About Metrics of Bibliometrics

Milan Kunz

Chemopetrol, Research Institute of Macromolecular Chemistry, 65649 Brno, Czechoslovakia

Received May 14, 1992

It is shown that bibliometric incidence matrices can be treated as vectors in nm-dimensional space and characterized by statistics of their singular values. A case of a personal bibliography is demonstrated.

INTRODUCTION

Without sound metrics, any science trying to measure its object is lost and resembles a soul without a living body. Horace knew that there is a measure in all things, but Prothagoras failed to explain how it came about that the measure of all things is man. For this reason, it is necessary to arrange conferences addressing this question.[1]

Bibliometrics is just one of many sciences whose name ends in "metrics". Such sciences include biometrics, technometrics, scientometrics, chemometrics—all disciplines whose purpose is the precise measurement of an object. The metric sciences have achieved many successes: some statistical patterns have been discovered and proclaimed as laws,[2] and this has led to the requirement for their theoretical and philosophical interpretation.

Haitun[3-5] tried to show that human activity differs from its physical base by its infinite moments of characteristic distribution functions. Unfortunately, this is not true because humans are not immortal and their work cannot be infinite.[6,7] Speculation about character of information leads to conjectures by Khursin,[8] who built a complete system of scientific hierarchies resembling Smyth's results[9] which were based upon measurements of the Great Pyramid.

Information theory already has a mathematical basis. Rashewsky[10] proposed long ago that information forms a hypersurface in multidimensional space. This technique is already used in coding theory,[11] factor analysis of citation studies,[12-14] and analysis of databases,[15,16] but the philosophical and conceptual consequences for information laws have not been extracted from the mathematical formalism.

This situation is caused by difficulties connected with the notion of multidimensional spaces. In quantum chemistry their application is common, but even so, specialists have difficulties with their antiintuitive properties.[17] Lengthy disputes about localization of microparticles did not evolve into questions concerning the localization of information.

It should be of interest to readers of this Journal that the mathematical formalism describing information, or indeed information itself in the form of messages, is identical with the formalism used to describe chemical compounds as graphs.[18] It is not surprising: Information in the form of messages is not only a result of physicochemical processes in our brains, but simultaneously a trigger of these processes. Likewise, a chemical compound can exert effects similar to those produced by a message. We can consider literature as a form of external memory, an extension of the brain which is accessible to direct inspection and thorough analysis.

INCIDENCE MATRICES

The essential problem connected with applications of algebra to the analysis of information strings is rooted in the fact that information strings are noncommutative; $p + a + t$ is different from $t + a + p$. To overcome this difficulty, we must see a suitable formal representation of information using matrices.

In linear algebra, matrices are linear operators that transform one vector into another:

$$y = Mx$$

We will limit ourselves to a specific case in which

$$x = \mathbf{I}$$

where $\mathbf{I}$ is the unit diagonal matrix vector. Then y is identical with the matrix M itself. The outcome of this is elementary, but not trivial. A message is an operator whose task is to change somebody's mind.

A string of symbols from an alphabet of n symbols forms a message which can be interpreted as a naive matrix N having in each row just one unit vector:

$$e_j = (0_1, 0_2, ..., 1_j, ..., 0_n)$$

corresponding to the given letter j. This string can be mapped into the space of words, notions, or names. When some bibliometric analysis is made, we select some vectors, e.g., authors, as characteristic features and count their occurrences in a given set. This can be formalized as a projection of the matrix N onto the unit vector row $\mathbf{J}^T$, where the superscript symbol T means the transposition. This naive formalism revealed[19,20] that information is governed by two groups of cyclic permutations S_m and S_n represented by the unit permutation matrices P_m and P_n, which act independently on the information matrix N from the left and from the right, respectively, P_nNP_m. These symmetries can be separated by finding two quadratic forms:

$$P_n{}^T N^T N P_n \text{ and } P_m N N^T P_m{}^T$$

This led to a simple proof that the Boltzmann (H_n) and Shannon (H_m) entropy functions are distinct and additive.

