Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 1534–1547
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Bibliometric maps of field of science

Irina Marshakova-Shaikevich

Institute of Philosophy RAN, 119842 Moscow, Russia
KW University, 85-090 Bydgoszcz, Poland

Received 27 September 2004; accepted 3 March 2005
Available online 6 July 2005

Abstract

The present paper is devoted to two directions in algorithmic classificatory procedures: the journal co-citation analysis as an example of citation networks and lexical analysis of keywords in the titles and texts. What is common to those approaches is the general idea of normalization of deviations of the observed data from the mathematical expectation. The application of the same formula leads to discovery of statistically significant links between objects (journals in one case, keywords—in the other). The results of the journal co-citation analysis are reflected in tables and map for field "*Women's Studies*" and for field "Information Science and Library Science". An experimental attempt at establishing textual links between words was carried out on two samples from SSCI Data base: (1) EDUCATION and (2) ETHICS. The EDUCATION file included 2180 documents (of which 751 had abstracts); the ETHICS file included 807 documents (289 abstracts). Some examples of the results of this pilot study are given in tabular form . The binary links between words discovered in this way may form triplets or other groups with more than two member words.

Keywords: Journal co-citation analysis; Lexical analysis of keywords; Network; JCR:SSE; SSCI

1. Introduction

Algorithmic classifications of sets of objects attracted researcher's attention in the1970s (Sparck Jones, 1971). It was hoped that classifications based on some algorithmic procedures on objects in a set would help to discover such properties of information flows that escape the possibilities of a priori logical classifications. Besides that algorithms promised greater flexibility as compared to rigid logical classifications, they might prove better suited to deal with ever-changing flow of information and thus a better tool to follow the dynamics of a science. Classificatory algorithms are usually applied to some easily identifiable entities in the texts, such as keywords, references to previous work, etc. The inductive character of those classifications

E-mail address: ishaikev@mail.ru

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.03.027

makes it difficult to predict the results of an empirical study; the fact may seem a curse or a blessing, depending on the tastes of the researcher. One important benefit of that approach is the exclusion of subjective factor in the process of classification: the results seem more 'natural' and 'objective'. It is believed that classificatory algorithms in diachronic studies may become a useful addition to synchronous classifications based on logical principles.

Various bibliometric methods fall into two major approaches. The first is based on the analysis of the dynamics of separate features—'plain bibliometrics'. The second is associated with the study of correlation between objects, their clustering and classification—'structural bibliometrics' (Marshakova-Shaikevich, 1996).

The second approach to the quantitative study of information has been gaining popularity since the early 1970s. It is aimed at getting structural (or qualitative) picture of state of science. The quantitative characteristics of information processes are not considered here as final results of the analysis, they are used as tools to describe the ever-changing structure of science in general and of particular fields of knowledge.

The present paper is devoted to two directions in algorithmic classificatory procedures: the journal co-citation analysis as an example of citation networks and lexical analysis of keywords in the titles and texts.

What is common to those approaches is the general idea of normalization of deviations of the observed data from the mathematical expectation. The application of the same formula leads to discovery of statistically significant links between objects (journals in one case and keywords in the other).

2. Citation networks

From the mathematical point of view a citation network is a set of documents with the relation of citing imposed on it. In other words it is a union of a set of citing papers and a set of cited papers. A citation network is a potential base for various classification of papers. It was M. Kessler who in 1962 formulated the concept of 'bibliographic coupling' as a measure of similarity of two documents based on the number of common references (Kessler, 1963). The logical opposite of bibliographic coupling is the concept of co-citation proposed in 1973 independently by H. Small in the USA and by the present author in the USSR (Small, 1973; Marshakova, 1973). The similarity of two documents depends on the number of papers citing both documents. When a new paper appears it is not linked to any other paper until it starts to be cited in scientific literature. This connection was called prospective by the author for it is based on citations in future literature. When applied to a vast bibliographic material (primarily to the SCI or SSCI databases) co-citational analysis serves as a means of getting clusters in the citation nets which can be interpreted as elements of a complex hierarchical structure of science—maps of science with very broad fields of knowledge at the top and many individual research fronts at the bottom.

Among numerous techniques covered by structural bibliometrics, the place of honor belongs to co-citation analysis of papers and corresponding mapping of science in ISI similarly.The journal co-citation analysis is aimed at monitoring of everchanging structure of science by means of algorithmic procedures.

The idea of co-citation may be concisely expressed by the following maxim: two papers are similar when they are simultaneously cited by other papers. Here (in case of the journal co-citation analysis) the maxim should be read as: two journal are similar when both of them are often cited in the same third journals. On the basis of JCR data a network of scientific journals is built. The proximity of a pair of journals is determined by their co-citation. The raw citation figures are normalized to exclude the overwhelming influence of most popular highly-cited journals. The journal to journal weighted links form an intricate network of a field of knowledge with possible clusters discernible in the map. The algorithmic stage should be followed by an informal analysis and possible verification by other approaches. This method was first tested on the field of Information Science and Library Science (Marshakova, 2003) and on the field of Women's Studies (Marshakova-Shaikevich, 2004).

Methodology of the journal co-citation analysis is given shortly below.

