Knowledge Discovery Through Co-Word Analysis

QIN HE

ABSTRACT

IN THE LAST HALF CENTURY, AS THE SCIENCE LITERATURE has increased dramatically, scientists found it increasingly difficult to locate needed data, and it is increasingly difficult for policymakers to understand the complex interrelationship of science in order to achieve effective research planning. Some quantitative techniques have been developed to ameliorate these problems; co-word analysis is one of these techniques. Based on the co-occurrence frequency of pairs of words or phrases, co-word analysis is used to discover linkages among subjects in a research field and thus to trace the development of science. Within the last two decades, this technique, implemented by several research groups, has proved to be a powerful tool for knowledge discovery in databases. This article reviews the development of co-word analysis, summarizes the advantages and disadvantages of this method, and discusses several research issues.

INTRODUCTION

Since World War II, the scope and volume of scientific research have increased dramatically. This is well reflected in the growth of the literature. In the 1960s, the amount of scientific literature was estimated to be doubling approximately every ten years (Price, 1963). Three decades later, in the 1990s, along with developments in information technology, especially in the area of data storage, the amount of information in the world is estimated to be doubling every twenty months (Frawley et al., 1991). In such a situation, it is hard for scientists to detect the subject areas and the

Qin He, Graduate School of Library and Information Science, University of Illinois, 501 E. Daniel Street, Champaign, IL 61820

linkages among these areas in their research fields, and policy makers have difficulties in mapping the dynamics of science to do research planning.

The traditional way to map the relationships among concepts, ideas, and problems in science is to seek the views of a relatively small number of experts. Even though such methods are indispensable for some purposes, as Law and Whittaker (1992) said, they also have certain drawbacks:

> First, they are extremely expensive unless the survey of experts is very small. Second, if the survey is small, then its representativeness is open to question. Third, the problem of collating a range of views about the way in which science has developed or is developing is complex. (pp. 417-418)

For these reasons, quantitative methods for mapping the structure of science have been developed; they include co-citation analysis, co-nomination analysis, and co-word analysis. This article reviews the development of the co-word analysis technique.

Co-word analysis is a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., words or noun phrases) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these texts. Indexes based on the co-occurrence frequency of items, such as an inclusion index and a proximity index, are used to measure the strength of relationships between items. Based on these indexes, items are clustered into groups and displayed in network maps. For example, an inclusion map is used to highlight the central themes in a domain, and a proximity map is used to reveal the connections between minor areas hidden behind the central ones. Some other indexes, such as those based on density and centrality, are employed to evaluate the shape of each map, which shows the degree to which each area is centrally structured and the extent to which each area is central to the others. By comparing the network maps for different time periods, the dynamic of science can be detected.

The co-word analysis technique was first developed in collaboration between the Centre de Sociologie de l'Innovation of the École Nationale Superieure des Mines of Paris and the CNRS (Centre National de la Recherche Scientifique) of France during the 1980s, and their system was called "LEXIMAPPE." For about twenty years, this technique has been employed to map the dynamic development of several research fields. One of the early studies was carried out by Serge Bauin (1986) to map the dynamics of aquaculture from 1979 to 1981. Based on the inclusion and proximity indexes, inclusion and proximity maps were created for 1979 and 1981.

With the decomposition of keywords into central poles and mediator words, the inclusion map for 1979 is shown in Figure 1 and that for 1981 is shown in Figure 2.
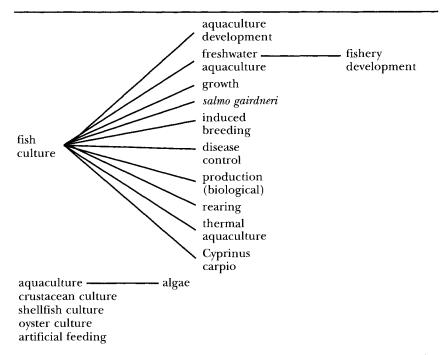
aquaculture
development

freshwater ———————— fishery
aquaculture   development

growth

salmo gairdneri

induced
breeding

disease
control

production
(biological)

rearing

thermal
aquaculture

Cyprinus
carpio

fish
culture

aquaculture ——————— algae
crustacean culture
shellfish culture
oyster culture
artificial feeding

Figure 1.   Main Central and Mediator Words, 1979   (Bauin, 1986, p. 127.)
Reprinted with kind permission of Macmillan Press, Ltd.

aquaculture ———————— algae
technique

freshwater
aquaculture

(salmonidae)
Salmo gairdneri

growth

rearing
(cyprinidea)

Cyprinus carpio

artificial feeding ——————— diets

induced breeding

fish culture
(pisces)

aquaculture   ———— fishery development
development   ———— brackishwater
                            aquaculture

crustacean culture
(malacostraca)

(bivalvia) ——————— oyster culture
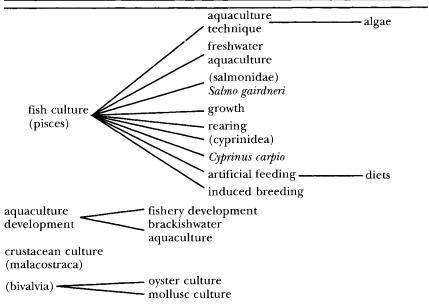   ———— mollusc culture

Figure 2.   Main Central and Mediator Words, 1981 (Bauin, 1986, p. 128.)
Reprinted with kind permission of Macmillan Press, Ltd.

In the map for 1979, "*Salmo gairdneri,*" a fish species which has been bred less extensively in Norway's seas since the 1950s, remained unexpectedly as a high frequency mediator word. However, in the map for 1981, this term was replaced by "salmonidae." One of the more significant changes is that the central pole "aquaculture" in the 1979 map has disappeared. It has been replaced by two new poles—"aquaculture development" and "aquaculture techniques." In addition, the word "artificial feeding" loses its status as a central pole in the map for 1979 and appears under "fish culture" in the map for 1981.

The proximity maps for 1979 and 1981 respectively are shown in Figures 3 and 4.
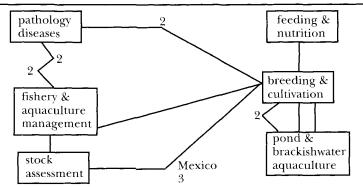

Figure 3. Proximity Map, 1979 (Bauin, 1986, p. 133). Reprinted with kind permission of Macmillan Press, Ltd.
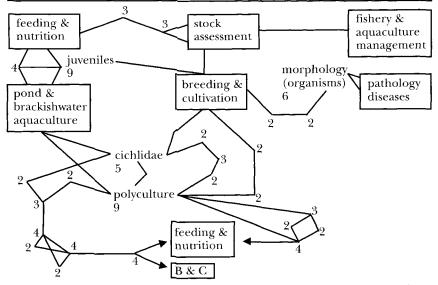

Figure 4. Proximity Map, 1981 (Bauin, 1986, p. 134). Reprinted with kind permission of Macmillan Press, Ltd.

Comparing the two maps, it is noted that, from 1979 to 1981, some clusters, such as "feeding and nutrition," become extended and more structured—i.e., the average number of links per word has increased. Overall, the average number of links per word in the complete maps has increased from 2.33 to 2.95. This might be an indication of the beginning of the integration of the whole field.

