 3. Author co-citation mining process.

5.2. Create raw co-citation matrix

After the creation of author co-citation pairs, the co-citation link strength (Garfield, 1980) is calculated using the following formula:

$$\text{Link Strength (AB)} = X/(Y - X),$$

where X is the number of co-citations of author A and author B, Y is the sum of the total number of citations of A and the total number of citations of B. This formula normalizes the co-citation link strength by taking into the account of the total number of citations for both A and B.

A threshold is set to filter out the author pairs with insignificant co-occurrences. Author pairs with co-citation link strength exceeding the threshold value are retained. The raw co-citation matrix is then formed by taking the list of authors as the entries for both row and column. The value of each cell in the matrix is the co-citation frequency count of the authors in the corresponding row and column. Such matrix is symmetric, as the lower triangular-matrix is identical to the upper triangular-matrix.

For each diagonal cell of the matrix, White and Griffith (1981) select the value based on the highest off-diagonal co-citation counts for each author. Chen and Carr (1999a) put the mean co-citation counts for the same author in the diagonal cell. As the diagonal cells store the self-citation counts, it does not contribute to the author clustering process. Here, the diagonal cell values are treated as zeros.

The value of the link strength threshold affects the final results of the author clustering process as the greater the frequency of a given pair, the greater the likelihood that the two authors belong to the same research area. Therefore, the scope of the research area can be adjusted by increasing or decreasing the threshold. Generally, the higher the threshold, the narrower the scope. Here, we have set the link strength threshold to be 0.45 as determined experimentally.

5.3. Convert into correlation matrix

The correlation is defined as a measure of similarity. The higher the positive correlation, the more similar two authors are from the perceptions of citers. Pearson's correlation coefficient is used to measure the similarity between author pairs (White & McCain, 1998). The formula used to calculate Pearson's correlation coefficient r is given as follows (Johnson, 1988):

$$r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[N(\sum X^2) - (\sum X)^2\right]\left[N(\sum Y^2) - (\sum Y)^2\right]}},$$

where X and Y are two input vectors and N is the dimension of the input vector. For example, if a raw co-citation matrix of 40 authors is generated, for an author co-citation pair A and B, each element of input vector X will be the co-citation count of author A with all other authors, each element of input vector Y will be the co-citation count of author B with all other authors, and N will be 39 as we do not take the self-citation count into consideration. Pearson's correlation coefficient r measures the strength of linear correlation. It is always between -1 and $+1$ inclusive. -1 means perfect negative linear correlation and $+1$ means perfect positive linear correlation.

5.4. Generate author clusters

The correlation matrix stores the inter-author proximity values. It serves as the input to the AHC algorithm (Kurita, 1991) to derive the author clusters. Fig. 4 shows the AHC process. It is essential to determine the number of clusters for hierarchical clustering. The clustering method begins by considering each author to be a cluster. At each stage of the analysis, the algorithm combines two clusters until, at the end, all of the authors are in a single cluster. The similarity measurement threshold needs to be defined such that at certain point, the algorithm will quit from the cluster merge process and return the clusters at that stage.

As the hierarchical clustering approach used is agglomerative, it is obvious that the first two clusters merged are the ones that show the greatest similarity. However, as the number of authors per cluster increases by merging the existing clusters, the authors within each cluster are increasingly dissimilar as the process goes on. In this step, we also need to consider the retrieval

AHC_Algorithm:

Input: The N authors to be clustered, and the corresponding N × N correlation matrix.

Process:

1. Start by assigning each author to its own cluster. That is, N clusters will be created initially.
2. Find the most similar pair of clusters and merge them into a single cluster.
3. Compute the similarities between the new cluster and each of the old clusters. There are four methods to compute the similarities, namely single link, complete link, average link, and Ward's method.
   - *Single link*. The inter-cluster similarity equals to the greatest similarity from any members of one cluster to any members of the other clusters.
   - *Complete link*. The least similarity pair between two clusters defines the inter-cluster similarity.
   - *Average link*. The inter-cluster similarity is the average pairwise similarity between two clusters.
   - *Ward's method*. The loss of information that results from the grouping of individuals into clusters can be measured by the total sum of squared deviations of every point from the mean of the cluster to which it belongs. This method joins two clusters that result in the smallest increase in the overall sum of the squared intra-cluster distance.
4. Repeat steps 2 and 3 until the desired number of clusters is reached.