To facilitate the calculation the bottom part of the list of citing journals was excluded from the analysis. The cut-off point was determined as

$$0.01 \times (\text{number of total cites} - \text{number of self-cites}).$$

The minimal cut-off point is arbitrarily set at 2.

The part of the list above this threshold will form the core of citing journals for the journal under analysis (with self-citation excluded).

For each citing journal its citing capacity (CC) is calculated. Citing capacity of the given citing journal (CC) is defined as the sum of its frequencies in all the core lists of cited journals of the category.

Assuming (for time being) the null hypothesis of uniform distribution of cites in journals within the category, let us calculate the mathematical expectation of a journal (a) being cited by a citing journal (i). This mathematical expectation (m) will be defined as the product of the share of journal 'a' in the total cites of all journals (p_a) by the citing capacity (CC) of the citing journal:

$$m_{ai} = p_a \times \mathrm{CC}_i,$$

$$p = \sum C_a \Big/ \sum\sum C,$$

$\sum C_a$ is the total cites for a journal 'a'; $\sum\sum C$ is the sum of all total cites of journals.

The real frequency of journal 'i' citing journal 'a' may deviate from m. The magnitude of this deviation will be calculated as

$$S = (x - m - 1)/\sqrt{m},$$

where x is the actual number of times that journal 'a' was cited by journal 'i'.

The (-1) in the numerator is introduced to cut off unique experimental events which should not be used as evidence no matter how small is m.

The resulting value of co-citation relatedness of two journals (a and b) will be calculated according to

$$R_{ab} = n \times \sum \lg(S_{ai} \times S_{bi}),$$

where n is the number of common o positive deviations (with $S \geqslant 2$) from mathematical expectation.

The journal co-citation analysis was tested on the ISI Journal Citation Reports: 2001 Social Sciences Edition categories *Information Science and Library Science* (IS&LS) and *Women's Studies* (WS). The category 'IS&LS' in 2001 edition included 55 journals, differing widely in the number of articles (36 articles on average, but *The Scientist* claiming 249), in the number of cites (with an average figure of 252 and the record of 1916 of *Journal of the American Society for Information Science and Technology*), in impact factor (with an average 0.5 and the record 2.02 of *Journal of Documentation*). The total number of cites is 13850, each journal making a small portion (p) of this sum. The category 'WS' in 2001 edition included 25 journals, differing widely in the number of articles (30 articles on average, but '*J Women Health Gen—B*' claiming 88), in the number of cites (with an average figure of 289 and the record of 1877 of journal *Sex roles*), in impact factor (with an average 0.4 and the record 2.28 of *J Women Health*). The total number of cites is 7224, each journal making a small portion (p) of this sum.

The results of the journal co-citation analysis are reflected in Table 1 and Fig. 1(a) 'Map of Journal Co-citation' for field '*Women's Studies*' and Table 2(a), (b) and Table 3, and Fig. 2(a), (b) and Fig. 3 for field 'IS&LS'.

3. Discussion

The two fields under study differ considerably. The Women's Studies field includes 25 journals, their co-citation relatedness is shown in Fig. 1(a) and Table 1. It would be more difficult to map the 55 journals,

Table 1
The values of co-citation relatedness R of pair of journals in the field 'Women's Studies'

Rang	Journal (a)	Journal (b)	Co-citation relatedness R
1	Sex Roles	Psychol Women Quart	26.75
2	Psychol Women Quart	Fem Psychol	6.76
		Fem Stud	1.00
		J Gender Stud	1.40
		Women Ther	1.56
3	Signs	Gender Soc	4.66
		Women Stud Int Fo	4.82
		Fem Rev	3.16
		J Women Aging	1.48
		Eur J Women Stud	1.07
		J Gender Stud	0.95
		Feminist Stud	0.80
4	Gender Soc	J Women Aginf	1.30
		Feminist Rev	1.38
		J Gender Stud	0.95
		Women Stud Int Fo	0.90
		Feminist Stud	0.78
5	Women Stud Int Forum	J Women Aging	1.00
6	J Women Aging	Women Health	0.78
		Women Health Iss	1.25
7	Feminist Rev	Feministt Stud	4.00
		J Gender Stud	1.38
		Eur J Women Stud	1.51
		Soc Polit	1.07
8	Feminist Stud	J Gender Stud	3.56
		Eur J Women Stud	0.90
		Soc Polit	1.3
9	Women Health	J Women Health	1.54
		Women Health Iss	1.43
10	J Gender Stud	Women Ther	1.54
		Eur J Women Stud	1.07
11	J Women Health	Women Health Iss	1.48
		J Women Health Gen-B	1.00
12	Women Health Iss	J Women Health Gen-B	1.07

belonging to 'Information Science and Library Science' field, here the results are given in tabular form which demands some spatial imagination on the part of the reader (for graphic representation see Marshakova, 2003, and Fig. 2(a), (b) and Fig. 3 where main co-citation links between journals are shown).

3.1. Women's studies field

Seventeen journals of the field under study of the 25 journals belonging to Women's Studies field showed positive values of R. The corresponding links are presented in Fig. 1(a).

