Information Services & Use 10 (1990) 135–145
Elsevier – Printed in The Netherlands

Advanced interfaces to analyse automatically online database set of answers

Henri Dou, Parina Hassanaly, Albert La Tela and Luc Quoniam *
Centre de Recherche Rétrospective de Marseille, UA126, 13397 Marseille cédex 13, France

The use of databases which contain an increasing number of references leads to a new concept in online retrieval. Because the number of answers increases with time, it is in many cases too important to be analysed by simple reading. Various hosts have introduced some sorting techniques on one reference field such as GET, ZOOM, MEM, but these techniques are not powerful enough to give to the reader a fast overview of the content of all the answers. This is the reason which prompts the CRRM to develop various analytical techniques to perform an automatic analysis of the answers obtained by the use of online databases. These techniques, co-word techniques and typological aspects of the information, will certainly be part of those which will be offered to the online user in the near future.

1. Introduction

Most host computers today offer numerous technical databases covering all aspects from applied to fundamental research. For this reason, most searches require several databases and very often, even from a single database a large amount of information is retrieved. The increasing number of references, often several million each year, results in the dispersion of literature among a large number of sources making it very difficult for the end-user to quickly identify the relevant key information [1]. Today, there is a ***paradox*** since very powerful tools, databases and softwares, are being used to restrict the answers of a system so that the end-user receives a manageable amount of information [2]. This common practice leads to an important and, sometimes vital, loss of information, since most of the innovations and improvements are often located in peripheral areas and not in the core references of the subject [3].

* We would like to thank Orbit Information Technologies for providing access to various database files.

0167-5265/90/$03.50

1.1. Classical uses of databases

The retrieval of a set of hits in a database generally results from the combination of several search terms with boolean operators [4]. These terms can be related to the Basic Index, or to any other inverted fields of the database. But, with the wide increase of reference numbers, the same search repeated at various periods of time leads to an increasing number of hits. Moreover, most general queries are now relevant to more than one database, and this leads again to a wide increase in the number of answers [5].

At the same time, the structure of research in laboratories and the process of decision making has not changed too much. Most of the references from a bibliography will end on the desk of people who will have to read them and select the most pertinent for further analysis. But our reading capacity, even if we are well trained, has not increased to the same extend as the amount of references we have to read.

Thus the most powerful software and databases are increasingly used not to retrieve all the possible references dealing more or less closely with a subject, but instead, ***to limit*** the amount of references to the reading capacity of one person. This ***paradox*** must be overcome, especially since most of the innovations remain at the fringe of disciplines, and not in the core of a bibliography.

The only way to solve the problem is to deal with information retrieval in a different way [6,7]. Several hosts have now included in their software some commands which allow the beginning of a statistical treatment of a set of answers. For instance: GET (Infoline) [8], ZOOM (ESA) [9], MEM (Télésystèmes) [10], PRINT SELECT (ORBIT) [11] are among the best, but they do not link the statistical results with the reference number of the indexed papers in the bibliography.

2. Link analysis, Datalink software

2.1. Principle and method

If words A, B, C, D are present in the same field, they will generate the pairs AB, AC, AD, BC, BD, CD. This mean that 4 items (ABCD) generate 6 new entities (e.g. the pairs). Each of these pairs has its own peculiar meaning, and for an expert of the science of the technical area analysed, the meaning of each pair of often more significant and precise than the meaning of each word taken separately.

In other words, this means that there is a physical link between the two items because they are present in the same field [12]. In our case we only considered a bibliographic reference, the abstracts being excluded.

This method allows us to:

- analyse the content of a set of documents giving word frequencies and the frequencies of word pairs,

- select all the references containing a significant pair of words,
- represent the existing networks between references.

This method may be criticized because the link between words may have a different strength or may be unstable over time [13]. These assumptions are true if we consider full text documents. Our data is not drawn from a natural language base but rather from indexed material.

What we describe are the links which exist between entities used in the same field. For time analysis, we use codes, such as the Inspec classification which remains constant from a period to another.

2.2. Description of Datalink

Datalink deals with the references in the following way: first the field to be analysed is extracted from the bibliography. Then all the items of the field are treated to eliminate stop words, replace some of them by synonyms, truncate them when necessary, eliminate plurals, etc. When this is done, all the items (entities between two separators), are coded (hash coding [5]), and then all the pairs of items present in the same field are also coded using the same method. When the coding process is finished, we can calculate the:

- occurrences of word pairs,
- draw the network of the references over or under a certain threshold
- retrieve all the references which contain a given word, pair or triplet (this is possible since all the words are coded),
- transfer the data in a suitable format for multivariate analysis.

This allows us to browse the bibliography to find new ideas, since words or pairs of words may be ranked or seen in alphabetical order, and directly related to the reference numbers.