Output: The desired number of clusters.

Fig. 4. The AHC algorithm to generate author clusters.

efficiency, and each resulting cluster obtained should contain a reasonable number of authors. In our work, the similarity measurement threshold is determined experimentally to get the best performance as will be discussed in Section 6.

5.5. Display author cluster map

After the clusters are generated, it is necessary to present the author cluster information in a graphical manner to make it more effective and intuitive. To achieve this, we propose a method to generate the author cluster map. It uses the MDS algorithm (Green et al., 1989) to map authors in the correlation matrix into two-dimensional space with each point representing an author. The

Modified_MDS_Algorithm:

Input: An N × N correlation matrix (D).

Process:

1. Assign author points to coordinates in two-dimensional space arbitrarily.
2. Compare Euclidean distances among all pairs of points to form a matrix, which is called a Dhat matrix.
3. Compare the Dhat matrix with the original correlation matrix D by evaluating the *stress* function. The smaller the *stress* value, the greater the correspondence between them. The general form of the *stress* function is given as:

$$stress = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(f(x_{ij}) - d_{ij})^2}{scale}}$$

   where N is the original dimension of the correlation matrix; d_{ij} refers to the Euclidean distance between points i and j on the two-dimensional map; $f(x_{ij})$ is the transformation function of the original input data of the correlation matrix; in metric scaling, $f(x_{ij}) = x_{ij}$; and in non-metric scaling, $f(x_{ij})$ is a weak monotonic transformation of the input data that maximizes the *stress* function; $scale$ refers to a constant scaling factor, it is used to keep the *stress* value between 0 and 1.

   If the MDS map reproduces the input data perfectly, the *stress* is zero. Thus, the smaller the *stress*, the better the representation.
4. Adjust the coordinates of each point in the direction that can best minimize the *stress* function.
5. Repeat step 2 through step 4 until the *stress* will not get any lower.

Output: XY-coordinates for each author.

Fig. 5. The modified MDS algorithm to generate author cluster maps.

modified MDS algorithm is given in Fig. 5. The output of the modified MDS algorithm is the *XY*-coordinates of each author. Then, based on the author cluster information and *XY*-coordinates of all the authors, a two-dimensional map is generated as shown in Fig. 6. Authors from various clusters are differentiated using points with different shapes and colors. Each cluster represents a research area. Authors within the same cluster are the experts or researchers of the same research area.

5.6. Author retrieval

The author retrieval process retrieves publications information published by an author, using author cluster information and cluster map. The user's query input will be the author name, either full name or partial name. The system will then display the map of the author cluster that the input author belongs to. For example, if the user's query is ''Belkin'', Fig. 7 shows the map of the cluster that contains the author ''Belkin''. Each point represents an author. The distance between each other roughly corresponds to the similarity among them. By clicking any author names, the list of papers written by that author are displayed as shown in Fig. 8. All paper titles are underlined to indicate the availability of the URL links for full-text access of the publications. Users are also allowed to view the author cluster map by clicking the button ''Click here for overall author cluster map'' in Fig. 7. The author cluster map will then be displayed as shown in Fig. 6.