We can continue to systematically build more complicated matrices and the corresponding multidimensional spaces. Matrices with two elements in each row, either the sums ($e_j + e_i$) or differences ($e_j - e_i$) are known as the incidence matrices of unoriented graphs (G) or the oriented graphs (S), respectively. Through these matrices or their quadratic forms, all applications of graph theory in chemistry are interrelated.[21] But in spite of the apparent simplicity of these matrices, some of their properties remained undiscovered until quite recently.[22]

All statistical linguistics and bibliometric studies are based upon the counting of distinctive words in sets of messages. When Lotka counted authors in the Chemical Abstracts Service Index, he ignored co-authors. Such simplification techniques are not generally allowable: as an example, in co-citation studies, it is necessary to be able to link papers

0095-2338/93/1633-0193$04.00/0

which are cited together. For this, we need matrices containing, in each row, an arbitrary number of nonzero elements representing papers in the rows and authors in the columns, or the citing papers in the rows and cited references in the columns. In conversation, we can express the importance of certain words by shouting them. This is because we have only a limited number of unit symbols. More generally, it is advantageous to express the weight of different terms by means of numbers.

MATRICES AND THEIR PROJECTIONS

A matrix M with elements m_{ij} is a vector in mn-dimensional space, referred to by statistical mechanics as the *phase space*. It describes a stochastic system completely but is too complicated and detailed a description. One can either see details without grasping the whole or get a picture of the whole while missing the details. For a system of molecules of a gas, for example, you cannot feel individual molecules, but you can feel their mean motion as temperature or wind. With information vectors, we can easily read all the words, but have more difficulty finding parameters such as the mean productivity of specific authors, or the importance of different fields, which characterize the information system as a whole. It should be possible to find these parameters by means of statistical treatment of the generalized incidence matrices M. As an example, three types of incidence matrices can be given

N^T	1	0	1	0	1	N^TJ	3
	0	0	0	1	0		1
	0	1	0	0	0		1

(– word ACABA). This matrix has just one nonzero symbol in each column (transposed row):

M^T	1	1	1	1	0	M^TJ	4
	1	0	0	0	0		1
	0	0	1	0	1		2

This matrix has $1 - n$ unit symbols in each row. This notation is used in music for different simultaneous tones:

${M_1}^T$	0.5	1	0.8	1	0	${M_1}^TJ$	3.3
	0.5	0	0	0	0		0.5
	0	0	0.2	0	1		1.2

(weighted matrix M^T). The column sums are always 1. The weights can be equal, as in 0.5 + 0.5, or unequal, as in 0.8 + 0.2.

The matrix vector is simplified if we determine its projections into its subspaces—either into the subspace of the columns or the subspace of its rows. This is easily done by finding its scalar products with the unit vector row $\mathbf{J}^T$, that is $\mathbf{J}^T\mathbf{M}$, and with the unit vector column $\mathbf{J}$, which is $\mathbf{MJ}$. These scalar products are merely the column or row sums of matrix elements, as in our examples, where the transposed form $(\mathbf{N}^T\mathbf{J})^T = \mathbf{J}^T\mathbf{N}$ is used to conserve space.

The relationship of a matrix vector to both projections is shown in Figure 1. Here, the original nm-dimensional vector $\mathbf{M}$ was somewhere in the Hilbert space on a sphere with diameter:

$$L = (\sum m_{ij}^2)^{1/2}$$

The traces of both quadratic forms have equal length $\mathbf{M}^T\mathbf{M}$ and $\mathbf{M}\mathbf{M}^T$. This follows from the rules for matrix multiplication or from the rules for finding the quadratic forms of vectors. Thus

$$L^2(M) = Tr(M^TM) = Tr(MM^T)$$

where $Tr(M)$ is the trace of the matrix, the sum of its diagonal elements.