Three clusters are clearly seen in this map. This result is quite meaningful which is demonstrated by the analysis of keywords (supplied by the authors or editor). For each journal a frequency list of keywords was

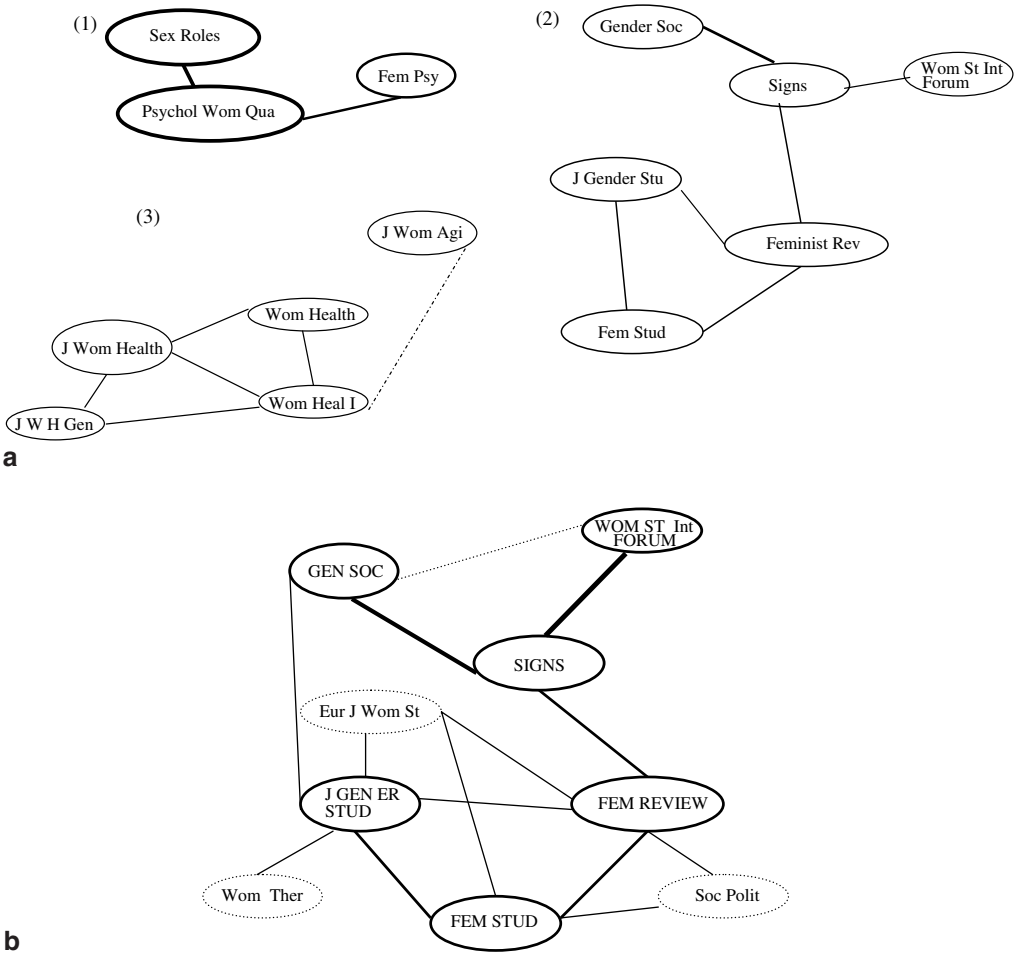

Fig. 1. (a) Map of journal co-citation (field 'Women's studies'). (b) Cluster (2): Sociology & Politics concerning women.

formed. Common keywords of a pair of journals (or of all journals of the cluster) show their thematic similarity. Let us turn to the three clusters in Fig. 1(a).

Cluster 1 includes three journals (*Sex Roles*, *Psychology of Women Quarterly* and *Feminism and Psychology*) with all R values exceeding 6. For the first two journals R is 26. The thematic coherence of this cluster is reflected in the list of keywords: *gender*, *girls*, *identify*, *knowledge*, *men*, *mental health*, *personality*, *psychology*, *rape*, *sex*, *students*, *women*.

The first pair of journals have many additional keywords: *dissatisfaction*, *psychology*, *stress*, *stereotypes*, *aggression*, *experiences*, *self-esteem*, *weight*. All these keywords belong to the central thematic component of the cluster, i.e., PSYCHOLOGY. The same conclusion is supported by keywords denoting methodological concept of psychological research: *college students*, *inventory*, *model*, *objectification*, *reliability*, *responses*, *scale*, *students*, *validation*, *validity*.

There are six journals in cluster 2: *Gender and Society*, *Signs*, *Women Studies International Forum*, *Feminist Review*, *Journal of Gender Studies*, *Feminist Studies*, but R values are low. Common keywords are few, nevertheless we find here: *gender*, *labor*, *movement*, *population*, *power*, *war*, *women*, *work*. All those words show the thematic center of the cluster: SOCIOLOGY and POLITICS concerning women. A few more

Table 2
Co-citation relatedness of some journals in 'Information and Library Science'

		1	2	3	4	5	6	7	8	9	10
(a)											
1	Annu Rev Inform Sci	+	36	35	86	41	48	18	4	–	–
2	Inform Process Manag		+	40	68	38	19	11	4	6	7
3	J Am Soc Inf Sci Tech			+	115	78	41	11	6	–	7
4	J Doc				+	78	57	59	12	6	17
5	J Inform Sci					+	27	25	15	5	12
6	Libr Quart						+	60	16	5	6
7	Libr Trends							+	24	4	6
8	LIBRI								+	–	–
9	P ASIS Annu Meet									+	6
10	Scientometrics										+