This and other examples (e.g., Turner & Callon, 1986; Callon, 1986; Courtial & Law, 1989; Law & Whittaker, 1992; Coulter et al., 1998) reveal that co-word analysis is a promising method for discovering associations among research areas in science and for revealing significant linkages that may otherwise be difficult to detect. It is a powerful tool that makes it possible to trace the structure and evolution of a socio-cognitive network (Bauin, 1986). As such, it offers a significant approach to knowledge discovery.

THE DEVELOPMENT OF CO-WORD ANALYSIS

In 1986, Callon, Law, and Rip (1986) edited a book titled *Mapping the Dynamics of Science and Technology*. This is a milestone work on co-word analysis. The first part of the book is an introduction on how to study the force of science. The second part is an analysis of the power of texts in science and technology, in which the authors have presented the theoretical foundation of co-word analysis, that is, "actor network." The third part is a detailed description of co-word analysis with examples. The last part is a conclusion.

Since publishing this book, co-word analysis has spread to researchers from not only France, but also the United Kingdom, the Netherlands, the United States, and some other countries. The process, measurement, and interpretation of co-word analysis has been improved to a great extent through these subsequent studies.

Theoretical Foundation—Actor Network

The co-word analysis technique was first proposed to map the dynamics of science. The most feasible way to understand the dynamics of science is to take the force of science in present-day societies into account. "Actor network" is the theoretical foundation for co-word analysis to map the dynamics of science (Callon, 1986).

Laboratories and literatures are considered as two powerful tools for scientists to change the world—they build complex worlds in laboratories and enforce them on paper (Latour, 1987). This implies that scientists attach particular importance to texts. They are not only using texts to publish their world built in the lab but also using texts as a way to build a world and enroll others. Even though science cannot be reduced to texts only, texts are still a prime source for studies on how worlds are created and transformed in the laboratory. Therefore, instead of following the

actors to see how they change the world, following the texts is another way to map the dynamics of science.

Based on the co-occurrence of pairs of words, co-word analysis seeks to extract the themes of science and detect the linkages among these themes directly from the subject content of texts. It does not rely on any a priori definition of research themes in science. This enables us to follow actors objectively and detect the dynamics of science without reducing them to the extremes of either internalism or externalism (Callon et al., 1986b).

Overall, co-word analysis considers the dynamics of science as a result of actor strategies. Changes in the content of a subject area are the combined effect of a large number of individual strategies. This technique should allow us in principle to identify the actors and explain the global dynamic (Callon et al., 1991).

Inclusion Index, Proximity Index, and Equivalence Coefficient

The first step of co-word analysis involves extracting keywords from records in indexing databases. After keywords are extracted from each document, a co-occurrence matrix of keywords can be constructed. Analyzing the interesting features of the co-occurrence matrix is the final and most important step of co-word analysis.

As different questions may be asked about the network of science, the co-occurrence matrix is subjected to various operations. A general co-word analysis is focused on two of these questions: one is to detect the hierarchies among the areas of a research problem, and the other is to detect the minor but potentially growing areas. In the early studies of co-word analysis, two indexes were introduced to address these two questions (Callon et al., 1986c).

The hierarchies of subject areas in a research problem can be detected by calculating an index, called the inclusion index (I_{ij}):

$$I_{ij} = C_{ij} / \min (C_i, C_j) \tag{1}$$

where,

C_{ij} is the number of documents in which the keyword pair (M_i and M_j) appears;

C_i is the occurrence frequency of keyword M_i in the set of articles;

C_j is the occurrence frequency of keyword M_j in the set of articles;

$\min (C_i, C_j)$ is the minimum of the two frequencies C_i and C_j.

I_{ij} has a value between 0 and 1, and it can be interpreted as a conditional probability. When $C_i > C_j$, that is, M_i is more general than M_j and includes M_j sometimes, I_{ij} measures the probability of finding M_i in an article given that M_j appears in it. An extreme case is that when $I_{ij} = 1$, M_j is fully included by M_i, that is, the M_j always co-occurs with M_i in the same article. The probability of finding M_i is 1, given M_j is found in the same article.

However, sometimes, even though I_{ij} has a low value, it is still significantly greater than the unconditional probability of finding M_i in any one of the N articles in the collection. Such a situation implies that there are some mediator keywords, which have a relatively low occurrence frequency but still have significant relationships with some of the peripheral keywords. To bring out such patterns, a proximity index P_{ij} is defined:

$$P_{ij} = (C_{ij} / C_i C_j) \bullet N \tag{2}$$

C_i, C_j, and C_{ij} have the same meaning as in formula (1). N is the number of articles in the collection. The mediator and peripheral keywords pulled out by P_{ij} represent minor but potentially growing areas.

In later co-word studies (e.g., Turner et al., 1988; Whittaker, 1989; Law & Whittaker, 1992; Coulter et al., 1996; Coulter et al., 1998), another index is employed to calculate the association values between word pairs. This coefficient is called the equivalence index (e-coefficient) (Callon et al., 1991) or strength (Coulter et al., 1998). It is defined as follows, where C_i, C_j, and C_{ij} have the same meanings as in formula (1):

$$E_{ij} = (C_{ij} / C_i) \bullet (C_{ij} / C_j) = (C_{ij})^2 / (C_i \bullet C_j) \tag{3}$$

E_{ij} has a value between 0 and 1. Similar to (1), E_{ij} measures the probability of word i appearing simultaneously in a document set indexed by word j and, inversely, the probability of word j if word i appears, given the respective collection frequencies of the two words. For this reason, E_{ij} is called "a coefficient of mutual inclusion" by Turner and his colleagues (Turner et al., 1988).

Inclusion Map, Proximity Map, and Sub-Networks

After the inclusion and proximity indexes are calculated, inclusion and proximity maps are created. The inclusion maps are designed to discover the central themes in a domain and depict their relationship to keywords that occur less frequently. The proximity maps are designed to discover connections between minor ideas hidden behind the central themes. These two kinds of maps correspond to two general types of studies. The first type of study involves getting more information about a certain theme. The second category of study concerns the analysis of the links between themes.

To create inclusion maps, the link that has the highest inclusion index value is selected first. These linked nodes become the starting points for the first inclusion map (subnetwork). Other links and their corresponding nodes are then added into the map in the decreasing order of their inclusion index until the threshold I_0 is reached. All nodes contained in the resulting cluster are removed from consideration as candidates in subsequent maps. The next map then starts with the link of highest inclusion index value of the remaining links. Keywords that appear on

the top level of inclusion maps are called "central poles" of the domain of research. Keywords that are included in the central poles, and themselves include some other words at lower levels, are called "mediator words" (Callon et al., 1986c).

The process to create proximity maps is similar to that for inclusion maps. The difference is that the proximity index is used instead of the inclusion index. If the threshold P_0 is lowered enough, more proximity connections between keywords will appear in the map and, eventually, the mediators and central poles found in inclusion maps will reappear. In this way, the relationship between minor issues and central poles can be studied (Callon et al., 1986c).

