Scientometrics, Vol. 11. Nos 3–4 (1987) 217–229

THE GOZINTO THEOREM: USING CITATIONS TO DETERMINE INFLUENCES ON A SCIENTIFIC PUBLICATION

R. ROUSSEAU

Katholieke Industriele Hogeschool West-Vlaanderen, Zeedijk 101, 8400 Oostende (Belgium)
and
Universitaire Instelling Antwerpen, Speciale Licentie Documentatie- en Bibliotheekwetenschap, Universiteitsplein 1, 2610 Wilrijk (Belgium)

(Received June 27, 1986)

This paper gives a mathematical technique to study influences, using citations. Taking into account both the publications that have a direct influence and those that have an indirect influence, we obtain the total influence measure on a fixed paper.

Introduction

Several bibliometric techniques are used to study relations between scientific publications. Bibliographic coupling and co-citation analysis are probably best known. Bibliographic coupling, a term introduced by *Kessler*,[1] is one of the first means of describing relations among scholarly papers. *Kessler* postulated that scientific papers bear a meaningful relation to each other when they have one or more references in common. His classic paper on bibliographic coupling[2] appeared in *American Documentation* in 1963. An excellent review on bibliographic coupling is given by *Weinberg*.[3]

Co-citation can be viewed as a variation of the idea of bibliographic coupling. It was proposed independently by *Small*[4] and *Marshakova*.[5] While bibliographic coupling focuses on groups of papers citing a source document, co-citation occurs when two (or more) documents are cited in the reference list of a third document. This technique has been extensively used [see f.i. *Marshakova*[6]].

In this paper we would like to explain a different, more elementary, method of citation analysis. Using the references of a particular paper we intend to determine what publications have had the greatest influence on the development of the paper under study. We take into account both the publications that have a direct influence and the ones that have an indirect influence. We claim that the publications mentioned in the reference list of a paper have a direct influence on this paper. Such publications are called the first generation. Publications taken from the reference list of one of the first generation papers, and not belonging to the first generation, form the second generation and so on. The direct influence a publication has, can be given a weight explicitly or not (in which case the weights may be set equal to 1).

In his paper on structural models of complex information sources *Zunde*[7] distinguishes three application areas of citation analysis:

(1) qualitative and quantitative evaluation of scientists, publications and scientific institutions;

(2) modeling of the historical development of science and technology;

(3) information search and retrieval.

Our work falls under the first category, being a method for the evaluation of a scientific publication, which exploits the idea that a path between two vertices of the citation graph can be interpreted as a chain of stimulation and fertilization (cf. *Zunde*[7] p. 13).

More specifically, we claim that by using total influence measures, we gain insight in the stream of ideas which led an author to the results given in a particular paper, even if he himself was perhaps not fully aware of it. (If he was he might have put the appropriate references in his reference list.) To obtain these results on total influences we assign weights to citations. The main subject of this paper, however, is a new interpretation of a known mathematical technique that will allow us to solve the following question: given the fact that we want to go as far as the n-th generation and given all the direct influence weights, how can we calculate the total influence of each publication resulting from these n generations on the paper (or papers) we have started from. Notice that very quickly one is confronted with hundreds of papers, so that doing the calculations by hand is downright impossible. We will argue that the Gozinto theorem (*Vazsonyi,*[8] *Staelens*[9]) combined with adequate computer techniques yields a solution to our problem.

This so-called Gozinto theorem (said to be developed by the celebrated Italian mathematician Zepartzat *Gozinto*) was originally used in connection with shipping schedules. In this context it says that the total requirement factor matrix (C) is obtained from the next assembly quantity matrix (A) by the formula: $C = (I - A)^{-1}$ [*Vazsonyi,*[8] p. 435, formula (10)].

In this paper we will not consider the problem of assigning influence weights nor will we discuss the full impact of our method: this is left for a next publication. We will however give a smallscale example, mainly to illustrate the mathematics involved. This provides a first (and partial) justification for our main claim.

The Gozinto theorem

While explaining the general ideas of the Gozinto theorem we will illustrate them by the following fictitious example (see Table 1 and Fig. 1). To make this example not to complicated we have considered only two generations.