2.3. *Examples of results*

We have used Datalink to analyse 620 INSPEC (Physics Abstracts) references concerning the following query:

ALL DATABASE# AND FRANCE/OS FROM 1977 TO DATE.

2.4. Example of references

AN – C88008572
TI – Couplage entre mecanismes de déduction et base de données (Coupling between deduction mechanisms and databases)
AU – Ellul, A.
OS – Bull S.A., Direction Sci. Groupe, Les Clayes-Sous-Bois, France
SO – RPT. NO. DSG/CRG/87021, 20 PP., June 1987, 14 REF.
DT – R (REPORT)
NU – DSG/CRG/87021

LA – French
CC – *C6160D; C6170
TC – PR (PRACTICAL)
IT – logic programming; relational databases
ST – deductive databases; Bull; database management system; deduction tools; logic programming language; Prolog; tuple oriented strategy; Alexander method; set-oriented strategy; relational DBMS; Oracle.

These references have been downloaded, and several fields have been analysed: the INSPEC code field (CC), the Index Term field (IT) and the Source field (SO).

The Code Field will be used in section 3.2. All the words of the Index Term Field have been coded, and the pairs of words analysed. This analysis allows a fast browsing of the downloaded references, showing for instance that in the field of database management about 90 papers have been produced in about 10 years in France. Other areas such as applications in database management, environment, geophysics nuclear devices, etc. are represented in the answers.

For clarity of the results, we will only develop in detail the multivariate (factorial analysis) analysis of the Index Term Field.

2.5. Factorial analysis

In many cases, we would like to be able to carry out a theme analysis. This is possible as soon as a matrix containing co-occurrence frequencies is built. The

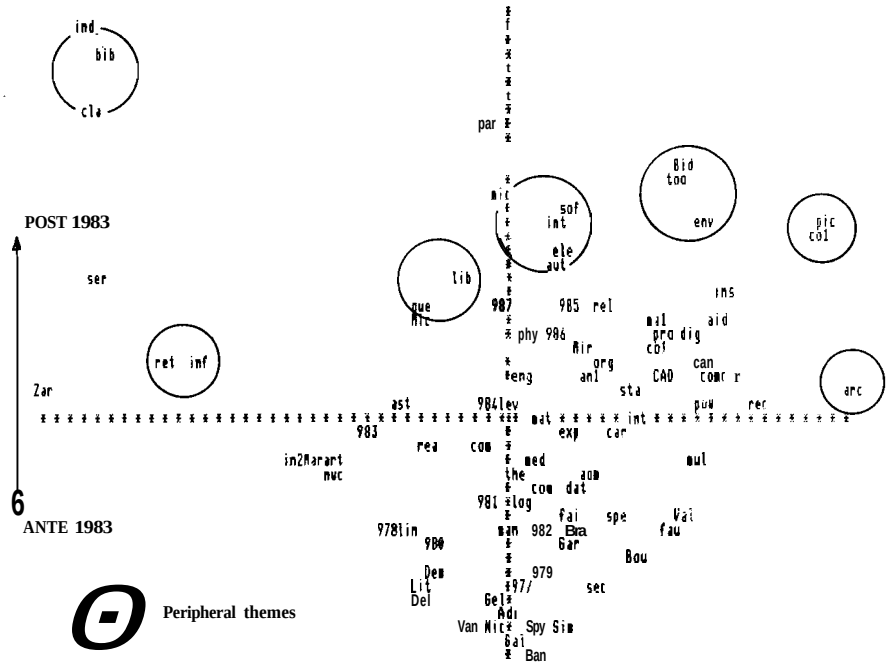


Figure 1. Simultaneous representation of rows and columns issued from factorial analysis.

words corresponding to themes in the data are used as column or row entries to build the matrix. We have developed a module which will automatically generate this matrix on the basis of an expert's choice of the theme he or she wishes to analyse.

In the present example, the themes, chosen for analysis by an expert, have been placed in the columns of the matrix (see the list of variables in Table 1), while the

Table 1
Words considered as variables with the abbreviated form used for drawing

(a) List of hidden points

Obvious point	Effective hidden points	List
inf	1	agr
car	1	fuz
Lit	1	Rol
dat	1	str
fau	1	tol
mat	1	geo
pow	2	gra, net
981	1	lan
986	1	eva
CAD	1	par
col	1	pro
Gal	4	Min, Abi, Ric, Pra
Ban	1	Spa
rea	1	fis
mic	1	app
Gar	1	976
Sim	1	Nij

(b) List of variables

Word	Abbreviation	Word	Abbreviation
CAD	CAD	library	libra
administrative	admin	management	manag
agriculture	agric	manufacturing	mal
application	appli	mathematic	mathe
astronomy	astro	medical	medic
automation	autom	microcomputer	micro
bibliographic	bibli	multiprocessing	multi
cartography	carto	nuclear	nucle
communication	col	organisation	organ
computer	compu	physic	physi
electronic	elect	picture	pictu
engineering	engin	power	power
environment	envir	reactor	react
fission	fissi	security	secur
geophysic	geoph	software	softw
information	infor	statistical	stati

Note: Dates are truncated to 1988 = 988 etc. Family names begin with a capital letter.

authors (present at least six times in the bulk of the reference set), dates, and other meaningful words are row headings.