There is another method to construct clusters (or subnetworks) consisting of keywords that are more strongly linked internally than with keywords external to this sub-network (Callon et al., 1991). Essentially, this is similar to the inclusion maps above. The clusters could correspond to centers of interest in the research problem that are intensively studied by researchers. However, instead of using the inclusion index and threshold I_0, an e-coefficient is used in this method to measure the strength between keywords, and a threshold of ten is used to limit the number of words in one subnetwork. The procedure still starts from the link with the highest e-coefficient. When a cluster already has 10 words in it, the next link will be refused. The value of this link that is first refused is called the saturation threshold. After a cluster saturates, a new cluster is started. The e-coefficient value of the first link of this new cluster is called the "ceiling threshold." Based on the association value of the inter-cluster link and external links and the value of the ceiling threshold and saturation threshold, three distinct categories of clusters can be identified. The first category is isolated clusters, which are characterized by an absence (or low intensity) of links with other clusters. The second is secondary clusters, whose external links with other clusters above the ceiling threshold are sufficiently strong that it is legitimate to consider that they are the natural extension of one of these. The third is principal clusters, to which one or more other (secondary) clusters are associated by links whose value is lower than the saturation threshold.

Coulter et al. (1998) have divided the process of constructing subnetworks into two "passes." During Pass-1, the network is constructed similar to the process of creating inclusion maps above, but the e-coefficient is used to measure the strength of association between two keywords. In Pass-2, the network is extended by adding Pass-2 links. To be a Pass-2 link, both nodes of the link must be included in some Pass-1 network.

Density, Centrality, and Strategic Diagram

An earlier study was carried out to compare citation, co-citation, and co-word analyses of the state of five disciplines (Healey et al., 1986). It was

found difficult to analyze and accept the preliminary co-word results, and some experts doubted the reliability of the findings. The co-word technique evaluated in this study is called the "first generation" of co-word analysis by Law et al. (1988). A "second generation" analysis is presented in the same article to overcome the problems encountered in the comparison study.

In the "second generation" co-word analysis, a strategic diagram is used to illustrate the "local" and "global" contexts of research themes. This diagram is created by putting the strength of global context on the X axis and putting the strength of local context on the Y axis. This diagram is used in many later co-word studies. Two kinds of indexes (i.e., density and centrality) are used to measure the strength of local context and global context respectively.

Density. Density is used to measure the strength of the links that tie together the words making up the cluster; that is the internal strength of a cluster. It provides a good representation of the cluster's capacity to maintain itself and to develop over the course of time in the field under consideration (Callon et al., 1991). Ranking subject areas (clusters) in terms of their internal coherence (density) is designed to provide information for systematic discussion of a major policy alternative. Further, sorting the keywords by decreasing order of density can provide a precise description of the areas (Bauin et al., 1991).

The value of the density of a given cluster can be measured in several ways. Generally, the index value for links between each word pair is calculated first. Then, the density value can be the average value (mean) of internal links (e.g., Turner et al., 1988; Coulter et al., 1998), the median value of internal links (e.g., Courtial et al., 1993), or the sum of the squares of the value of internal links (e.g., Bauin et al., 1991). An internal link means both of the words linked by it are within the cluster.

Centrality. Centrality is used to measure the strength of a subject area's interaction with other subject areas. Ranking subject areas (clusters) with respect to their centrality shows the extent to which each area is central within a global research network. The greater the number and strength of a subject area's connections with other subject areas, the more central this subject area will be in the research network (Bauin et al., 1991).

For a given cluster (area), its centrality can be the sum of all external link values (e.g., Turner et al., 1988; Courtial et al., 1993) or the square root of the sum of the squares of all external link values (e.g., Coulter et al., 1998). More simply, it can be the mean of the values of the first six external links (e.g., Callon et al., 1991). An external link is a link that goes from a word belonging to a cluster to a word external to the cluster.

Strategic Diagram. A strategic diagram that offers a global representation of the structure of any field or subfield can be created by plotting centrality

and density into a two-dimensional diagram (Law et al., 1988). Typically, the horizontal axis represents centrality, the vertical axis represents density, and the origin of the graph is at the median of the respective axis values. This map situates each subject area within a two-dimensional space divided into four quadrants.

The strategic diagram is used in many co-word analysis studies (e.g., Turner et al., 1988; Courtial & Law, 1989; Turner & Rojouan, 1991; Callon et al., 1991; Coulter et al., 1998) and the analysis based on it is similar among these studies. Generally, the subject areas in quadrant 1 are both internally coherent and central to the research network in question. However, those areas in quadrant 4 seem to be of only marginal interest to work in the global research network. Coherent subject-specific areas always appear in quadrant 3 of the diagram. These areas are internally well structured and indicate that a constituted social group is active in them. However, they appear to be rather peripheral to the work being carried out in the global research network. Weakly structured areas are found in quadrant 2. These subjects, individually, are linked strongly to specific research interests throughout the network but are only weakly linked together. In other words, work in these areas appears to be underdeveloped, but it could potentially be of considerable significance to the entire research network. All these characteristics of a strategic diagram can be summarized in Figure 5.

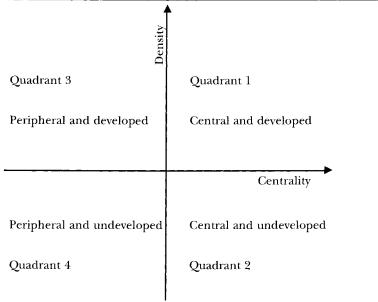|	Density (↑)	
Quadrant 3		Quadrant 1
Peripheral and developed		Central and developed
		Centrality (→)
Peripheral and undeveloped		Central and undeveloped
Quadrant 4		Quadrant 2

Figure 5. Strategic Diagram (Callon et al., 1991, p. 166.) Reprinted with kind permission from Akadémiaí Kiado publishers from an article in *Scientometrics*, 22(1), 155-205.

Comparative Analysis of Networks

The Stability of Networks. A striking feature of some strategic diagrams is the radical change in the configuration of the research network at two periods. This reflects the dynamics of science. Based on the strategic diagram, we can analyze the stability of the networks and foresee their changes in the future. This issue is addressed in many studies, and the methods used in these studies fall into two categories.

The first method used to study the stability of networks is directly based on the strategic diagrams (e.g., Callon et al., 1991; Turner & Rojouan, 1991). The findings can be summarized as showing that the probability for the research content of themes situated in quadrants 2 and 3 to change over time is significantly higher than it is for themes which are situated in quadrant 1. With a low density, the unstructured themes in quadrant 2 tend to undergo an internal structuring to improve their cohesiveness. With a low centrality, the scope of themes in quadrant 3 is likely to be extended in order to better articulate what is being done in the rest of the network. The reason as well as the goal for all these changes is to situate their work at the heart of their research network (quadrant 1). This can be done either by enlarging its scope or by improving its visibility through conceptual developments in the definition of a research program.