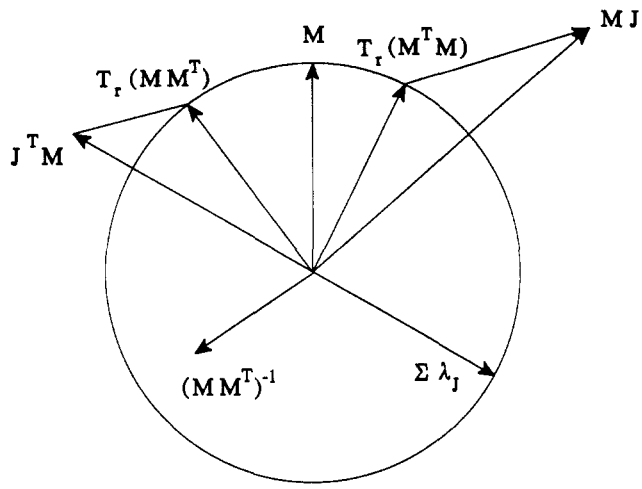

Figure 1. System of vectors from the information incidence matrix M. $\mathbf{M}$, the information vector, is a string of words which leads our mind to some state. We can choose just some parts, such as authors, references, or keywords which then replace the original information. MJ is the projection of the matrix M into m-dimensional space. If the matrix M is a text, it has in each row just one symbol, or word, and then $MJ = J.J$ is the unit vector which makes row sums of matrix elements. J^TM is the projection of the matrix into n-dimensional space. Its elements are the column sums of elements in the information matrix M. In both projections, we abstract some features of the original information and get statistics. $Tr(M^TM)$ and $Tr(MM^T)$ are the trace vectors of the corresponding quadratics forms. They have the same length as the matrix vector, $\mathbf{M}$. $\sum\lambda_j$ is the vector of singular values of the matrix M or eigenvalues of both quadratic forms. It is the matrix vector $\mathbf{M}$ in rotated coordinates. $(MM^T)^{-1}$ are inverse vectors if they are finite.[22,23,26] Their importance with respect to information vectors has not been investigated. MJ^T, $Tr(M^TM)$, and off-diagonal elements of M^TM form a right triangle in the Hilbert space. The second triangle is formed by J^TM, $Tr(MM^T)$, and the off-diagonal elements of MM^T.

The difference between the trace vectors and both projection vectors is in the off-diagonal elements of both quadratic forms. The diagonal and the off-diagonal elements form, in multidimensional space, a right triangle.

If an information matrix is "naive" in that all its columns are orthogonal and all the off-diagonal elements of the quadratic form N^TN are zero, the right triangle reduces to a straight line. By deducing this quadratic form, we transform a message into its statistics, and thus we now know which words were used and how many times, but we cannot determine the meaning of the message. If off-diagonal elements exist in the quadratic form, the trace will have the same length as the original vector $\mathbf{M}$, as is shown in Figure 1, but it will not coincide with it. Such a matrix vector is better represented by the eigenvalues of the quadratic forms M^TM or MM^T (both forms have equal eigenvalues). These eigenvalues are known as the singular values of the original matrix M. They are obtained by diagonalization of matrices, a process that is painless when carried out by some computer programs.

When we speak of symmetry of information matrices, we have rotated an information vector in a fixed coordinate system and considered all the matrices obtained by such permutations to be equivalent and lying on a spherical orbit. When we search eigenvalues, we leave the matrix vector in position and rotate the coordinate system in an attempt to find combinations of unit vectors $\mathbf{e}_j$ in which the matrix M appears as a diagonal

vector. Such unit vectors are known as eigenvectors, or factors, and are explained in most chemometrics textbooks.

There is still another way in which the relationships between both quadratic forms can be interpreted. We can take as the foundation stone of the information space, the adjacency matrix[21] A formed as a block matrix in which the diagonal blocks 0 are zero matrices and the off-diagonal blocks are the matrix M and its transpose. The adjacency matrix will be symmetrical, and its quadratic forms will coincide with its square, A^2, which will split into two diagonal blocks, M^TM and MM^T. They form separate orthogonal components of the original $(m + n)$-dimensional space.

$$A = \begin{array}{c|c} 0 & M \\ \hline M^T & 0 \end{array}$$

There seems to be a paradox in that the matrix M^TM with n rows and columns corresponds to the projection of the matrix M into m-dimensional space while the m-dimensional square matrix MM^T corresponds to the projection of the matrix M into n-dimensional space. This discrepancy can be explained in terms of the elements of both matrices.