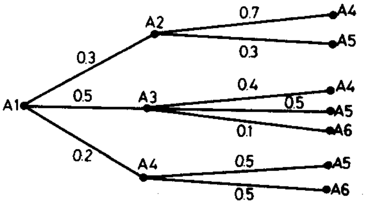

Fig. 1. Tree structure of Table 1

Table 1
Example

Paper	References	Weight
A1	A2	0.3
	A3	0.5
	A4	0.2
A2	A4	0.7
	A5	0.3
A3	A4	0.4
	A5	0.5
	A6	0.1
A4	A5	0.5
	A6	0.5

Let V be the set of all publications under consideration, this is: the set of all articles appearing at least once in one of the $n + 1$ generations under study. (The paper or papers we have started from form the 0-th generation.) We denote by z the number of elements in the set V. In our example $V = \{A1, A2, A3, A4, A5, A6\}$ and $z = 6$.

In V we consider two relations: the relation R, which means "is cited by" and which will be interpreted as "has a direct influence on" and the relation $\bar{R}$ which is the transitive closure of R, this is the transitive relation on V containing all ordered pairs in R and the smallest possible number of ordered pairs. This relation will be interpreted as "has a direct or indirect influence on". For simplicity we will assume that the relation R, hence also $\bar{R}$, has no circuits. This means that if A_i cites A_j then A_j may not cite A_i, also if A_i cites A_j and A_j cites A_k, then A_k may not cite A_i, and so on. We will also say that a publication has no direct influence on itself but still has an indirect influence on itself (with weight 1 for instance). This last assumption is only made for mathematical purposes. Under these assumptions $\bar{R}$ becomes a partial order (i.e. is reflexive, antisymmetric and transitive); R on the other hand is irreflexive, antisymmetric (even asymmetric) and intransitive. It is an incitence relation in the sense of *Zunde.*[7] The weighted incitence graph of our example is given by Fig. 2.

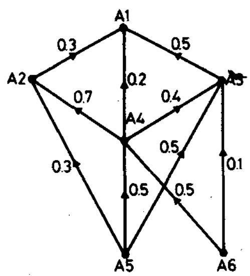

Fig. 2. Weighted incitence graph of Table 1

We consider now the matrices A and C, where the matrix entry a_{ij} is the weight of A_i in A_j (direct influence) and c_{ij} is the total influence weight of A_i on A_j (direct or indirect influence). The matrix A (associated with the relation R) is known, the matrix C (associated with $\bar{R}$) is the one we want to know. If $i \neq j$, then the total influence of A_i on A_j is the sum over all publications (A_k's) of the direct influence weight of A_i on A_k times the total influence of A_k on A_j (see Fig. 3). This gives $c_{ij} = \sum_{k=1}^{z} a_{ik}c_{kj}$ $(i \neq j)$. However, as we have agreed that A_i has an indirect influence on itself with weight 1, we need also that $c_{ii} = 1$. But if $i = j$ in the preceding formula then we always find that a_{ik} or c_{ki} is zero, for $a_{ik} = 0$ if A_i does not influence A_k directly and if A_i does influence A_k directly (so that $a_{ik} \neq 0$) then A_k can not influ-

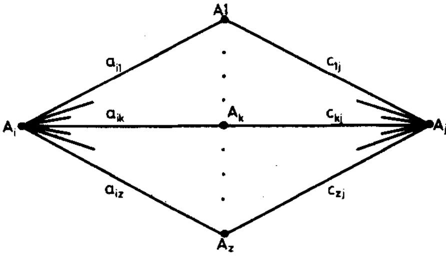

Fig. 3. Calculation of c_{ij}

ence A_i, neither directly nor indirectly as there are no circuits in R, so $c_{ki} = 0$. To correct for this we have to write:

$$c_{ij} = \sum_{k=1}^{z} a_{ik} c_{kj} + \delta_{ij},$$

where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$). Using matrix notation this formula becomes: $C = A \cdot C + I$, where I is the identity matrix. This yields: $C = (I - A)^{-1}$.

So, the Gozinto theorem says that the problem to find C has a solution if the matrix $(I - A)$ is invertible. It can be shown (see appendix) that this is always the case if the weighted incitence graph has no circuits and if we agree that a document has only an indirect influence on itself.