		1	2	3	4	5	6	7	8	9	10	11	12
(b)													
1	ASLIB Proc	+	23	3	–	3	2	–	4	8	2	12	6
2	Coll Res Libr		+	29	17	118	4	19	29	33	13	70	39
3	Electr Libr			+	–	7	–	2	4	1	–	18	20
4	Inform Tech Li				+	19	10	6	17	4	4	4	27
5	J Acad Libr					+	–	22	23	32	23	61	51
6	Libr Inf Sci						+	4	8	10	3	26	10
7	Libr Inform Sci Res							+	6	22	–	18	–
8	Libr J								+	12	3	45	30
9	Libr Quart									+	1	60	57
10	Libr Resour Tech Ser										+	14	8
11	Libr Trends											+	57
12	Online												+

Table 3
Co-citation relatedness of some journals in 'Information Systems and Information Management'

		1	2	3	4	5	6	7	8
1	Inform Manag—Ams	+	6	10	20	–	53	14	246
2	Inform Soc		+	4	7	4	15	3	–
3	Inform Syst J			+	28	25	37	–	54
4	Inform Syst Res				+	38	80	34	378
5	Int J Inform Manage					+	–	–	66
6	J Inform Technol						+	4	119
7	J Manage Inform Syst							+	28
8	MIS QUART								+

journals may be added to the cluster (with R values < 2): *European Journal of Women's Study*, *Social Politics*, *Women & Therapy*. Those additional journals support the same thematic conclusion (Fig. 1(b)).

Cluster 3 includes five journals: *Women & Health*, *Journal of Women Health & Gender-Based*, *Women Health Issues*, *Journal of Women Aging*, *Journal of Women Health*; all devoted to the theme WOMEN's HEALTH. There are 145 keywords, common at least to one of pair of journals. They are often medical terms: *Alzheimer's disease*, *blood pressure*, *breast cancer*, *cancer*, *coronary heart diseases*, *disorders*, *epidemiology*, *estrogen replace therapy*, *hip fracture*, *HIV infection*, *hormone replace therapy*, *intervention*,

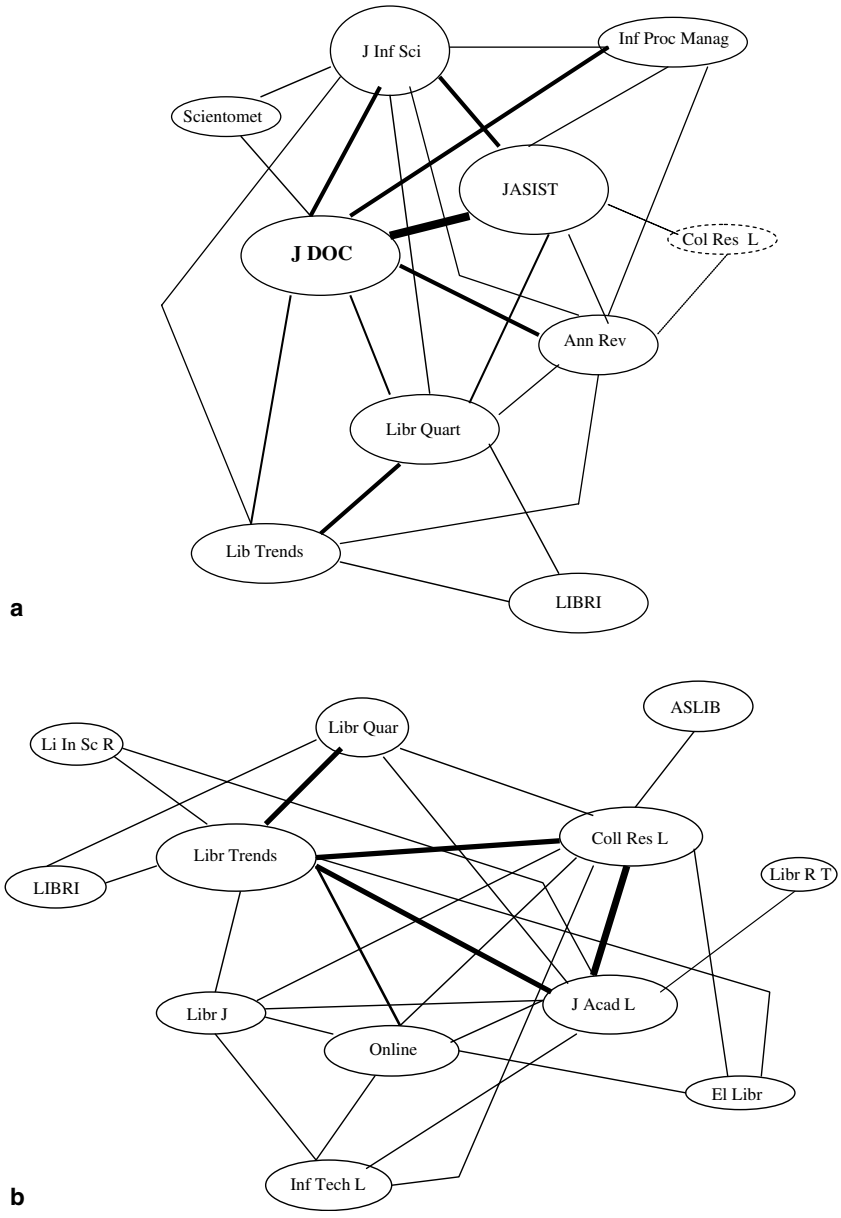

Fig. 2. Co-citation relatedness of some journals in 'Information and Library Science'.