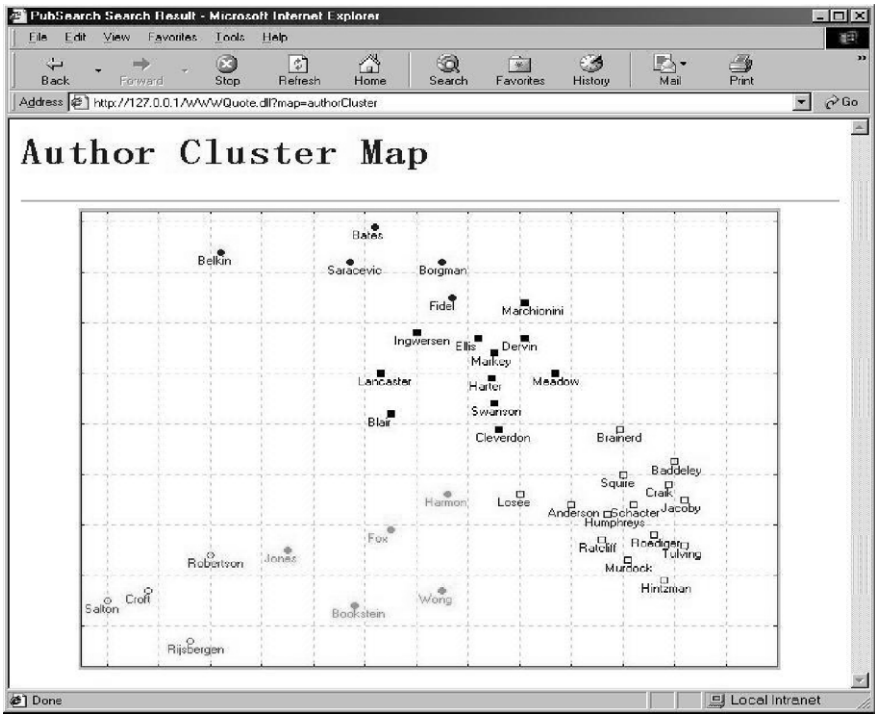

Fig. 6. Author cluster map.

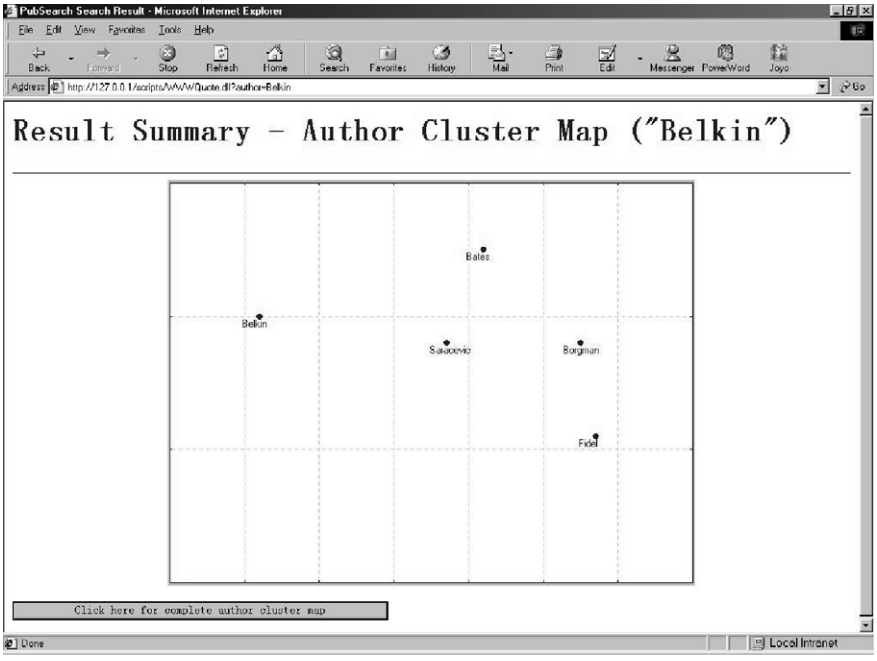

Fig. 7. Author cluster map for the search query on ''Belkin''.