The second method is based on the ratio of centrality to density (c/d) (e.g., Courtial et al., 1993; Turner et al., 1994). The ratio (c/d) is considered as a meaningful indicator of the development stage of science and technology by many researchers. On the one hand, the findings show that, if this ratio tends toward 1, it indicates that this area is serving as a mainstream in the research network and is capable of redefining the global configuration of the system. On the other hand, if this ratio tends away from 1, it indicates the theme is falling out of favor and could well disappear as a subject of interest in the research network. However, Leydesdorff (1992a) claims "the c/d ratio is indeed a measure of the mutual information provided between the word distribution and the document distribution in that part of the structure" (p. 310) and cannot be used for this purpose.

Network Comparison. In co-word analysis studies, several subnetworks can be constructed concurrently while each network changes over time. To detect the difference among subnetworks simultaneously or subnetworks at different times is another issue studied by many researchers.

The comparison of two networks, N_1 and N_2, which might be two networks at different times or two distinct networks at the same time, can be done by a three-stage method (Callon et al., 1991).

The first stage is to compare the clusters. Let C_{1i} be the set of clusters of network N_1 and C_{2i} be the set of clusters of network N_2. A transformation

index (also called a dissimilarity index) is defined to measure the degree of dissimilarity between two given clusters. This index is defined as:

$$t = (W_i + W_j) / W_{ij} \qquad\qquad (4)$$

where,

   W_i is the number of words in cluster C_i;
   W_j is the number of words in cluster C_j; and
   W_{ij} is the number of words common to C_i and C_j.

For example, if the cluster C_i is defined by seven words and the cluster C_j by four words, and if four words among these eleven words are common to the two clusters, the transformation index is t = 11 / 4 = 2.8.

The second stage is to compare the positions in the strategic diagrams of those clusters demonstrated to be similar in stage one. This comparison can go beyond a simple enumeration of correspondences between clusters and bring out the relative position and degree of development of similar clusters within their respective networks.

The third stage is to create the life cycle curve of clusters in the case of dynamic analysis for a research network at times T_0, T_1, T_2 ... T_{10}. Suppose a set of similar clusters is identified in a comparative analysis at different times where C_{10} from T_0 corresponds to C_{11} from T_2; C_{11} corresponds to C_8 from T_3; and so on. This set of similar clusters is called a series. Clearly, the more stable a network is, the more series there are indicating the temporal propagation of its clusters. The existence of these series provides information about the progressive transformation of the clusters through time.

The transformation of networks and their intersections with other networks across time periods provides insights into the emergence of research themes. The similarity of networks in different time periods is also studied by Coulter et al. (1998). In this study, the authors employ the similarity index (SI), which comes from Callon's dissimilarity (or transformation) index above. It is defined as follows:

$$SI = 2 \bullet (W_{ij} / (W_i + W_j)) \qquad\qquad (5)$$

where,

   W_i is the number of descriptors in network N_i;
   W_j is the number of descriptors in network N_j; and
   W_{ij} is the number of descriptors common to N_i and N_j.

A constant 2 is multiplied to make the maximum value of SI to 1, which occurs when N_i and N_j have identical nodes. SI is used to measure the intersection of the descriptors in two networks and to examine the emergence of a network during a particular period.

Index of Influence and Provenance. Another comparative analysis is done by Law and Whittaker (1992) to highlight the overlap between themes on

similar subjects in succeeding time periods. Two indexes, the Index of Influence (I) and the Index of Provenance (P), are employed to measure the degree of continuity between themes in generations. These two indexes are calculated as follows:

$$I_{ij} = (2 \cdot M_{ij} + Ln_{ij}) / (2 \cdot N_i); \tag{6}$$

where,

M$_{ij}$ is the number of words in both theme i and (succeeding) theme j;

Ln$_{ij}$ is the number of words in both theme i and linked to subsequent theme j but belonging to no other theme in this generation; and

N$_i$ is the number of words in theme i.

$$P_{ij} = (2 \cdot M_{ij} + Ln_{ij}) / N_j; \tag{7}$$

where,

M$_{ij}$ is the number of words in both theme j and (preceding) theme i;

Ln$_{ij}$ is the number of words in both theme j and linked to preceding theme i but belonging to no other theme in this generation; and

N$_j$ is the number of words in theme j.

The index I$_{ij}$ shows the proportion of the words within a theme in one generation attached to any given theme in the next generation. A high I$_{ij}$ means that the "influence" of a first generation theme on one of the second generation is high. The P$_{ij}$ index shows the proportion of words within a second generation theme that come from any given theme in the preceding generation. A high P$_{ij}$ means that the "provenance" of a second generation theme primarily lies in a single theme of the first generation. Using these two indexes, the authors analyzed the continuities between themes and identified the lines of work in the field of acidification research. They were satisfied that they had detected a number of relatively stable themes by means of I and P indexes.

Frequency Analysis, Proximity Analysis, and Database Tomography

Database Tomography is a patented system for analyzing large amounts of textual computerized material (Kostoff et al., 1995). It can be considered as another generation of co-word analysis. Algorithms for extracting multi-word phrase frequencies and performing phrase proximity analysis are included in this system. Phrase frequency analysis can be used to discover the pervasive themes of a database while the phrase proximity analysis can be used to detect the relationships among these themes and between these themes and subthemes (Kostoff et al., 1997a). The indexes used in Database Tomography are similar to those used by traditional co-word analysis, such as the E$_{ij}$ equivalence index. But the co-occurrence of keywords is limited to ± 50 words within the text.

Similar to co-word analysis, Database Tomography can identify the main intellectual thrust areas and the relationship among these thrust areas. It provides a comprehensive overview of a research network and allows specific starting points to be chosen rationally for more detailed investigations into a topic of interest. Kostoff and his colleagues have employed Database Tomography tools to study chemical literature (Kostoff et al., 1997a). There are two appendixes in their article that show Database Tomography can be used for the generation of taxonomies and the identification of promising research directions.

Based on the term co-occurrence information, Database Tomography can also be used to expand the initial query in information retrieval (IR) systems and, in turn, allow the retrieval of relevant documents that would not have been retrieved with the initial query (Kostoff et al., 1997b). Simulated nucleation is the name given to the form of Database Tomography adapted to IR. In simulated nucleation, a core nucleus is developed first, and similar material is added as time develops until the desired amount of material is obtained. Then the main algorithms of Database Tomography (phrase frequency and phrase proximity analysis) operate on this core group of documents to identify patterns of word combinations in existing fields and generate new search term combinations that follow the newly identified patterns. The process is repeated until convergence is obtained, where relatively few new documents are found even though new search terms are added. Thus, simulated nucleation is running in a self-correcting cybernetic homeostatic model. It continually expands the coverage and improves the quality of the retrieval results.

Rotto and Morgan (1997) have employed frequency analysis and phrase proximity analysis techniques to study if the work in a dissertation abstract is potentially applicable to industry. The study first counts the frequency of every single-, double-, and triple-word phrase. Then, using the most frequently occurring technology-related word phrases as theme words, phrase proximity analysis is applied to construct clusters of word phrases that co-occur within abstracts. These clusters are then examined to investigate whether research subspecialties or related research focuses could be identified.