In our example the matrices A and C are:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0.7 & 0.4 & 0 & 0 & 0 \\ 0 & 0.3 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.1 & 0.5 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 0 \\ 0.61 & 0.7 & 0.4 & 1 & 0 & 0 \\ 0.645 & 0.65 & 0.7 & 0.5 & 1 & 0 \\ 0.355 & 0.35 & 0.3 & 0.5 & 0 & 1 \end{bmatrix}$$

which in this unrealistically simple case could also be inferred directly from the incitence graph. Remark that in descending order of influence we obtain: $A1, A5, A4, A3, A6, A2$. This shows that $A5$ is the paper with the greatest influence on $A1$. Note also that although we have formulated the problem in such a way that we only needed the first column of C, the other columns give further insight in the development of $A1$ and in the flow of ideas in these publications. Indeed, considering for instance the second column, we have also obtained the total influence the other elements of the graph exert upon $A2$.

An experiment

As an experiment we have studied the paper: T. S. BLYTH, J. C. VARLET, Fixed points in MS-algebras, *Bulletin de la Société Royale des Sciences de Liège*, 53 (1984) 3–8.

Weights are given according to the number of times a publication is cited in the paper under consideration and the place where the citation occurs. More precisely, we have given a weight 1 for every time a publication is cited in the introductory or preliminary part and a weight 3 for every time it is cited in the main part of the paper. A rationale for this choice is given by Susan *Bonzi*,[10] who found that "the number of times a work is cited in text may be an excellent predictor of relevance to the citing article". Moreover, she noticed that citations of the type "Several studies have dealt with. . ." (i.e. so-called perfunctory citations) tend to cluster at the beginning of an article. So we have valued these citations less than those in the main body of the article.

As an alternative we have followed a suggestion of *Jones*,[11] using the formula:

$$w = 1 - 0.9 \exp(-0.5\, x)$$

where w is the Jones-weight and x is the weight described in the previous paragraph. This reduces all weights to numbers between zero and one. In this way all total influence weights become comparable, which is not the case when using whole numbers as in the first way of assigning weights. There we can only compare papers of the same generation. However, as the problem of how to determine the weights is not the subject of this paper, we have chosen only these two reasonable ways of assigning weights, without studying the merits of possible alternatives. Probably more refined weighting schemes, perhaps based on citation context analysis (such as described by *Small*[12]) would give results that could be better interpreted.

Table 2 gives the set V of all publications under consideration: $A1$ is the *Blyth-Varlet* paper we are analyzing, $A2 - A5$ are the first generation papers, $A6 - A28$ are the second generation papers. The weighted incitence graph for the first way of assigning weights is given by Fig. 4. Table 3 and Table 4 give the associated matrices A and C. The total weights for the second method are given in Table 5. Recall that the total weights for the first method can be found in the first column of the associated matrix C.

Discussion of the results. The paper $A1$ by *Blyth* and *Varlet* is one in a series on *MS*-algebras; $A2$ is the first paper, in which the notion was introduced, $A3$ the