The body of the matrix is constituted by the number of co-occurrences between words present in columns and rows. The French factorial analysis (A.F.C.) [14] was performed with the STATITCF [15] program. This analysis allows a representation of a dual space matrix in a unique space optimized by calculating the eigen values and eigen vectors of a KHI^2 distance matrix issued from the original matrix. The main plan represents 27% of the information present in the matrix, and concerns the axes 1 and 2 (Figure 1).

The graphic interpretation is as follows:

Three main periods can be considered: 1976–1980, 1980–1983 and 1983 to date.

First period: Most of the authors are present in the first period of the study. The dominant themes concern database management, administration, medicine, nuclear devices, and computers.

Second period: The data corresponding to the period 1980–1983 show no emerging trends.

After 1983: The themes are still present in a core, but peripheral information not present in the former period appears. The subjects are very different, dispersed and deal with a large variety of subjects. There are no authors present above the threshold of 6. The main themes concerned: engineering, physics, astronomy, statistics, and database organization.

At the periphery of this period, new trends appear: bibliometrics (which is related to indexation and classification), micro-computers and related software, tools for architecture and environment of databases, and image processing.

Figure 1 and Table l a represent the plan of the analysis, determined by the 1 and 2 axes.

3. Topology of a set of answers

3.1. Method

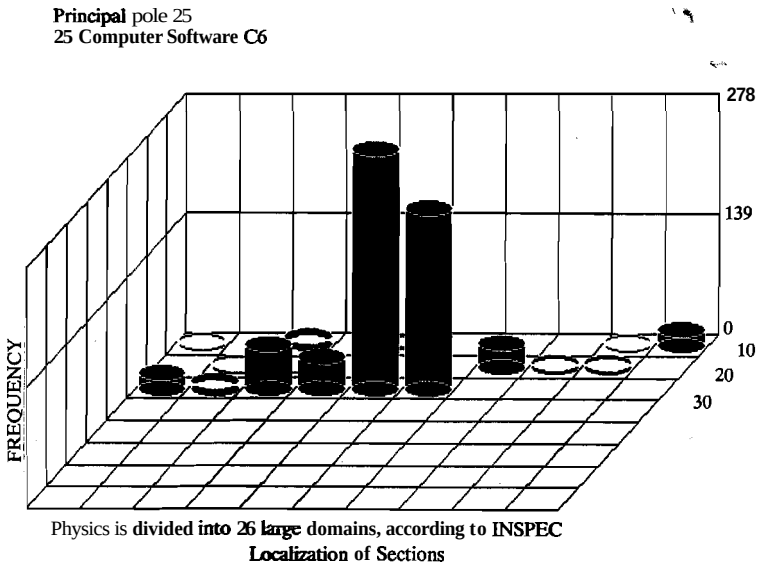
Most of the scientific and technical databases contain category codes which describe the subject to which the references are related. For instance, chemistry is divided into 80 sections, physics, according to the INSPEC file code, into 145 sections or 26 larges themes, the patents into 124 main groups according to the Derwent classes, etc.

The preceding method dealt with words, mainly present in Index Terms, Supplementary Terms, Authors and other fields.

Pole is the name of the software used now, to analyse topologically a set of answers after a query deals with the various codes used by the indexor to describe the areas of sciences related to the references.

The code field is extracted from the references, and the codes are limited to a certain number of digits corresponding to the way in which the information is coded in the various databases. We generally limit the number of topics, in order to obtain a fast infographic representation of the results on a microcomputer graphic screen.

It is also possible, with certain databases, to use special codes which are directly related to very precise items such as the Registry Numbers in the Chemical Abstracts [16].



Condensed matter: structure, thermal and mechanical properties A6, freq. 1.
Electromagnetic fields B5, *freq.* 1.
Cross-disciplinary physics and related areas of science and technology A8, freq. 3.
Physics, general AO, freq. 3.
Electrical. Components, electron devices and materials B2, freq. 4.
Electrical. Circuit theory and circuits B1, freq. 4.
Classical areas of phenomenology A4, freq. 5.
Electricals. Instrumentations and special applications B7, *freq.* 5.
Electrical Power systems and applications B8, freq. 5.
Computer and control. Control technology C3, *freq.* 8.
Nuclear physics A2, *freq.* 9.
Computer and control. Systems and control theory C1, *freq.* 17.
Geophysics, astronomy and astrophysics A9, freq. 18.
Communications B6, freq. 28.
Computer hardware C5, freq. 36.
Numerical analysis and theoretical computer topics C4, freq. 51.
Computer applications C7, freq. 208.
Computer software C6, freq. 278.