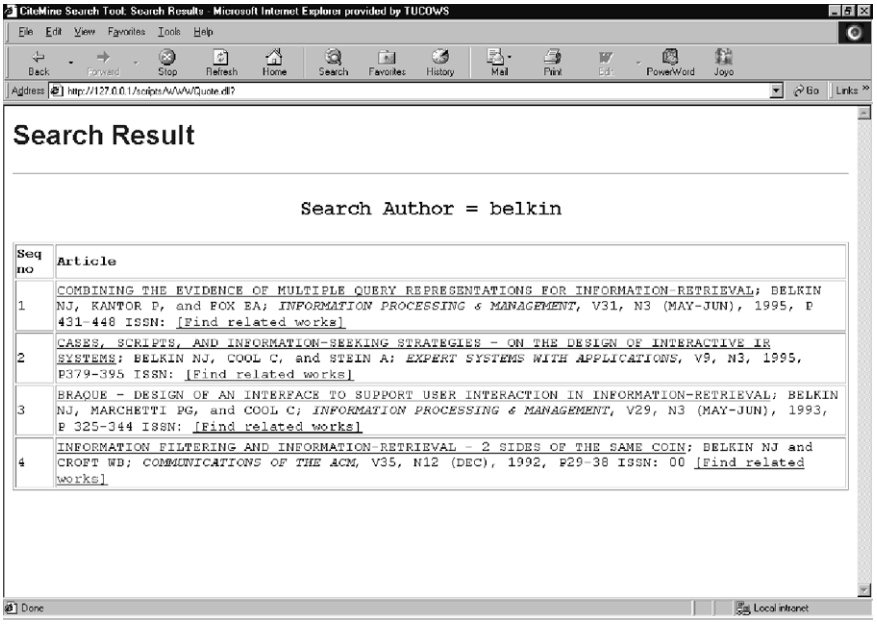

Fig. 8. List of papers by ''Belkin''.

Besides author IR, the author cluster maps can also be used in monitoring the change of the research focus of a particular author as well as the newly emerged research areas throughout the years. This can be shown as follows. The forty most highly cited authors from the Web Citation Database in IR field from 1987 to 1997 are used as author samples. Users are allowed to enter the range of the publication year such that author cluster maps are generated in different time frames. For example, the 10-year period can be divided into two time frames, 1987–1991 and 1992–1997. The records with the publication date falling within these two time frames are analyzed separately and two author cluster maps can then be formed as shown in Figs. 9 and 10, respectively.

For illustration purposes, the boundaries of every author cluster are manually drawn and the description for each research area is given based on the common research topics of every author within a cluster. Authors from different research areas, such as the general IR theory, IR techniques, IR model, user information seeking and retrieving behavior, are displayed. The area of computerized IR system and mathematical model during the period of 1987–1991 is more or less similar to the area of IR system design and evaluation during the period of 1992–1997. The newly emerged research fields include user perspectives of IR and IR theory research. Some authors' research interests have changed over these two periods. For example, Belkin shared the same research interest with Van Rijsbergen, Croft and Sparck Jone during the period of 1987–1991. But his research focus subsequently shifted to user information seeking and retrieving behavior.

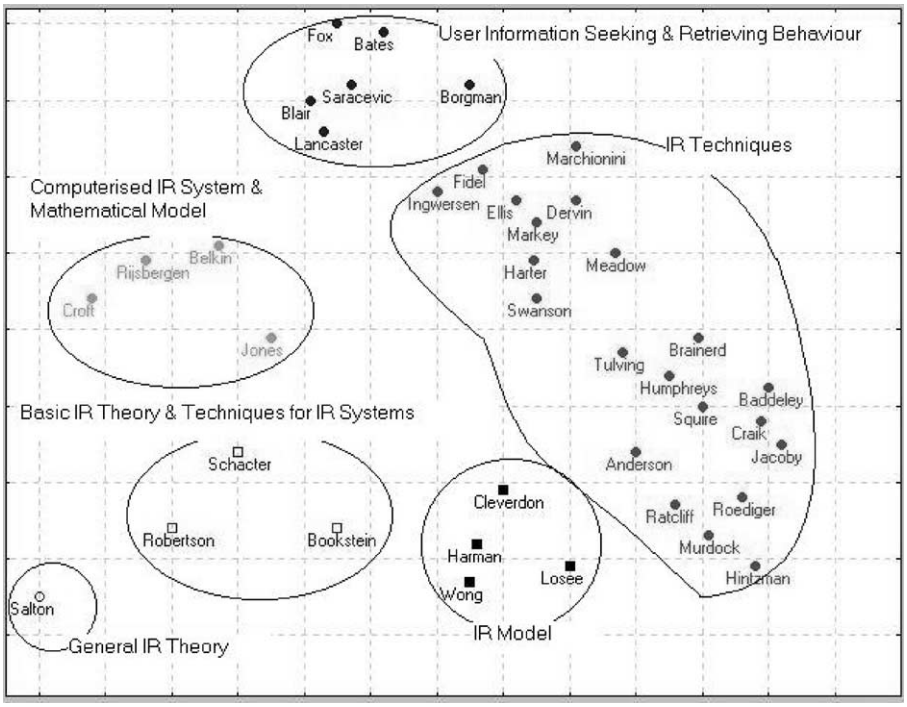

Fig. 9. Author cluster map (1987–1991).