If the matrix M is the incidence matrix of publications and authors, then columns and row vectors of the correlation matrix M^TM are authors, but elements of these vectors are publications. Authors are represented by their publications, which are measures of their productivity. Shares of individual authors appear on the main diagonal, and off-diagonal elements show publications common to the given pair of authors. Conversely, the elements in the space of publications, MM^T, are authors. This provides a formal explanation for Prothagoras' remark that man is the measure of all things. It is well-known that in both subspaces, it is possible to determine distances between authors or publications as paths in a graph,[22] which gives the local properties of the system described by the matrix M. These distances are connected in an intricate manner to the inverses of both quadratic forms.[23-25]

It is customary to characterize the position of an information vector by a function. Lotka[26] derived a statistic from the Author Index of Chemical Abstracts. He counted the number n_k of authors having m_k publications and then expressed the first number as a function of the second:

$$n_k = f(m_k)$$

Such a function should describe the matrix vector M and its position in multidimensional space. This approximation is good when dealing with a naive matrix N, whose column sums J^TN coincide with the column sums of its quadratic form J^TN^TN. In a general case, the diagonal values of the quadratic form do not coincide with its eigenvalues. For our examples, we have

$$M^TM \begin{vmatrix} 4 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{vmatrix}$$

This matrix has the diagonal values 4, 2, and 1 and eigenvalues 5.40, 1.32, and 0.28. For the weighted matrix $M_1{}^TM_1$, the corresponding diagonal values are 2.89, 1.04, and 0.25, while the eigenvalues are 2.93, 1.03, and 0.23. Here the difference is small, but none of the values are simple sums. In the unweighted matrix, the difference is sufficiently large to merit investigation.

Now, to the point of this exercise. We know that eigenvalues of matrices characterizing physical objects describe the objects' physical and chemical properties and are, thus, more important than any of the explicit matrix elements. This being the case, these parameters can also be more important in information systems which make use of incidence matrices. Because of this, distributions of singular values could be more interesting than distributions obtained by direct counts. Studies of the distributions of singular values have already been reported,[27] but were done for other purposes, namely to determine the rank of the correlation matrices. In the present case, these distributions identify the struture of the information field.[28,29]

If, in an incidence matrix of authorships, more than one entry occurs in a row, publications are authored by groups. Many studies of different aspects of group authorships exist,[2] and we can ask how collective authorship affects the extremely skewed statistical distribution known as the Lotka Law. Such matrices can be treated in three different ways. Full authorship may be attributed to each co-author; they can be weighted, evenly or unevenly; and, as an extreme case, the full merit can be assigned to a single author, as was done by Lotka,[26] for pragmatic reasons. Pao recommends[29] that this single author be the senior author. An incidence matrix is naivized by such a procedure, but the picture of the system is simultaneously distorted, and it is necessary to find some techniques that will overcome this deficiency.

In some instances such as personal bibliographies, it is not possible to select a single author from the various co-authors. Personal bibliographies are not linked by a common subject, but by a common author. In earlier papers, he is usually a junior author and is the senior author only in later papers. Not only do his papers have co-authors, but his papers will appear in the bibliographies of his co-authors. As a test set, we can examine the first 150 publications of I. Gutman,[30] who is a chemist in the Zagreb group, working on eigenvalue problems of chemical graphs.[31] A total of 32 co-authors are found in this bibliography, and so the bibliography matrix, with its 150 rows (publications), has 32 columns (co-authors). The numbers of co-authors are summarized in the table below:

no. of co-authors	1	2	3	4	5	Σ32
no. of papers	64	49	26	8	3	Σ150

The arithmetic mean of the number of co-authors is 1.91. The incidence matrix is sparsely populated, with approximately only two nonzero elements per row. More than two-fifths of the publications were by Gutman alone, but there were 48 papers published with his tutor, Trinajstić. With 13 of his co-authors, Gutman has published only one paper, as may be seen from the full breakdown in Table I, which also contains the distributions of unweighted and evenly weighted authorships on a logarithmic scale, together with the corresponding singular values.

The distribution reveals a typical pattern of extremely skewed information distribution. If such shapes are common for personal bibliographies, or whether they are specific for an exceptional author, cannot be determined in the absence of comparisons with other cases. Consequently, it is more important to show how the distribution of singular values differs from the distribution of authorships than to find some analytical function.