THE ADVANTAGES AND LIMITATIONS OF CO-WORD ANALYSIS

Advantages of Co-Word Analysis

Quantitative Over Qualitative. The drawbacks of qualitative methods have already been addressed at the beginning of this discussion. The advantages of co-word analysis over qualitative analysis were recognized by researchers from the time of its introduction. In the book by Callon et al. (1986a), the advantages of co-word analysis over qualitative methods have been shown at several points, for example:

> The problem of distinguishing between the successful and the less successful strategies of translation in qualitative analysis is solved by quantitative means: the aggregation of the co-occurrences of signal-words across a population of texts and the depiction of significant levels of such co-occurrences by graphical methods . . . . Using the quantitative in pursuit of the qualitative, we are also able to highlight features of scientific fields that have not always been recognized. The heterogeneity of scientific world-building is preserved in co-word analysis, where experimental findings, research methods, concepts, social problems, material artifacts and locations may appear together on the maps. (Callon et al., 1986d, pp. 225-26)

Qualitativists often jump from detailed analyses of scientific controversy to general explanations posed in terms of social interests:

> [Qualititativists] are unable to make the connection in a more detailed and less perilous manner. By contrast, the co-word approach, by summarizing articles in terms of forceful words and counting occurrences and co-occurrences to trace developments at an aggregate level, not only allows successful translations to be traced and distinguished from those that quickly disappear; it also makes it possible to uncover the many direct and indirect links that exist between translations whether or not these lead rapidly to social problems and interests. (Callon et al., 1986c, p. 108)

Flexibility. Compared with other methods of analysis that focus on texts, co-word analysis is much more flexible in that it shows the research network with graphs. On the one hand, these graphs can be simplified to show the overall structure of the network. On the other hand, one can zoom in on certain areas and trace the co-word patterns in as much detail as one wishes.

When Callon (1986) studied a collection of patents using co-word analysis, he took the flexibility advantage of co-word analysis and applied two techniques to analyze the maps. The first technique was to simplify the maps. It was found that certain words unified the whole of the field without really adding new information. When they were deleted, the structure of the maps was simplified but not altered. The second technique was to zoom in on a pole. The author used zooms as a means to carry out a detailed study of why a concept (i.e., "enzyme") totally disappeared from the inclusion maps in 1981 after having been a central pole in 1980. This zooming technique showed that this abrupt change between the two periods was more than a simple fluctuation; it was linked to the appearance of a small number of patents that introduced new centers of interest and reorganized existing relationships.

The technique of zooming in on certain areas to get more information on a specific word of medium frequency has also been used in other studies (e.g., Callon et al., 1986d). In addition, there is a technique proposed to do variable level clustering, which is another flexible way to show the maps of research areas at different levels (Turner et al., 1988).

Limitations of Co-Word Analysis

It is obvious that the quality of results from co-word analysis depends on a variety of factors, such as the quality of keywords and index terms, the scope of the database, and the adequacy of statistical methods for simplifying and representing the findings (Law et al., 1988). Solely making use of keywords and index words was the biggest problem of early co-word analysis. It was called "indexer effect" and was addressed by many researchers.

Callon et al. (1986d) mentioned one such problem when dependence was on indexing:

> Indexing is an intervention between the text and the co-word analysis, and the validity of the map will depend, to a certain extent, on the nature of the indexing. Yet since indexers try to capture what it is about a text that is interesting, they partially reproduce the readings that the texts are given within the field itself. Thus, despite the fact that indexing is not entirely reliable, validity is never totally absent. (p. 226)

Turner et al. (1988) questioned the schemes used in co-word analysis as follows:

> However, most of the work done in this area has used the classification schemes of the data base producers to draw conclusions. Designed for document retrieval, these schemes are generally not suited for monitoring changes in the state of technological art at any given moment in time. (pp. 320-21)

Whittaker (1989) has pointed out that the results of co-word analysis are dependent on how the indexers choose the keywords to conceptualize the scientific fields. However, the results from indexing are more akin to the conceptualizations of indexers than to those of the scientists whose work is being studied.

In addition to the "indexer effect," some other limitations of co-word analysis are also recognized by researchers, which include, but are not limited to, the following:

1) The representations of the results given by co-word analysis are too difficult to read. (Whittaker, 1989)
2) The coverage of database is incomplete. Certain types of literature, such as patents and the"grey literature," lie outside of the publication circuit and are not indexed in the database. This means that the results from co-word analysis cannot reflect the whole picture of the research field in question.
3) The delay between the writing of a document and the moment when it is indexed and entered into a database causes the co-word analysis to fail to detect emerging research themes at an early stage. (Callon et al., 1986)

SOME ISSUES IN CO-WORD ANALYSIS

Assumptions

The assumptions of co-word analysis are presented by Whittaker (1989) as follows:

> It [co-word analysis] relies upon the argument that (i) authors of scientific articles choose their technical terms carefully; (ii) when different terms are used in the same articles it is therefore because the author is either recognizing or postulating some non-trivial relationship between their referents; and (iii) if enough different authors appear to recognize the same relationship, then that relationship may be assumed to have some significance within the area of science concerned. (p. 473)

A fourth premise, that the keywords chosen by trained indexers as descriptors of the contents of articles are in fact a reliable indication of the scientific concepts referenced in them, makes it possible to use the keywords as the basic data for co-word analysis.

Later, Law and Whittaker (1992) have restated two of the assumptions above. First, co-word analysis assumes that the keywords used by indexers to index a paper reflect the present stages of the scientific research in question. Then, co-word analysis assumes that arguments received by other scientists will lead to the publication of further scientific papers that are indexed by similar sets of keywords.

If all these are reasonable assumptions, it is then possible for co-word analysis to make use of the frequencies of word pairs in an article set as a way to map the structure of concepts embodied in the articles.

Indexer Effect

As noted before, the indexer effect is one serious problem with co-word analysis. The main criticism against co-word analysis is also because of this. Many researchers tried to address this issue, and some tests have been done to overcome this problem.

One result of the indexer effect is that keywords assigned to the articles by the indexers are out of date. There are three sources contributing to this problem: (1) the lexicon used by indexers is itself out of date; (2) the indexers may use combinations of keywords that reflect the conventional views of science as they were previously; and (3) the inevitable delay between the publication of an article and the appearance of an entry in the database causes a problem (Whittaker, 1989).

In Law and Whittaker's study (1992), some experts are asked to evaluate the keywords assigned by indexers in the PASCAL database. Even though most comments are positive, three kinds of complaints are posed. First, some keywords assigned by indexers are too general. Second, one or two specific terms had been omitted from the satisfactory list. Third, errors and misplaced specificity are found—i.e., the indexer puts the wrong emphasis, or even a mistaken emphasis, in keywording.

Some tests have been carried out to overcome the indexer effect of co-word analysis. Generally, these tests make use of some mechanism to automatically index the database. Two examples are as follows:

- *Test with QUESTEL-PLUS.* QUESTEL-PLUS is a full text information retrieval system used by TELESYSTEMES in France. In collaboration with TELESYSTEMES, Callon and his colleagues (Callon et al., 1986) combined different techniques with QUESTEL-PLUS and ran LEXIMAPPE together with them. This established a completely computerized chain of procedure running from a QUESTEL-PLUS treatment of full-text literature to the automatic establishment of inclusion or proximity maps by the LEXIMAPPE.