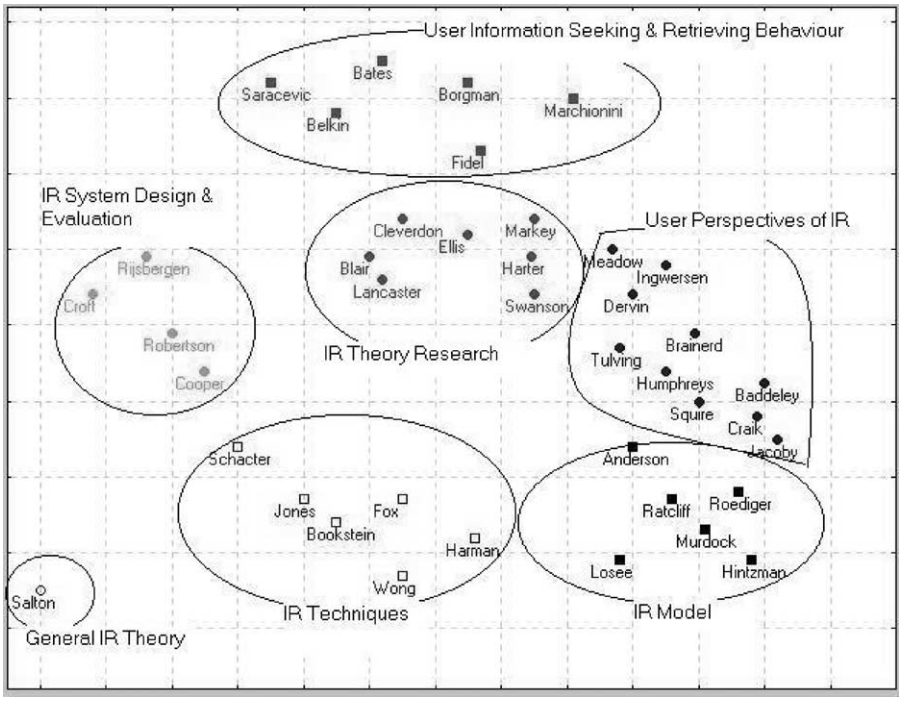

Fig. 10. Author cluster map (1992–1997).

6. Performance analysis

To evaluate the performance of the author clustering algorithm, the entropy measure that uses labels as defined in (Boley, 1998) is applied. Each author is manually given a label according to the research area the author belongs to. The labels are used to measure the entropy of the resulting cluster as a measure of quality. The entropy of a given cluster C is defined by the following formula:

$$e_C = -\sum_i \left(\frac{c(i,C)}{\sum_i c(i,C)} \right) \log \left(\frac{c(i,C)}{\sum_i c(i,C)} \right),$$

where $c(i,C)$ is the number of times label i occurs in cluster C. The entropy for a cluster is zero if the labels of all the authors are the same, that is, all authors are in the same research area. Otherwise, it is positive. The total entropy is the weighted average of the individual cluster's entropies:

$$e_{\text{total}} = \frac{1}{M} \sum_C (e_C \times N_C),$$

where M is the total number of clusters and N_C is the number of authors in cluster C. Therefore, the lower the entropy, the better the quality of the author clustering algorithm.

In the proposed author clustering process, two factors can affect the final clustering results. One is the co-citation link strength threshold. As recalled from the previous sections, only author pairs with the link strength greater than the pre-defined threshold are used to form the correlation matrix. Other author pairs are discarded from the measurement. Experiments need to be conducted in order to get the optimal co-citation link strength threshold. Another factor that can affect the cluster results is the method used to calculate the similarity between clusters. Four different methods are implemented to compare the performance. They are the single link, complete link, average link and Ward's method.

6.1. Experiment

The whole Web Citation Database consists of papers on IR studies published from 1987 to 1997. For illustration purpose, we have chosen 40 most highly cited authors as the input data to the experiment. These authors have been classified into six different research areas (Ding, 1998) using the SPSS tool with manual processing. Table 1 shows the results of the classification.