Both singular values have a singularity of five zero values, and the skew of the other values is much less pronounced than for n_k values. This differs from a case with, say, 13 collaborators having only one publication in common, but it can be compared to a case of five co-authors where the weighted

Table I. Co-authorship Statistics of Publications by I. Gutman[a]

logarithmic scale $\log_2 m_k$	unweighted authorships		weighted authorships	
	matrix sums J^TM	singular values	matrix sums J^TM_i	singular values
< −9	0	5	0	5
−4 to −9	0	0	0	0
−3	0	0	5	4
−2	0	0	9	8
−1	0	0	4	9
0	13	5	2	1
1	7	2	5	3
2	3	11	4	0
3	5	7	1	0
4	2	1	0	1
5	0	1	1	0
6	1	0	0	0
7	0	0	1	0
8	1	0	0	0

[a] This table covers Gutman's first 150 publications.[29] Extremely skewed information distributions such as this are most simply modeled by the truncated log normal distribution.[7] The logarithmic scale is used to form classes according to the sums m_k, of unweighted or evenly weighted authorships. The distribution of co-authorships is modeled satisfactorily, but singular values show a singularity corresponding to authors with the lowest degrees of collaboration.

co-authorship values are only 0.2. These co-authors collaborated with Gutman only once, and the resulting publication had five other co-authors.

DISCUSSION

The foregoing exercise in linear algebra has shown, we believe, that man is a measure of all things only in the subspace of "things". In the subspace of "man", "things" such as words, publications, citations, or money measure the importance of people. Of course, such subspaces must first be related by some incidence matrices. These relationships already exist: the problem is that we are unable to formulate the corresponding matrices.

Significant parameters hide behind the apparent parameters obtained from simple bibliometric counts, and they can be calculated from the corresponding implicit or explicit matrices. The problems of eigenvalues of chemical graphs are important and have been studied for many decades[30] by a branch of mathematical chemistry.

Factor analysis introduced many years ago by psychometricians seeks to interpret eigenvalues of correlation matrices as the proportion of the dispersion that is explained by corresponding factors. According to this interpretation, Gutman authored 150 publications for which the unweighted authorship is 287, and his share in the bibliography is therefore 52.7%. This is essentially the same as the figure of 53.8% of the first eigenvalue of 80.7 derived from 150 weighted authorships. But the first eigenvalue 173 of unweighted authorships from 287 is 60.5%. This gives greater merit to the first author which, unfortunately, is not always Gutman himself. The zero eigenvalues seem to represent co-authors who participated on a single Gutman publication with five authors, but it is dangerous to draw firm conclusions from a single case.

The differences between the simple counts and the eigenvalues could be used for evaluation of matrix structures and are similar to the complex structures measures studied by *Kretschmer*.[32]

If patterns of scientific research are becoming more complicated, bibliometric analysis cannot progress by simplifying the problems but rather by improving its methods. Evaluation of computer-developed statistics by simple correlation will become as obsolete as manual analysis. When, in a sparsely populated matrix of co-authorships, a significant difference is discovered between simple counts and singular values, then a much greater difference must be expected in citation matrices with tens of nonzero elements in each row. It is relatively easy to count singular values and to study their distributions. Computers need not be used only to register millions of compounds and to estimate their properties but also to unveil the mysteries of the chemical literature and its authors.