  The study has tested the chain on a small dietary fiber file. In comparison with the manually indexed file, the results obtained are encouraging in three aspects. First, the general and redundant words, which complicate the maps without adding new information, are eliminated. Second, a much larger number of specific peripheral issues appear in the inclusion maps. Third, the structure of the proximity maps is much richer and more detailed.

- *Test with LEXINET.* Turner et al. (1988) have carried out a test to overcome the limits of manual indexing through the use of a computer-assisted indexing system known as LEXINET. The goal of the LEXINET system is to help an expert construct an indexing vocabulary suitable for a particular area of study by an interactive validation process between the expert and the system.

  The study shows that, with LEXINET, the indexing process is considerably accelerated. Since part of the delay between the writing of a document and the moment it is available for analysis is caused by manual indexing, using LEXINET can reduce the time lag considerably. Consequently, it improves the quality of the information available for a co-word analysis and essentially reduces the indexer effect.

During the last two decades, much progress has been made in the field of automatic indexing. With the development of automatic indexing, we should be able to considerably reduce, if not eliminate, the "indexer effect" found in the results of co-word analysis.

Related Statistical Methods

The statistical method used in co-word analysis is similar to the single linkage cluster algorithm. This method is simple and considered unreliable. Some other statistical methods have been studied to consider the possibility of using them in co-word analysis.

Courtial (1986) has compared correspondence analysis and multidimensional scaling with co-word analysis and indicated the limitation of the first two methods as follows:

- Since the goal of correspondence analysis is to extract a set of dimen-
  sions of decreasing importance in the same way as principal compo-
  nent analysis does for quantitative characteristics, the representation
  of objects or characteristics is limited to the space created by the first
  two dimensions. Applying this method in a test, the two first dimen-
  sions merely "explain" 11.2 percent of the total distances between key-
  words. The reason is that keyword coded scientific articles never have
  the usual features of characteristics attributed to objects. There is an
  inherent difference between keywords and characteristics. Keywords
  cannot be treated as characteristics if the associations between these
  characteristics must be the combinations of a small number of inde-
  pendent dimensions.
- Multidimensional scaling suffers from the same sort of difficulties. The
  goal of this method is to identify a configuration of words such that
  the calculated distances between the words can reflect the geometric
  distances as much as possible. This is done within a space, which is set
  at two or three dimensions beforehand. When applying this method
  in a test, it is possible to find some global properties of the field. But
  the results do not allow any more detailed analysis because the stress
  is far from negligible (pp. 190-92).

Leydesdorff has employed factor and clustering analysis techniques
in his co-word analysis. The approach is described as follows:

> Co-word analysis generates a symmetrical matrix with an empty di-
> agonal, i.e. word A AND word B happens as many times as B AND A.
> The matrices are factor-analyzed using both orthogonal and oblique
> rotations (to check for inter-factorial relations). For graphic repre-
> sentation, cluster analysis was pursued using Wards' mode of analysis
> with the cosine as the similarity coefficient. (as cited in Whittaker,
> 1989, p. 489)

There is an important difference between Leydesdorff's co-word analysis
and the co-word analysis we have described here. The former uses some
complicated statistical techniques to assign words into clusters while the
latter does not. The latter rests more upon the assumption that there is a
cluster-type structure and its algorithm is set to build those clusters link by
link according to the relative frequencies of words and co-words in the
document. The goal of the former method is to identify, list, and mea-
sure the distance between classes to create distinction rather than empha-
sizing connection and continuity. By contrast, the goal of the latter is to
describe a network of words and explore the qualitative character of the
links between them by concentrating on, and tracing out, connections
and crossroads in that network. So, the two methods are actually doing
different jobs and are appropriate for different purposes. Whittaker (1989)
compared these two methods in his study. He thinks that, if one is dealing

with a relatively homogeneous set of documents, it may be reasonable to assume that all the nontrivial title words should be included in the cluster structure. If the task is more complicated and the analysis is on a large and heterogeneous set, such an assumption seems unwarranted, and the method we have described here offers significant advantages.

Turner et al. (1994) have studied the co-word analysis techniques in connection with the local components analysis (LCA) in the GEODE (La Gestion Optimisee des Documents Electroniques) project. LCA is a neural algorithm designed to identify "data poles" and their "influence areas" in a document set. It can reveal local data structures in a very large data set. In this study, LCA was used to produce a data pole map. Each data pole is a constructed object in the GEODE system and described in the same way as the LEXIMAPPE generated objects—i.e., by a list of keywords. However, the LCA technique can supply additional information: it ranks the documents in a data flow according to their contribution to defining the emergence of a data pole.

Nederhof and van Wijk (1997) analyzed a co-occurrence matrix of the 104 most frequent nontrivial topics and 63 SSCI journal groups. They computed and transformed a discipline by topic correlation matrix into a discipline by discipline matrix. Disciplines with high correlation (Pearson $r > 0.88$) were merged. Two data sets were analyzed in this study. One set consisted of topics on which publication changed greatly, and that gives rise to a "dynamic" map. The other set consisted of a matrix of about 100 nontrivial topics that most frequently occurred in SSCI in 1986-1990, through which the "static" map was generated. Both sets of matrices are analyzed by means of combined cluster analysis and correspondence analysis. Both topics and disciplines were clustered separately but analyzed jointly in the correspondence analysis. Compared to co-word methods, this set mapping method has the important advantage that it related not just words to words, but also, in one single map, disciplines to disciplines and topics to disciplines.

Measurements

In addition to the measurements described in previous sections, researchers have also studied the probability of making use of some other measurements.

The usability of the Jaccard index and "statistical coefficient" was studied by Courtial (1986) as follows:

- The Jaccard index is often used to express the degree of intersection between two document sets and it is defined as:

$$J_{ij} = (C_{ij}) / (C_i + C_j - C_{ij}) \tag{8}$$

  This index can be used to measure the relative degree of overlap between "semantic areas" of words within a given database. However, it

cannot handle associations between low-frequency and high-frequency words very well, because it will have low values even when the low-frequency word always appears together with the high-frequency word. Therefore, the author thinks, this index can only be used to explore overlap between medium-frequency words.

- "Statistical coefficient" is similar to the proximity coefficient. It can be used to compare the observed frequency (C_{ij} / N) of a pair of words with the expected frequency of that pair if the words were independent (C_i / N $\bullet$ C_j/ N). Compared with the proximity coefficient, this coefficient has the advantage of being symmetrical and normalized. It is calculated as:

$$S_{ij} = 1 \: / \: S \bullet (C_{ij} - C_i C_j / N) \qquad (9)$$

where S is the standard deviation of the hypergeometrical distribution. According to Courtial, this coefficient is not usually used because the strength of association is not an important variable in the graphs. In addition, the computation of this coefficient takes a long time, while the extra information is not essential for interpretation.