Table 2
Bibliographic data for the publications used in Fig. 4

A1	Blyth, T. S. and Varlet, J. C. "Fixed points in *MS*-algebras," Bulletin de la Société Royale des Sciences de Liège. 53: 3–8; 1984.
A2	Blyth, T. S. and Varlet, J. C. "On a common abstraction of de Morgan algebras and Stone algebras, "Proceedings of the Royal Society of Edinburgh. 94A: 301–308; 1983.
A3	Blyth, T. S. and Varlet, J. C. "Subvarieties of the class of *MS*-algebras," Proceedings of the Royal Society of Edinburgh. 95A: 157–169; 1983.
A4	Varlet, J. C. "Congruences on de Morgan algebras," Bulletin de la Société Royale des Sciences de Liège. 50: 331–342; 1981.
A5	Varlet, J. C. "Fixed points in finite de Morgan algebras," Discrete Mathematics. 53: 265–280; 1985.
A6	Balbes, R. and Dwinger, P. Distributive Lattices. University of Missouri Press; 1974.
A7	Berman, J. "Distributive lattices with an additional unary operation," Aequationes Mathematicae. 16: 165–171; 1977.
A8	Matsumoto, K. "On a lattice relating to the intuitionistic logic," Journal. Osaka Institute of Science and Technology. 2: 97–107; 1950.
A9	Varlet, J. "Fermetures multiplicatives," Bulletin de la Société Royale des Sciences de Liège. 38: 101–115; 1969.
A10	Berman, J. and Dwinger, P. "De Morgan algebras: free products and free algebras," preprint.
A11	Davey, B. A. "On the lattice of subvarieties," Houston Journal of Mathematics. 5: 183–192; 1979.
A12	Varlet, J. "On the greatest boolean and stonean decompositions of a p-algebra," Colloquid, Mathematica Societas János Bolyai. 29: 781–791; 1977.
A13	Anderson, A. R. and Belnap, N. D. Entailment: the Logic of Relevance and Necessity, Vol. 1. Princeton University Press; 1975.
A14	Bauer, H. and Kamara, M. "Priestley duality for distributive polarity lattices," preprint.
A15	Belnap, N. D. and Spencer, J. H. "Intensionally complemented distributive lattices," Portugaliae Mathematica. 25: 99–104; 1966.
A16	Białynicki-Birula, A. and Rasiowa, H. "On the representation of quasi-boolean algebras," Bulletin de l' Académie Polonaise des Sciences. V3: 259–261; 1957.
A17	Cornish, W. H. and Fowler, P. R. "Coproducts of Kleene algebras," Journal of the Australian Mathematical Society. 27: 209–220; 1979.
A18	Grätzer, G. General Lattice Theory. Basel: Birkhäuser; 1978.
A19	Kalman, J. "Lattices with involution," Transactions of the American Mathematical Society. 87: 485–491; 1958.
A20	Katrinak, T. "Essential and strong extensions of p-algebras," Bulletin de la Société Royale des Sciences de Liège. 49: 119–124; 1980.
A21	Rasiowa, H. An algebraic approach to non-classical logics. Amsterdam: North-Holland; 1974.
A22	Sankappanavar, H. P. "A characterization of principal congruences of de Morgan algebras and its applications," Mathematical Logic in Latin America, North-Holland: 341–349; 1980.
A23	Traczyck, T. "On the variety of bounded commutative BCK-algebras," Mathematica Japonica. 24: 283–292; 1979.
A24	Varlet, J. "A strenghtening of the notion of essential extension," Bulletin de la Société Royale des Sciences de Liège. 48: 440–445; 1979.
A25	Monjardet, B. "Eléments ipsoduaux du treillis distributif libre et familles de Sperner ipsotransversales," Journal of Combinatorial Theory. 19: 160–176; 1975.

Table 2 (cont.)

A26	Monteiro, A "Construction des algèbres de Nelson finies," Bulletin de l'Académie Polonaise des Sciences. 11: 359–362; 1963.
A27	Rivière, N. M. "Recursive formulas on free distributive lattices," Journal of Combinatorial Theory. 5: 229–234; 1968.
A28	Varlet, J. "Relative de Morgan lattices," Discrete Mathematics. 46: 207–209; 1983.

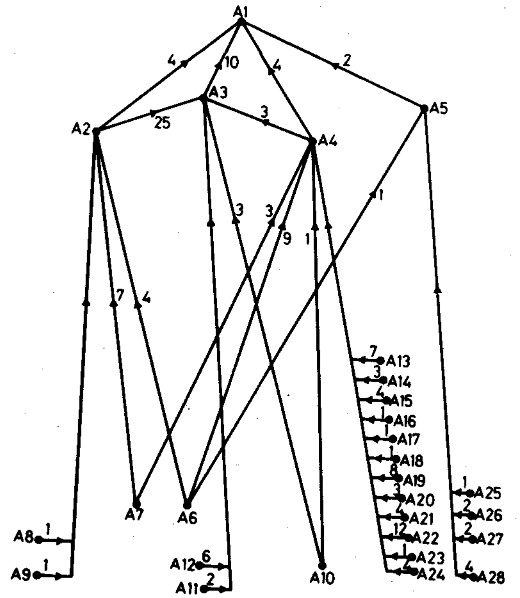

Fig. 4. Weighted incitence graph of the experiment

second. *MS*-algebras form a generalization of de Morgan algebras and Stone algebras, which both are special distributive lattices.

Considering the total weights of the first generation we see that $A2$ has the highest weight, before $A4$ and $A3$. So, although $A3$ has the greatest direct influence on $A1$, our analysis using the Gozinto theorem, rightly indicates that $A2$, being the paper where this new notion was introduced has a greater total influence (see also professor *Varlet*'s comments). $A4$ is a paper on de Morgan algebras which has served as an inspiration source to study analogous phenomena in the context of *MS*-algebras.