6.2. Co-citation link strength threshold

In this section, experiments were conducted to show how the co-citation link strength affects the final clustering result. Table 2 gives an illustration on how to calculate the entropy values using complete link method with co-citation link strength set to 0.3. The clusters would stop merging when the similarity value was below the similarity measurement threshold of 0.5. The entropy (e_C) for each cluster was also calculated.

By varying the co-citation link strength threshold, different clustering results are obtained. Fig. 11 shows a comparison of the entropy values obtained by varying the co-citation link strength threshold from 0.1 to 0.7 for the four similarity measurement methods, namely the single link, complete link, average link and Ward's method. In the figure, the similarity measurement threshold is set to 0.5. It can be seen that the co-citation link strength threshold ranging between

Table 1
The categorized author cluster results

Research area	Author names
General IR theory	Salton G
Computerised IR system & mathematical model	Croft WB, Jones KS, Robertson SE, van Rijsbergen CJ
IR model	Bookstein A, Cooper WS, Fox EA, Harman D, Losee RM, Wong SKM
Psychology	Anderson JR, Baddeley AD, Brainerd CJ, Craik FIM, Hintzman DL, Humphreys MS, Jacoby LL, Murdock BB, Ratcliff R, Roediger HL, Schacter DL, Squire LR, Tulving E
IR theory & techniques (less mathematical & computerized)	Blair DC, Cleverdon CW, Dervin B, Ellis D, Harter SP, Ingwersen P, Lancaster FW, Marchionini G, Markey K, Meadow CT, Swanson DR
User searching behavior	Bates MJ, Belkin NJ, Borgman CL, Fidel R, Saracevic T

Table 2
Statistics of entropy for each cluster with co-citation link strength = 0.3

Research area	Clusters			
	1	2	3	4
General IR theory	0	0	1	0
Computerised IR system & mathematical model	1	3	0	0
IR model	0	0	6	0
Psychology	9	1	0	3
IR theory & techniques (less mathematical & computerized)	0	1	1	9
User searching behavior	0	0	0	5
Entropy (e_C)	0.2611	0.2741	0.0947	0.2183
Total entropy (e_{total})	2.1126	2.1126	2.1126	2.1126

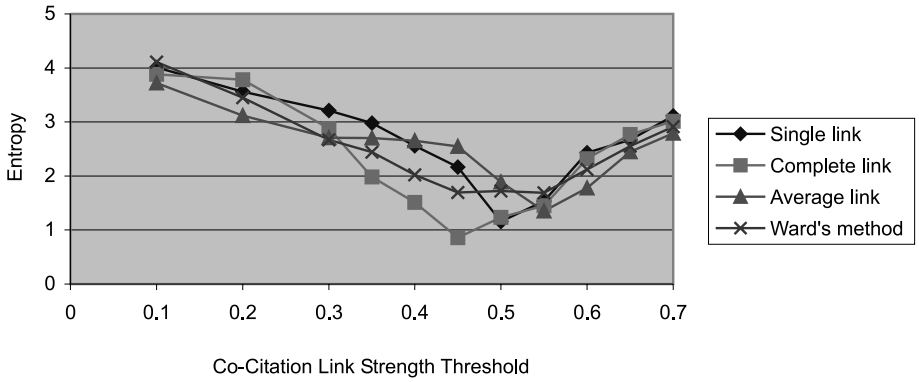

Fig. 11. Entropy values by varying the co-citation link strength thresholds.

0.4 and 0.55 gives better clustering results. Similar results have also been obtained by setting other values for the similarity measurement threshold.

6.3. Comparison of different similarity measure methods

For the AHC algorithm, the similarity measurement threshold is used to determine when to stop merging the existing clusters. Experiments were also conducted in order to find the optimum threshold. Fig. 12 shows the comparison of the four similarity measurement methods with the co-citation link strength threshold varying between 0.4 and 0.55. The best performance is obtained using the complete link method with entropy value of 0.8562 when the co-citation link strength threshold is set to 0.45 and the similarity measurement threshold is set to 0.5. The complete link method produces tightly bound or compact clusters (Baeza-Yates, 1992). The single link method, in contrast, suffers from a chaining effect (Everitt, 1986) and produces poor performance. It has a tendency to produce clusters that are straggly or elongated. The average link method is very efficient when the objects form natural distinct "clumps" and it performs equally well with elongated, "chain" type clusters. Ward's method is different from the above three methods because it