mammography, *menopause*, *osteoporosis*, *risk*, *risk factor (s)*. Demographic terms are reflection of standard experimental research procedure: *adolescent*, *African–American*, *age*, *black*, *education*, *employment*, *older women*, *population, race*. Keywords denoting behavior traits are also conspicuous here, e.g., *abuse*, *assault*, *cocaine*, *contraception*, *domestic violence*, *exercise*, *homeless*, *pregnancy*, *smokers*, *smoking*, *victims*.

However, it should be kept in mind that there are a few keywords common to all three clusters (*gender*, *women*, United States) or to the first and third clusters (*attitudes*, *behavior*, *depression*, *inventory*, *knowledge*, *men*, *risk factor*, *sex*, *stress*). The keyword *work* is common to the second and third clusters.

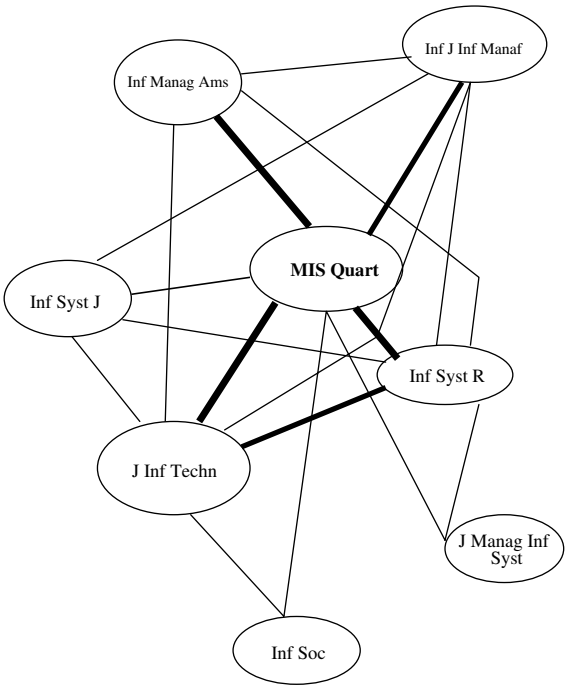

Fig. 3. Co-citation relatedness of some journals in 'Information Systems and Information Management'.

3.2. Information and library science field

Table 3 and Fig. 3 clearly demonstrate the existence of a separate cluster of eight journals, in no way connected with the rest of our category. Among the citing journals of this cluster, many more journals can be found which are not part of JCR: Social Science Edition 'Information Science and Library Science' category, but which appear on the list of JCR: Science Edition *Computer Science, Information Systems* category. These are: *Decis Support Sys*, *Eur J Inform Syst*, *Inform Software Tech*, *Internet Res*, *J Comput Inform Syst*, *Wirschaftsing*. This list may be extended by many journals in Social Science outside our category (Information and Library Science): *Comput Educ*, *Comput Hum Behav*, *Int J Electr Co*, *Int J Hum-Comput*, *Int J Oper Prod M, IEEE T Eng. Manage*, *J Comput Assist L*, *Leadership Quart*, *J Bus Res*, *Omega- Int J Manag*, *Organ Sci*.

As to Table 2(a) and (b) (and Fig. 2(a) and (b)) their analysis gives an opportunity to discuss some general problems. Researchers in scientometrics (or bibliometrics) usually see a map of discrete set with subsets in them as the final result of their study. Failure to attain such a picture seems a disappointment. However, another alternative to thinking in customary cluster terms may be suggested. Table 2(a) and (b) (and Fig. 2(a) and (b)) evoke a metaphor of space: the two tentative clusters, represented here, are not separated by any barrier—they are bridged by *Libr Quart, Libr Trends*, (and to a certain degree—LIBRI, *Coll Res Libr* and *ASLIB Proc*).

At the same time, the opposite poles: e.g., *Libr J*, *Online*, *Electr Libr* on the one hand, and *JASIST*, *J Inform Sci*, *Annu Rev Inform Sci*, *Inf Proc Man* on the other are not mutually connected anywhere. The journals in the upper part of the combined picture (Fig. 2) are mostly preoccupied with theoretical problems, while the journals at the bottom part more often try to solve practical problems. Nevertheless, the

composite character of this large combined cluster is evident. This conclusion is supported by journals with low *R* values, which can be seen at the periphery of Fig. 2. *Scientometrics*, *Proc ASIS Annu Meet* (to a lesser extent *Can J Inform Libr Sci*) are in the upper part of the picture. *Database*, *Interlend Doc*, *Supply*, *Libr Collect Acquis*, *Online Inform Rev*, *J Scholarly Publ*, *Knowl Organ* evidently belong to Library Science (the bottom part of the picture).