Going one step further, we see that $A7$, $A6$ and $A22$ have the greatest total weight among the second generation papers. $A6$ is an important book on distributive lattices; the authors use it as a reference for all basic facts on distributive lattices. Our analysis indicates that $A7$ also has a very great indirect influence on $A1$. This is

Table 3
The matrix A

A1	0	0	0	0	0
A2	4	0	25	0	0
A3	10	0	0	0	0
A4	4	0	3	0	0
A5	2	0	0	0	0
A6	0	4	0	9	1
A7	0	7	0	3	0
A8	0	1	0	0	0
A9	0	1	0	0	0
A10	0	0	3	1	0
A11	0	0	2	0	0
A12	0	0	6	0	0
A13	0	0	0	7	0
A14	0	0	0	3	0
A15	0	0	0	4	0
A16	0	0	0	1	0
A17	0	0	0	1	0
A18	0	0	0	1	0
A19	0	0	0	8	0
A20	0	0	0	3	0
A21	0	0	0	4	0
A22	0	0	0	12	0
A23	0	0	0	1	0
A24	0	0	0	4	0
A25	0	0	0	0	1
A26	0	0	0	0	2
A27	0	0	0	0	2
A28	0	0	0	0	4

All other matrix entries a_{ij} are 0.

correct: in their first paper on *MS*-algebras ($A2$), the authors write in a comment: "The idea of generalising de Morgan algebras and Stone algebras is not entirely a novelty. In fact, in $A7$, Berman considers. . ."

Finally, about $A22$, which directly influenced $A4$, the author writes: "In $A22$, Sankappanavar gives a characterization of the principal congruences of a de Morgan algebra. We intend to supplement his results. . ."

Using *Jones*-weights allows us to rank all the publications according to their total influence. $A6$ and $A7$ become the most influential documents, before $A2$, $A4$ and $A22$.

We have presented our results to Professor *Varlet* who gave the following comments: "Generally speaking, I think that your conclusions are correct. Moreover, it seems that the first method is better than the second. Without any doubt, $A7$ has been the catalyst of our series of papers. The papers $A2$ and $A3$ constitute the basis

Table 4
The matrix C

A1	1	0	0	0	0		
A2	254	1	25	0	0		
A3	10	0	1	0	0		
A4	34	0	3	1	0		
A5	2	0	0	0	1		all other matrix
A6	1324	4	127	9	1	1	entries c_{ij} are 0
A7	1880	7	184	3	0		
A8	254	1	25	0	0		
A9	254	1	25	0	0		
A10	64	0	6	1	0		
A11	20	0	2	0	0		
A12	60	0	6	0	0		
A13	268	0	21	7	0		
A14	102	0	9	3	0		
A15	136	0	12	4	0		
A16	34	0	3	1	0		
A17	34	0	3	1	0		
A18	34	0	3	1	0		
A19	272	0	24	8	0		
A20	102	0	9	3	0		
A21	136	0	12	4	0		
A22	408	0	36	12	0		
A23	34	0	3	1	0		
A24	136	0	12	4	0		
A25	2	0	0	0	1		
A26	4	0	0	0	2		
A27	4	0	0	0	2		
A28	8	0	0	0	4	1	

of our study and their knowledge is indispensable for a right comprehension of our papers." (Translated from the French.)

We may conclude that although weights were assigned in a rather ad hoc way, our method clearly singles out those publications that have an important underlying influence on the paper we have analyzed.

Final remarks and conclusions

In the examples of the preceding sections we started from one particular paper, but there is no compelling reason to do so. From a mathematical point of view the Gozinto theorem works equally well starting from several papers. The set V could

Table 5
Total influence weights
and rank using Jones-weights

A1	1	Rank
A2	1.872	3
A3	0.994	14
A4	1.673	4
A5	0.669	22
A6	3.604	1
A7	3.158	2
A8	0.850	16
A9	0.850	16
A10	1.554	8
A11	0.665	23
A12	0.949	15
A13	1.628	7
A14	1.337	12
A15	1.469	9
A16	0.760	18
A17	0.760	18
A18	0.760	18
A19	1.645	6
A20	1.337	12
A21	1.469	9
A22	1.669	5
A23	0.760	18
A24	1.469	9
A25	0.304	27
A26	0.448	25
A27	0.448	25
A28	0.587	24

also be formed starting "at the bottom", looking for publications that cite a particular paper, directly or indirectly.