In the study of Coulter et al. (1998), co-word analysis has been used to get an evolutionary perspective of software engineering. In order to measure the complexity of networks, they use the ratio of links to nodes L/N as a measurement. As (N-1)/N L/N (N-1)/2, the minimum value for L/N is ½. "Percentage of connectivity" is another related and normalized measure of a network's complexity. It is based on the ratio of the number of links in a network to its maximum possible number of links, that is, 2L/(N(N-1)). This value will be greater for simple stand-alone networks or subnetworks than that of larger networks because the numbers of nodes and links are fewer.

How to Interpret the Map

The maps obtained by co-word analysis are generally considered very difficult to understand by themselves. They have to be interpreted with caution. It is suggested by Callon et al. (1986) that the interpretation must be active and based on the comparison of inclusion and proximity maps. In some cases, it is necessary to make use of zooms and examine the original documents (or at least their descriptors). Collaborating with the experts is another way to improve the interpretation.

As the goal of co-word analysis is not to photograph a field of knowledge but to reveal the strategies by which actors mutually define one another, Callon and his colleagues (1991) suggest that the maps cannot be considered statistically and they must be interpreted dynamically. Attention should be paid to not only the internal dynamics of each network but also the interactions between networks. For the internal dynamics, we need to analyze the appearances, disappearances, transformations, and

movements in the series of clusters and the overall life cycles of clusters. For the interactions between networks, possible interactions include academic networks to general network, applied research to academic research, and some other complex interactions.

Where Should the Words Come from?

In 1987, Leydesdorff criticized the co-word analysis technique for the indexer effect, and his answer to this problem was to use title words instead of keywords as the basis of co-word analysis (as cited in Whittaker, 1989). This idea looks attractive because it might allow more direct access to the views of authors, and the descriptions can give more confidence to those who have doubts about the indexing process.

However, Whittaker (1989) points out that there are two difficulties in using title words. One is that authors might choose their title words deliberately in order to address a particular readership and produce an "audience effect." The other concerns the usage of nonstandard titles such as those in the form of a rhetorical question. To discover whether title words are preferable to keywords for co-word analysis, Whittaker has carried out a comparative study. He found that keyword analysis generates a picture similar to, but substantially more detailed than, that created by title word analysis. It does not show that either form of analysis is superior to the other. To some extent, this also proves the indexer effect is not a problem, at least in this case.

A literature review shows that the words used in co-word analysis are expanding from keywords in a lexicon to words in the full-text. In early studies, only keywords from a lexicon are used (e.g., Bauin, 1986). Later, the documents are indexed by title, summary, and a certain number of restricted keywords (descriptors) drawn from a lexicon from the study by Callon et al. (1991). Recently, Rotto and Morgan (1997) suggested co-word analysis could be performed on abstracts using words suggested by industry experts to help identify more specific research focuses within the research area of need. Finally, in Database Tomography, full-text words are used. One of the many advantages of full text over key or index words is the ability to retain low frequency but highly important phrases, since the keyword approach ignores the low frequency phrases (Kostoff et al., 1997a).

To What Extent Should the Words be Normalized?

Even after we have decided where to get the words, we still need to "normalize" them before we do a co-word analysis. Normalization has been addressed and done in several previous studies.

Turner et al. (1988) note that databases for information retrieval generally have to be "cleaned up" when they are used as a science evaluation tool. Strategies have to be devised to normalize institutional addresses, and country and author names, overcome the limits of manual indexing, and deal with multi-authored papers.

In the study by Courtial et al. (1993), the normalized title is used as a list of keywords. The WPIL patent database used in their study provides a normalized title for each patent family of the database, which is given by WPIL editors. These improved titles are based on the whole text of the priority documents. In addition, WPIL also makes use of thesaurus terms. Word processing is even used to improve the list of uniterms by joining a set of two succeeding words, such as joining "ice cream" to make "ice*cream." All these pre-processes on keywords enable the authors to obtain meaningful results in the study.

Nederhof and van Wijk (1997) have studied the association among topics in a discipline. The topics in the study are derived from words in the title of each article. To exclude idiosyncratic terms, only topics occurring at least ten times in a five-year period are analyzed. Many words for which British and American spellings differ have been standardized to the American spelling by the Institute for Scientific Information when they are put into the citation index databases.

Validation of the Method

Validation of LEXIMAPPE. Leydesdorff (1992a) has employed information theory to evaluate the LEXIMAPPE method of co-word analysis of scientific texts. LEXIMAPPE is criticized as follows:

> In LEXIMAPPE, only the strength of the association is computed, the strength of the association is not tested against an expected value for significance in terms of the distribution. As a consequence, two words which most strongly differ in terms of "structural equivalence" may occur in one cluster, and two words which do correlate significantly in terms of their distribution over the document set may occur in different clusters. The basic model is a graph-analytic relational model limited to diadic relations only. (p. 297)

In summary, Leydesdorff has shown that LEXIMAPPE uses rather different mechanisms to cut the "cake": on the basis of the comparisons of distributions, one finds completely different groupings than those abstracted on the basis of single co-occurrences.

In reply to Leydesdorff, Courtial (1992) first questions whether information theory can be used as an evaluation tool. He thinks information theory, when dealing with codes in a universe that are infinite, such as knowledge, confuses equiprobability and information, thus confusing disorder and information. In general, Courtial thinks Leydesdorff's article seems to deplore the attention paid by co-word analysis to infrequent but strong links to the detriment of global statistics.

Leydesdorff (1992b) gives the following reply to Courtial's comments. He thinks the relational algorithm (in LEXIMAPPE) informs us only about how the system reconstructs the information in the data and nothing about what this change means within the network. The relational approach

exhibits relations and hierarchies, not position and dimensions. In order to assess change and continuity, Leydesdorff thinks, one needs a hypothesis with respect to dimension (e.g., in order to know how to assess the author correlation in the data).

Representativeness of Co-Word Analysis. In another article, Leydesdorff (1997) questions co-word analysis again. He thinks words and co-words cannot map the development of science, because words change position not only in terms of the dimensional scheme of "theory," "methods," and "observation results," but also change in meaning from one text to another. By using the distribution of words over the sections, a clear distinction among "theoretical," "observational," and "methodological" terminology can be made in individual articles but not at the level of the set.

Courtial (1998) has given some comments on Leydesdorff's article above. He claims words are not used as linguistic items to mean something in co-word analysis, but as indicators of links between texts, whatever they mean. In co-word analysis, words are chain indexes, allowing one to compute translation networks. What is important for co-word analysis is not the exact meaning or definition of a word, but the fact that this word is linked to word X in one case and word Y in another case.

Again, Leydesdorff (1998) insists relational indexes cannot warrant inferences about strategic positions and that information calculus provides a useful tool for combining the static and dynamic analysis.

Selection of Method

The appearance of co-word analysis added another choice in the area of bibliometrics and provided another way to discover knowledge in databases. Similar to co-citation analysis, co-word analysis is also a kind of relational study based on the idea that publications should not be considered as discrete units. Instead, each is built upon others (Turner et al., 1988).

Co-citation and co-word analysis are the two most common methods used for constructing the thematic and strategic map of a field. Then which one should be selected?