The ISI category of Information and Library Science comprises Social Science journals of which 30 are represented in Table 2(a), (b) and Table 3. Some of the remaining journals (e.g., *Econtent*, *J inform Ethics*, *NFD Inform-Wiss Prax*, *Z Bibl Bibl*) have so few citations to their credit that it is virtually impossible to place them anywhere on the map. An extension of the study in time-depth would probably make it possible to locate them with some precision.

Much more interesting are those journals of the group that are extensively cited, but still fail to show any *R* values. They are *Int J Geogr Inf Sci* (with 531 cites), *Scientist* (462), *Telecommun. Policy* (247), *Soc. Sci. Inform.* (193), *Restaurator* (100), *J Health Commun* (91), *Law Libr J* (89), *Soc Sci Comput Rev* (86), *J Am Med Inform Ass* (second best in popularity, with 1658 cites) should also be included in this group, for it has only one weak connection with *B Med Libr Assoc* (387), which has three very weak links with IS&LS journals. All these journals are obvious outsiders. Their inclusion in the category *Information and Library Science* under study is quite mistaken, e.g., the core list of *Int J Geogr Inf Sci* has 29 journals, of which only one journal (*Lect. Notes Comput.*) has some connection with our category, but it did not demonstrate any *S*-values. All these journals should be relegated to their corresponding fields of knowledge (geography, health studies, law, etc.).

The method of calculation of co-citation relatedness described in the present paper to some extent depends on the initial composition of the category to be analyzed. One may wonder what would happen to resulting picture, if one starts with a somewhat different initial composition. Let us presume that 10 journals, just discussed, were actually deleted from the category Information and Library Science. The total sum of CC would diminish by 8%, and the values of *p* would grow accordingly. This change would slightly lower co-citation relatedness *R*-values, without significant changes in the configuration of links.

If the initial group is expanded considerably, one may expect greater changes in configuration. The JCR: Science Edition category of Computer Science, Information Systems may serve as an example. As a pilot test the Social Science category was enlarged by eight journals with the highest impact rating from JCR: Science Edition. These were: (1) *IEEE Netw* (with 657 cites), (2) *IEEE ACM T NETWORK* (2150), (3) *ACMT Inform Syst* (470), (4) *IEEE Pers Commun* (827), (5) *IEEE T Inform Theory* (1635), (6) *Inform Syst* (1062), (7) *Med Inform Internet* (363), (8) *J Chem Inf Comp Sci* (3561). The growth of overall cites to 30,500 diminished the *p* values of the former journal members with the corresponding growth of *S*- and *R*- values. It seems that additional journal members will add new separate clusters, rather than change configuration of clusters in former Social Science category. The first two journals, for example, have a record *R* of 415 between them and are strongly connected to the fourth journal ($R = 67$ and $R = 45$, respectively). This finding gives support to the supposition that good clusters will keep stable even under changes in initial composition of a group.

4. The analysis of keywords in the titles and texts

The research of statistical dictionary of keywords showed that change in the frequency of keywords may be due to various factors. Sometimes it is a purely linguistic process, for example, a competition of terms, such as ‘optical generator of stimulated radiation’, ‘generator of stimulated radiation’, ‘optical quantum generation’. All these terms were very frequent in 1961–1963 (see The Soviet ‘Physics’ abstracts journal, division of ‘Generators of stimulated radiation’). At the same time, English abbreviations LASER and

MASER entered the field. The period of 1964–1965 was marked by lexical stabilization when only two absolute synonyms LASER and OKG (optical quantum generator) struggled for survival. The victory of 'laser' was due to its compactness and possibility to form an adjective 'lasernyi' (which is important in scientific Russian) (see Marshakova, 1974b).

The technique of statistical analysis of keywords might be aimed at discovering new directions of science. In this sense it might prove important for monitoring development of science and for evaluation of science and programs. Many of such directions were manifested in Research Fronts on co-citation maps of science. Thus, analysis of keywords may be applied to any subject indexes of databases.

Possible objectives of analysis of keywords in the titles and texts may be classed in two main directions.

1. Document clustering on the basis of lexical similarity of titles.
2. Word clustering on the basis of co-occurrence of words in the same documents.

The details of the procedure of clustering documents on lexical similarity were given in the paper (Marshakova, 1974a). Only the second task will be considered here.

Late years witnessed the efforts of many researchers trying to develop methods of analysis of keywords in the titles and texts of scientific publications (Amudhavalli & Raghavan, 1995; Kopsa & Schiebel, 1995; Leydesdorf, 1997; Marshakova-Shaikevich, 2001; Noyons & Van Raan, 1995; Zitt & Bassecoulard, 1994) and many others.

Clustering based on co-occurrence of words in the documents is a formal technique for construction of lexical maps (this approach may be termed lexical bibliometric classification). Lexical maps may facilitate general orientation in the conceptual framework of science or in a given field of science. To solve this large and complex task, it seems necessary (1) to identify the list of lexical units used in the language of the given field of science and (2) to identify systemic relationships between these units. As a practical step in this direction, a study of secondary scientific literature was carried out, here information is represented in a more compact form and lexical redundancy of the primary text has been eliminated appreciably.

The possibility of an accurate calculation of frequency values for all terms makes it easier to create a mathematical model for statistically significant character of co-occurrence of terms.