The main topic of this paper is a mathematical technique to study influences, using citations. We will continue our investigations to study the problem of the determination of the weights. Should one give a weight equal to one, to every cited item, or give an equal share to every publication, or use yet another, more intricate formula, similar to the one we have used, or the one suggested by *Jones*.[11]

Another problem is the number of generations to consider. Probably two to four will be reasonable, especially considering the large matrices one has to handle. Of course, there is also the technical problem of the inversion of such matrices. However, the matrices we have to consider have a lot of zero entries. Such matrices are said to be sparse and there exist special techniques to handle sparse matrices (see f.i. *Tewarson*[13]).

A further problem is the exact meaning of the total influence weights one obtains. In this paper we have only considered their rank (or the rank within each generation), but it might be interesting to look for an intrinsic meaning. Finally, there is also the question of how to treat books, theses and other documents that have a large bibliography.

A biographical note on Z. Gozinto (*Staelens*[9])

The interested reader might wonder who that famous Italian mathematician actually was. However, no bibliographical tool will be able to help him for Z. Gozinto never existed. Indeed, during a lecture on the problem of parts listing (production scheduling) Andrew *Vazsonyi* let out roguishly that this problem had been studied by Z. Gozinto much earlier. Afterwards, George *Dantzing* (the famous inventor of the simplex method in linear programming) asked who Z. G. actually was. *Vazsonyi* answered with a straight face that the theorem he had just proved was indeed discovered by the "celebrated Italian mathematician Z. Gozinto". George *Dantzig*, not satisfied, asked what the initial Z. meant. "Well, Zepartzat, of course" replied *Vazsonyi* laconically. So, Zepartzat Gozinto, *Vazsonyi*'s mysterious brain-child was born!

Appendix

If the relation R (or equivalently, the weighted incitence graph) has no circuits then one can order the elements of V in such a way that $(A_i, A_j) \in R$ (i.e. A_i directly influences A_j), A_j comes before A_i. This is called topological sorting (cf. *Knuth*,[14] p. 258–265). Using this order for the rows and columns of the matrix A one obtains a lower triangular matrix with zeros on the diagonal. If one agrees upon giving a document an indirect influence weight equal to one with respect to itself, the matrix $I - A$ also becomes a lower triangular matrix with 1's on the diagonal. This shows that $I - A$ is invertible under these conditions. Remark however that in actual calculations it is not necessary to reorder the rows and columns to obtain this triangular form: changing rows and columns has no influence on the invertibility of the matrix. One may also remark that it is not necessary to chose the influence weight of a publication equal to one with respect to itself: any strict positive number will do.

*

I would like to thank Susan *Bonzi* and Donald *Kraft* for helpful observations.

References

1. M. M. KESSLER, An experimental study of bibliographic coupling between technical papers, *IEEE Transactions PTGIT*, IT-9 (1963) 49.
2. M. M. KESSLER, Bibliographic coupling between scientific papers, *American Documentation*, 14 (1963) 10.
3. B. H. WEINBERG, Bibliographic coupling: A review, *Information Storage and Retrieval*, 10 (1974) 189.
4. H. SMALL, Co-citation in the scientific literature: A new measure of the relationship between two documents, *Journal of the American Society for Information Science*, 24 (1973) 265.
5. I. MARSHAKOVA, System of document connections based on references, *Nauchno-Tekhnicheskaya Informatsiya*, Seriya II, (1973) 3.
6. I. MARSHAKOVA, Citation networks in information science, *Scientometrics*, 3 (1981) 13.
7. P. ZUNDE, Structural models of complex information sources, *Information Storage and Retrieval*, 7 (1971) 1.
8. A. VAZSONYI, *Scientific Programming in Business and Industry*, New York, John Wiley, 1958, Chapter 13.
9. H. STAELENS, Het Gozintoprobleem: Een mooie toepassing van matrixrekenen, *Wiskunde en Onderwijs*, 8 (1982) 307.
10. S. BONZI, Characteristics of a literature as predictors of relatedness between cited and citing works, *Journal of the American Society for Information Science*, 33 (1982) 208.
11. W. JONES, A fuzzy set characterization of interaction in scientific research, *Journal of the American Society for Information Science*, 27 (1976) 307.
12. H. SMALL, Citation context analysis, in: B. J. DERWIN, M. J. VOIGHT (Eds), *Progress in Communication Sciences*, Vol. 3, 1982, pp. 287–310.
13. R. TEWARSON, *Sparse Matrices*, New York, Academic Press, 1973.
14. D. KNUTH, *The Art of Computer Programming*, Vol. 1.: Fundamental Algorithms, Reading, Addison–Wesley, 1969.