In the study by Bauin et al. (1991), there were two reasons for choosing the co-word analysis method. The first reason is because they wanted to study the knowledge structure of the field rather than the relationship between researchers. Co-word analysis is based on the scientific content of publications and it serves their purpose directly. The second reason was methodological. They wanted to test the usefulness of co-word analysis in the process of strategic planning to see if it could be used as a tool in science management.

Callon et al. (1991) have shown why co-word is better than co-citation to study interactions between academic and technological research. The reason is the indicators used by co-citation only show the existence of a

link and cannot give any information on the subject or problem area in question. In order to know if it is scientific research or technology that has been the prime mover of an invention or an innovation, it is necessary to return to the documents themselves and to read the contents of the articles and patents identified. As the indicators used in co-word analysis can reflect the subject themselves, it is not necessary to go back to the original documents in all cases.

A review of the previous studies on co-word analysis shows the technique has been employed in the following types of studies:

- Mapping the dynamics of science (Callon et al., 1986a; Courtial & Law, 1989; Coulter et al., 1998).
- Mapping the structure of scientific inquiry (Whittaker, 1989).
- Mapping interaction between basic and technological research (Callon et al., 1991).
- Evaluating input/output relationships in a regional research network (Turner & Rojouan, 1991).

In addition, it is suggested by Callon and his colleagues (Callon et al., 1986) that co-word analysis should also be useful in the documentation field. It can be employed as a means to classify documents in terms of their evolving centers of interest. From this point of view, it should be useful both for retrospective retrieval and the construction and updating of thesauri.

CONCLUSION

Previous studies have shown co-word analysis to be a powerful tool to discover knowledge in databases. It has been used to detect the themes in a given research area, the relationship between these themes, the extent to which these themes are central to the whole area, and the degree to which these themes are internally structured. In the last twenty years, co-word analysis has been improved in many aspects. The main progress can be found in two fields:

1. *Source of words.* The early tests used the keywords assigned by indexers. Later, words in the title, summary, and abstract are used. Currently, the technical developments in full-text indexing make it possible to use words in full-text to do a co-word analysis. This will reduce the indexer effects greatly.
2. *Measurements.* The measurements used in co-word analysis have improved. The early co-word analysis used the inclusion and proximity indexes. A more general index, e-coefficient, was proposed later. Density and centrality are two other important measures that enable us to draw a strategic diagram.

However, there are still various kinds of problems remaining in the use of this method. One of the problems is the clustering algorithm. The clustering algorithm in current co-word analysis is very simple. Perhaps the other statistical clustering algorithms would work better. Another problem is the measurements. There are many ways to calculate the value of each index or coefficient. Research is needed to determine the relative effectiveness of the approaches. In addition, the procedure to select the files in the test collection, the elimination of "noise" from the data files, and so on also need further study. Improvements in the method will essentially depend on how it is used.

ACKNOWLEDGMENTS

REFERENCES

Bauin, S. (1986). Aquaculture: A field by bureaucratic fiat. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 124-141). London: The Macmillan Press Ltd.

Bauin, S.; Michelet, B.; Schweighoffer, M. G.; & Vermeulin, P. (1991). Using bibliometrics in strategic analysis: "Understanding chemical reactions" at the CNRS. *Scientometrics, 22*(1), 113-137.

Callon, M. (1986). Pinpointing industrial invention: An exploration of quantitative methods for the analysis of patents. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 163-188). London: The Macmillan Press Ltd.

Callon, M.; Courtial, J-P.; & Turner W. (1986). Future developments. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 211-217). London: The Macmillan Press Ltd.

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), 155-205.

Callon, M.; Law, J.; & Rip, A. (Eds.). (1986a). *Mapping the dynamics of science and technology: Sociology of science in the real world.* London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986b). How to study the force of science. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 3-15). London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986c). Qualitative scientometrics. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 103-123). London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986d). Putting texts in their place. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 221-230). London: The Macmillan Press Ltd.

Coulter, N.; Monarch, I.; & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science, 49*(13), 1206-1223.

Courtial, J-P. (1986). Technical issues and developments in methodology. In M. Callon, J.

Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 189-210). London: The Macmillan Press Ltd.

Courtial, J-P. (1992). Comments on Leydesdorff's "A Validation Study of LEXIMAPPE." *Scientometrics, 25*(2), 313-316.

Courtial, J-P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science, 49*(1), 98.

Courtial, J-P.; Callon, M.; & Sigogneau, A. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics, 26*(2), 231-242.

Courtial, J-P., & Law, J. (1989). A co-word study of artificial intelligence. *Social Studies of Science* (London), *19*, 301-311.

Frawley, W. J.; Piatetsky-Shapiro, G.; & Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge discovery in databases* (pp. 1-27). Cambridge, MA: AAAI Press.

Healey, P.; Rothman, H.; & Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*, 233-251.

Kostoff, R. N.; Miles, D. L.; Eberhart, H. J. (1995). *System and method for Database Tomography.* U. S. Patent Number 5440481. August 8, 1995.

Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R.; & Pellenbarg, R. (1997a). Database Tomography for technical intelligence: Comparative roadmaps of the research impact assessment literature and the *Journal of the American Chemical Society. Scientiometrics, 40*(1), 103-138.

Kostoff, R. N.; Eberhart, H. J.; & Toothman, D. R. (1997b). Database Tomography for information retrieval. *Journal of Information Science, 23*(4), 301-311.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press.

Law, J.; Bauin, S.; Courtial, J-P.; & Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. *Scientometrics, 14*(3-4), 251-264.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics, 23*(3), 417-461.

Leydesdorff, L. (1992a). A validation study of "LEXIMAPPE." *Scientometrics, 15*(2), 295-312.

Leydesdorff, L. (1992b). A reply to Courtial's comments. *Scientometrics, 15*(2), 317-319.

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the science. *Journal of the American Society for Information Science, 48*(5), 418-427.

Leydesdorff, L. (1998). Reply about using co-words. *Journal of the American Society for Information Science, 49*(1), 98-99.

Nederhof, A. J., & van Wijk E. (1997). Mapping the social and behavioral sciences worldwide: Use of maps in portfolio analysis of national research efforts. *Scientometrics, 40*(2), 237-276.

Price, D. S. (1963). *Little science, big science.* New York: Columbia University Press.

Rotto, E., & Morgan, R. P. (1997). An exploration of expert-based text analysis techniques for assessing industrial relevance in U.S. engineering dissertation abstracts. *Scientometrics, 40*(1), 83-102.

Turner, W. A., & Callon, M. (1986). State intervention in academic and industrial research: The case of macromolecular chemistry in France. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 142-162). London: The Macmillan Press Ltd.

Turner, W. A.; Chartron, G.; Laville, F.; & Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In A. F. J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 291-323). Netherlands: Elsevier Science Publishers.

Turner, W. A., & Rojouan, F. (1991). Evaluating input/output relationships in a regional research network using co-word analysis. *Scientometrics, 22*(1), 139-154.

Turner, W. A.; Lelu, A.; & Goergel, A. (1994). GEODE: Optimizing data flow representation techniques in a network information system. *Scientometrics, 30*(1), 269-281.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science, 19*, 473-496.