The application of lexical bibliometric methods in building up the terminological system provides for the solution of two tasks: (1) to follow the quantitative dynamics of groups of terms in some scientific field and (2) to discover specific links of terms both in static and dynamic states. From the development of science viewpoint, the general task can be formulated as producing the terminological map of science at the given time-point. Such terminology maps may help to solve two problems: monitoring dynamics of scientific terminology and to visualize the conceptual links between terms and accordingly between science fields.

4.1. Pilot study

An experimental attempt at establishing textual links between words was carried out on two samples from SSCI data base: (1) EDUCATION and (2) ETHICS. The EDUCATION file included 2180 documents (of which 751 had abstracts); the ETHICS file included 807 documents (289 abstracts). The present paper deals mostly with the latter sample.

Some linguistic remarks:

Lexical analysis of a text corpus often calls for some pre-study editing.

The depth of pre-editing depends on the linguistic typology of the language under study. Slavic languages demand extensive morphological procedures (lemmatisation). In the case of English, morphological treatment is minimal.

Here the following steps were taken:

1. Some mistakes due to translation effect were corrected (e.g., sometimes *educational* was changed to a more idiomatic *education*).
2. With the help of some program tools in conjunction with manual editing the most frequent and least informative grammar words were excluded from the analysis (e.g., *the*, *has*, *and*, *should*, *by*, etc.).
3. New lexical items were created in the text. On the basis of a frequency dictionary of binary word combinations, frequent collocations were changed into new lexical units (business ethics would become business-ethics, decision making would be turned to decision-making, etc.). The pre-editing stage was followed by a recalculation of word frequencies.

Formal discovery of textual links between lexical units to a great degree depends upon the size of the fragment (or 'window'), within which the co-occurrence of two words is studied, and the corresponding mathematical calculations are made. On every stage of the analysis, the factual co-occurrence is compared to mathematical expectation, calculated on the basis of some null hypothesis. If the size of the fragment is minimal (i.e., one adjacent word), the results would be mostly grammatical and phraseological (collocational). With the growth of the size of the 'window', through which we look at co-occuring words, the information obtained would become more semantically general and thematic. In the present study, the length of the fragment was 50 words, i.e., the corpus was divided into 50 word fragments.

The calculation of statistical significance of co-occurrence figures was made according to the same formula that was used for journal co-citation:

$$S = (X_{ij} - m - 1)/\sqrt{m},$$

X_{ij}, real frequency of co-occurrence of words i and j; m, mathematical expectation of co-occurrence of words i and j

$$m = (F_i \cdot F_j)/N,$$

F_i, frequency of word i in the corpus; F_j, frequency of word j in the corpus; and N, number of fragments in the corpus (EDUCATION file $N = 2478$, ETHICS file $N = 883$).

The -1, introduced in the numerator, cuts off the cases of single co-occurrence of two words.

4.2. Results

Some examples of the results of our pilot study are given below in tabular form. They include six lexical units with their links and corresponding S-values:

ETHICS file

	X_{ij}	F_j	S
Adolescent F_i = 10			
child	9	13	22
forensic	3	10	6
psychiatry	9	12	23
Bioethics F_i = 46			
bedside	7	10	8
catholic	5	8	6

(*continued*)

	X_{ij}	F_j	S
church	3	4	4
theological	4	7	5
Catholic $F_i = 8$			
bioethics	5	46	6
church	2	4	5
Quebec	2	3	6
theological	4	7	5
Drugs $F_i = 9$			
control	4	29	5
disease	2	9	3
drug	3	18	4
policy	3	31	3
vaccine	2	4	5
war	2	5	4
Drug $F_i = 18$			
abuse	4	9	7
control	6	29	6
drugs	3	9	4
Nigeria	3	4	7
problem	4	21	4
reduction	2	3	4
solution	3	6	6
trafficking	2	3	4
trials	3	4	7
vaccine	3	4	7
HIV $F_i = 21$			
acquired	2	4	3
aids	12	31	13
contact	2	3	3
immunodeficiency	2	3	3
infection	8	10	15
patients	7	34	6
risk	3	15	3
testing	3	14	3
vaccine	2	4	3

The binary links between words discovered in this way may form triplets or other groups with more than two member words. Here are some examples of the resulting clusters (Fig. 4).

The scale of this pilot study was rather limited, some of the clusters listed depend on a single text in the database. But most links are so clear semantically, they are shure to hold in a larger database. The corresponding clusters would form parts of an intricate semantic networks of words.

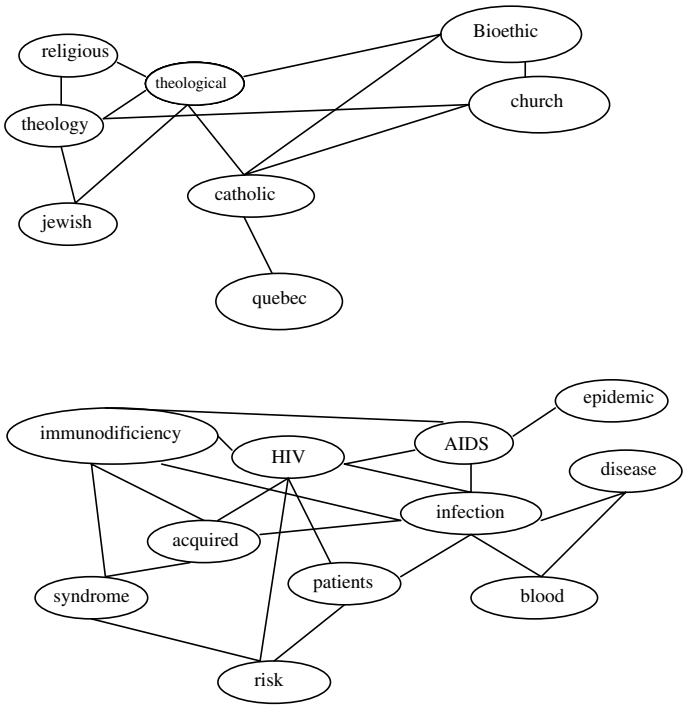

Fig. 4. Two fragments of semantic networks of keywords.