REFERENCES AND NOTES

(1) Elkana, Y.; Lederberg, J.; Merton, K. R.; Thackray, A.; Zuckerman, H. *Towards a Metric of Science*; Wiley: New York, 1978.
(2) White, H. D.; McCain, K. W. Bibliometrics. *Annu. Rev. Inf. Sci. Technol.* **1989**, *24*, 119–186.
(3) Haitun, S. D. Stationary Scientometric Distributions. I. Different Approximations. *Scientometrics* **1982**, *4*, 5–25.
(4) Haitun, S. D. Stationary Scientometric Distributions. II. Non-Gaussian Nature of Scientific Activities. *Scientometrics* **1982**, *4*, 89–101.
(5) Haitun, S. D. Stationary Scientometric Distributions. III. The Role of the Zipf Distribution. *Scientometrics* **1982**, *5*, 375–395.
(6) Kunz, M. A Case Study Against Haitun's Conjectures. *Scientometrics* **1988**, *13*, 25–33.
(7) Kunz, M. Can the Lognormal Distribution be Rehabilitated? *Scientometrics* **1990**, *18*, 179–191.
(8) Khursin, L. A. About the Substance of Information Flows as the Reflection of the Dynamic Structure of the Weighted Base of the Short-Lived Memory of the Human Brain. *Nauch. Tekh. Inf.* **1970**, *2*, 10–19, (in Russian).
(9) Edwards, I. E. S. *The Pyramids of Egypt*; Penguin Books; Harmondsworth, 1961; pp 295–296.
(10) Rashewsky, N. Some Bio-sociological Aspects of the Mathematical Theory of Communication. *Bull. Math. Biophys.* **1950**, *12*, 359–378.
(11) Hamming, R. W. *Coding and Information Theory*; Prentice-Hall: New York, 1980.
(12) Pinski, G.; Narin, F. Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Applications to the Literature of Physics. *Inf. Process. Manage.* **1976**, *12*, 297–312.
(13) Noma, E. Untangling Citation Networks. *Inf. Process. Manage.* **1982**, *18*, 43–53.
(14) Noma, E. Co-Citation Analysis and the Invisible College. *J. Am. Soc. Inf. Sci.* **1984**, *35*, 29–33.
(15) Dou, H.; Hassanaly, P. Automatic Generation of the Strategic Matrices from Online Databases. *World Patent Inf.* **1991**, *4*, 223–229.
(16) Quoniam, L. Bibliometrics of Bibliographic References: Methodology. In *La Veille Technologique*, Desvals, H., Dou, H., Eds.; Dunod: Paris, 1992; pp 243–262 (in French).
(17) Mezey, P. G. *Potential Energy Hypersurfaces*; Elsevier: Amsterdam, 1987.
(18) Randić, M. Representation of Molecular Graphs by Basic Graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57–69.
(19) Kunz, M. Information Processing in Linear Vector Space. *Inf. Process. Manage.* **1984**, *20*, 519–525.
(20) Kunz, M. Natural Phase Spaces. In *Problems in Quantum Physics. II. Gdansk, 89*; Mizerski, J., Posiewnik, A., Pykacz, J., Zukovski, M., Eds.; World Scientific: Singapore, 1990; pp 377–389.
(21) Rouvray, D. H. The Topological Matrix in Quantum Chemistry. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1975; pp 175–221.
(22) Odda, T. On Properties of a Well-Known Graph or, What is Your Ramsey Number? *Ann. N.Y. Acad. Sci.* **1979**, *328*, 166–172.
(23) Kunz, M. On Topological and Geometrical Distance Matrices. *J. Math. Chem.*, in press.
(24) Kunz, M. Path and Walk Matrices of Trees. *Coll. Czech. Chem. Commun.* **1989**, *54*, 2148–2155.
(25) Kunz, M. A. Moebius Inversion of the Ulam Subgraphs Conjecture. *J. Math. Chem.* **1992**, *9*, 297–305.
(26) Lotka, A. The Frequency Distribution of Scientific Productivity. *J. Wash. Acad. Sci.* **1926**, *16*, 317–323.
(27) Malinowski, E. R. Theory of the Distribution of the Error Eigenvalues Resulting from Principle Component Analysis. *J. Chemom.* **1987**, *1*, 33–40.
(28) Quoniam, L.; Dou, H.; Hassanaly, P.; Mille, G. Bibliometrics and Chemistry. A Case of Fat Acids and Phospholipides. *Analusis* **1991**, *19*, 148–152 (in French).
(29) Pao, M. L. Lotka's Law: A Testing Procedure. *Inf. Process. Manage.* **1983**, *21*, 305–320.
(30) Gutman, I. Scientific Publications of Ivan Gutman (1-100), (101-150), personal communication.
(31) Trinajstić, N. The Characteristic Polynomial of a Graph. *J. Math. Chem.* **1988**, *2*, 197–215.
(32) Kretschmer, H. Representation of a Complex Structure Measure for Social Groups and its Application to the Structure of Citations in a Journal. *Scientometrics* **1983**, *5*, 5–30.