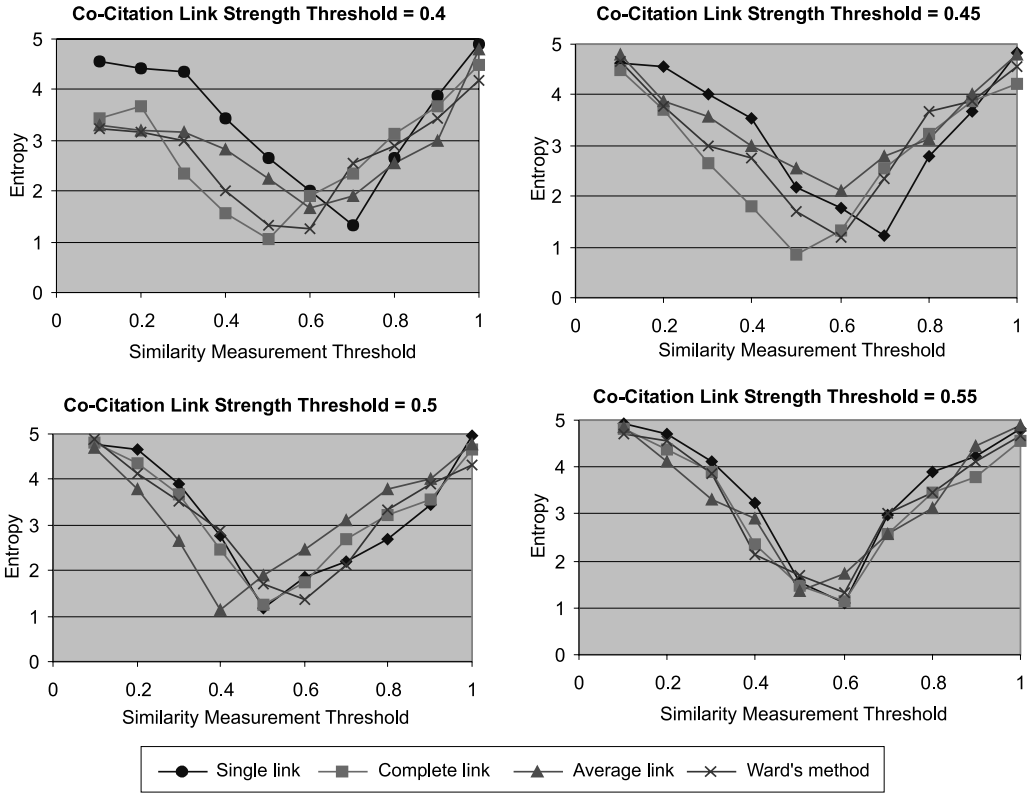

Fig. 12. Performance measure of entropy using different clustering methods.

uses an analysis of variance approach to evaluate the distances between clusters. In general, this method is quite efficient, however, it tends to create clusters in small size.

7. Conclusion

The World Wide Web has become an important medium for disseminating scientific publications. Web Citation Database is a data warehouse used for storing citation indices of Web publications. ACA can be used to identify authors from the same or similar research fields. In this paper, we have proposed a mining process for ACA from the Web Citation Database. The clustering results and author cluster map have been incorporated into the PubSearch system to support author retrieval of Web publications. The mining process automates the ACA process, which traditionally relies on statistical tools to do the analysis and requires human interpretation. As the Web Citation Database also contains other useful information, we are currently investigating other mining techniques for document clustering (Carpenter, Grossberg, & Rosen, 1991; Kohonen, 1995) and co-word analysis (Callon, Courtial, & Laville, 1991) to support publication retrieval. In addition, other mining tasks, such as identifying experts of a particular research area,

predicting the research trends, categorizing journals and locating the leading journals, can also be performed on the Web Citation Database.

References

Baeza-Yates, R. A. (1992). Introduction to data structures and algorithms related to information retrieval. In *Information retrieval: data structures and algorithms* (pp. 13–27). Englewood Cliffs, NJ: Prentice-Hall.