5. Conclusion

(1) The journal co-citation analysis, carried out in this study, discovered three clusters of journals within the JCR:SSE category of Women's Studies. Co-citation procedure provide its greater differentiating as compared to use of keywords. To be sure the keywords *gender* and *women* may be considered characteristic of the field as a whole. On the other hand, most keywords of singe clusters would unite them with psychology, sociology or medicine. Perhaps, a better solution for a thematic cluster would be to use a group of keywords, rather than single terms.

The journal co-citation analysis, carried out in study of Information and Library Science, discovered two clusters of journals within that category. The lesser of the clusters (Information Systems and Management) has consistent inner links, the greater cluster show some characteristics of a space. The well-cited journals, which did not show links to those clusters (Int J Geogr Inf Sci, Scientist, Telecommun. Policy, Law Libr and others), turned out to be outright mistakes and should be excluded from the category of Information and Library Science.

(2) While constructing lexical maps of science, one encounters many obstacles. Perhaps, the most important prerequisite is a frequency dictionary of the texts of abstracts of a certain field of science.

Back in the 1960–1970s there was a widespread growth of statistical lexica and databases for information purposes. SCI / Permuterm Subject Index contains more than 7 million pairs of keywords, an American statistical dictionary of texts for schools is built on a corpus of 5 million words [Carroll J.B. e.a. WORD FREQUENCY BOOK., Boston, 1971], the record corpus of French literary texts exceeds 70 million words [DICTIONNAIRE DES FREQUENCES Paris, 1971] After that the interest for statistical dictionaries declined considerably. While electronic corpora of texts (including secondary information literature) kept on growing, there were no statistical dictionaries for such corpora. This situation is due to several factors, one

of them should be mentioned. The statistical software is convenient and effective dealing with graphic words (chains of symbols between spaces). It is exactly that kind of information that made possible relatively large frequency dictionaries (such as Carrol's) in the 1960s. However, you need more than that to build a lexical map: your graphic words should be lemmatized, which is not easy to do in a fully automatic manner. The relative share of manual processing is declining, but it still grows in absolute figures with the astounding growth of corpora to be treated.

The results of the present pilot study clearly showed that formal quantitative analysis of large collections of secondary documents may discover specific (semantic) links between words. This pilot study was only a first step in an attempt to build a semantic map of a field of science.

References

Amudhavalli, A., & Raghavan, K. S. (1995). Co-word analysis of literature on information retrieval. In *Proceedings of the fifth biennial conference of international society for scientometrics and informetrics*. River Forest, IL, USA, June 1995.

Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.

Kopsa, A., & Schiebel, E. (1995). Science and technology mapping: a new interaction model for representing multidimensional relationships. Paper presented at the *Fourth international conference on science and technology indicators*. Antwerp, Belgium, October 5–7, 1995.

Leydesdorf, L. (1997). Co-words and citation relations between documents sets and environments. In *Proceedings of the first international conference on bibliometrics and theoretical aspects of information retrieval*. Diepenbeek, Belgium, August 24–28, 1997.

Marshakova, I. (1973). System of documentation connections based on references (SCI). *Nauchno-Tekhnicheskaya Informatsiya Seriya 2, 6*, 3–8.

Marshakova, I. (1974a). Classification of documents on the basis of keywords. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2, 5*, 3–10.

Marshakova, I. (1974b). The study of frequency dictionary of keywords. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2, 11*, 7–13.

Marshakova, I. V. (2003). Journal co-citation analysis in the field of information science and library science. In P. Nowak & M. Gorny (Eds.), *Language, information and communication studies* (pp. 87–96). Poznan: Adam Mieckiewicz University.

Marshakova-Shaikevich, I. (1996). The standard impact factor as an evaluation tool of science fields and scientific journals. *Scientometrics, 35*(2), 283–290.

Marshakova-Shaikevich, I. (2001). Scientometric perspectives of the analysis of chemical terminology. *Scientometrics, 52*(2), 323–336.

Marshakova-Shaikevich, I. (2004). Journal co-citation analysis in the field of women's studies. In H. Kretshmer, Y. Singh, & R. Kundra (Eds.), *International workshop on webometrics, informetrics and scientometrics* (pp. 247–259). Roorkee, March 2–5, 2004.

Noyons, Ed., & Van Raan. (1995). *Mapping the development of neural network research*. Report to the German Federal Ministry for Education and Science (BMBF), Report CWTS 95-06, May 1995.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of American Society of Information Science, 24*, 256–269.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics, 30*(1), 333–351.