Boley, D. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, *2*(4), 325–344.

Bollacker, K., Lawrence, S., & Giles, C. (1998). CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the third ACM conference on digital libraries, Pittsburgh, PA* (pp. 116–123).

Bollacker, K., Lawrence, S., & Giles, C. (2000). Discovering relevant scientific literature on the web. *IEEE Intelligent Systems*, *15*(2), 42–47.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics*, *22*(1), 153–203.

Carpenter, G., Grossberg, S., & Rosen, D. (1991). Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, *35*(3), 401–420.

Chen, C., & Carr, L. (1999a). Trailblazing the literature of hypertext: author co-citation analysis (1989–1998). In *Proceedings of the 10th ACM conference on hypertext and hypermedia: returning to our diverse roots (Hypertext '99), Darmstadt, Germany* (pp. 51–60).

Chen, C., & Carr, L. (1999b). Visualizing the evolution of a subject domain: a case study. In *Proceedings of IEEE visualization 99, San Francisco, CA, USA* (pp. 499–502).

Chen, C., Chennawasin, C., & Yu, Y. (2000). Visualising scientific disciplines on the web. In *Proceedings of the 16th IFIP world computer congress. International conference on software: theory and practice, Beijing, China* (pp. 720–725).

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

Ding, Y. (1998). Visualization of intellectual structure in information retrieval: author co-citation analysis. *International Forum on Information and Documentation*, *23*(1), 25–36.

Everitt, B. (1986). *Cluster analysis* (2nd ed.). Hampshire, England: Gower Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1–34). Cambridge, MA: AAAI/MIT Press.

Garfield, E. (1980). ABCs of cluster mapping, Part 1, Most active fields in the life sciences in 1978. *Current Comments*, *40*, 5–12.

Gorsuch, R. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Green, P. E., Carmone, F. J., & Smith, S. M. (1989). *Multidimensional scaling: concepts and applications*. Boston: Allyn and Bacon.

He, Y. (2001). *Mining a web citation database for the retrieval of scientific publications over the WWW*. M. Eng. Thesis, School of Computer Engineering, Nanyang Technological University, Singapore.

ISI (2000). Institute for scientific information. http://www.isinet.com.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data (Prentice-Hall advanced reference series)*. Englewood Cliffs, NJ: Prentice-Hall.

Johnson, A. G. (1988). *Statistics*. Orlando, FL: Harcourt Brace Jovanovich.

Kohonen, T. (1995). *Self-organizing maps*. New York: Springer.

Kruskal, J. (1977). The relationship between multidimensional scaling and clustering. In *Classification and clustering* (pp. 17–44). New York: Academic Press.

Kurita, T. (1991). An efficient agglomerative clustering algorithm using a heap. *Pattern Recognition*, *24*(3), 205–209.

Larson, R. (1996). Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the 1996 Annual ASIS (The American Society for Information Science) meeting, Baltimore, US*.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, *48*, 40–54.

Lin, X. (2000). *Map of information scientists*. http://faculty.cis.drexel.edu/sitemap/sm2/citation.html.

McCain, K. (1990). Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science*, *41*(6), 433–443.

Mitchell, T. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 31–36.

Perry, C. A., & Rice, R. E. (1998). Scholarly communication in developmental dyslexia: influence of network structure on change in a hybrid problem area. *Journal of the American Society for Information Science*, *49*(2), 151–168.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation, Vol. 24* (pp. 249–284).

SPSS Inc. (2000). *Statistical package for the social sciences*. http://www.spss.com.

White, H. D. (1990). Author co-citation analysis: overview and defense. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 84–106). Newbury Park: Sage.

White, H. D., & Griffith, B. C. (1981). Author co-citation: a literature measure of intellectual structure. *Journal of the American Society for Information Studies*, *32*, 163–171.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: an author co-citation analysis in information science, 1992–1995. *Journal of the American Society for Information Science*, *49*, 327–356.

White, H. D., Buzydlowski, J., & Lin, X. (2000). Co-cited author maps as interfaces to digital libraries: designing pathfinder networks in the humanities. In *Proceedings of IEEE international conference on information visualization (IV'00). London, England* (pp. 25–30). Los Alamitos, CA: IEEE Computer Society.