Figure 2. INSPEC—databases France: 1977–1988.

3.2. The different steps

The following example shows two coded fields from the above references dealing with the concept of databases in France between 1977 and 1988, and are issued from the former query. The host used was Orbit Information Technologies [11].

CC – *C6160D; C6170

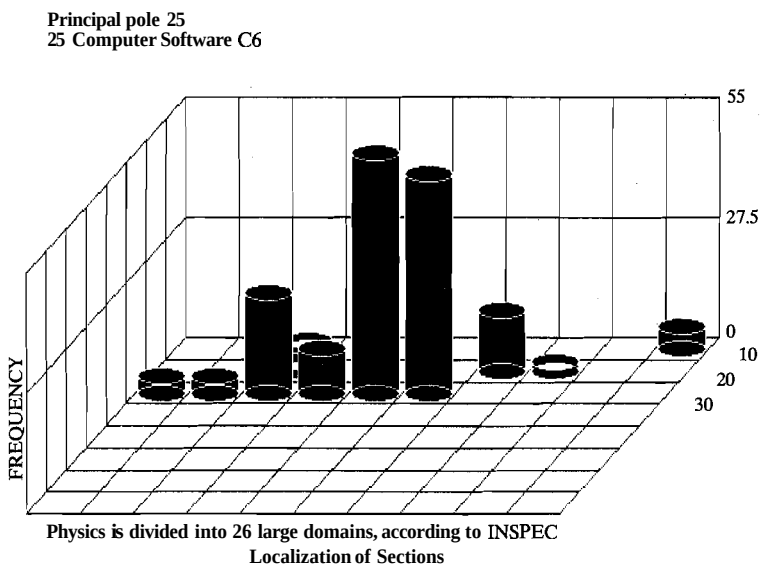
CC – *C6160D; C4250; C6140D; C6150C; C6120

In these references, the codes are a combination of figures and letters: They are also weighted. We extract from the coded field all the weighted codes which represent the main research areas. These codes are sorted, and their frequencies calculated. The file obtained allows the infographic representation of these main research areas.

To avoid discrepancies coming from a change in indexing policy between 1977 and 1988, we have truncated the codes to 2 digits. In this case all the codes remain the same for this period. A period can be analysed more completely by truncating the codes to three digits, then 145 different topics will be used.

Table 2
The 26 main areas of physics according to INSPEC

01	Physics, general A0
02	The physic of elementary particles and fields A1
03	Nuclear physics A2
04	Atomic and molecular physics A3
05	Classical areas of phenomenology A4
06	Fluids, plasmas and electric discharges A5
07	Condensed matter: structure, thermal and mechanical properties A6
08	Condensed matter: electronic structure, electrical, magnetic and optical properties A7
09	Cross-disciplinaryphysics and related areas of science and technology A8
10	Geophysics, astronomy and astrophysics A9
11	Electrical. General topics, engineering mathematics and materials science BO
12	Electrical. Circuit theory and circuits B1
13	Electrical. Components, electron devices and materials B2
14	Electrical. Magnetic and superconducting materials and devices B3
15	Optical materials and applications, electro-optics and optoelectronics B4
16	Electromagnetic fields B5
17	Communications B6
18	Electricals. Instrumentations and special applications B7
19	Electrical. Power systems and applications B8
20	Computer and control. General and management topics CO
21	Computer and control. Systems and control theory C1
22	Computer and control. Control technology C3
23	Numerical analysis and theoretical computer topics C4
24	Computer hardware C5
25	Computer software C6
26	Computer applications C7



Physics, general AO, freq. 1.
Classical areas of phenomenology A4, freq. 1.
Electrical. Circuit theory and circuits B1, freq. 1.
Electrical. Power systems and applications B8, freq. 1.
Electrical. Components, electron devices and materials B2, freq. 2.
Electricals. Instrumentations and special applications B7, freq. 2.
Nuclear physics A2, freq. 2
Computer and control. Systems and control theory C1, freq. 4.
Computer and control. Control technology C3, freq. 4.
Geophysics, astronomy and astrophysics A9, freq. 5.
Computer hardware C5, freq. 10.
Communications B6, freq. 14.
Numerical analysis and theoretical computer topics C4, freq. 23.
Computer applications C7, freq. 50.
Computer software C6, freq. 55.

Figure 3. INSPEC—databases France: very recent period (1985 to date).

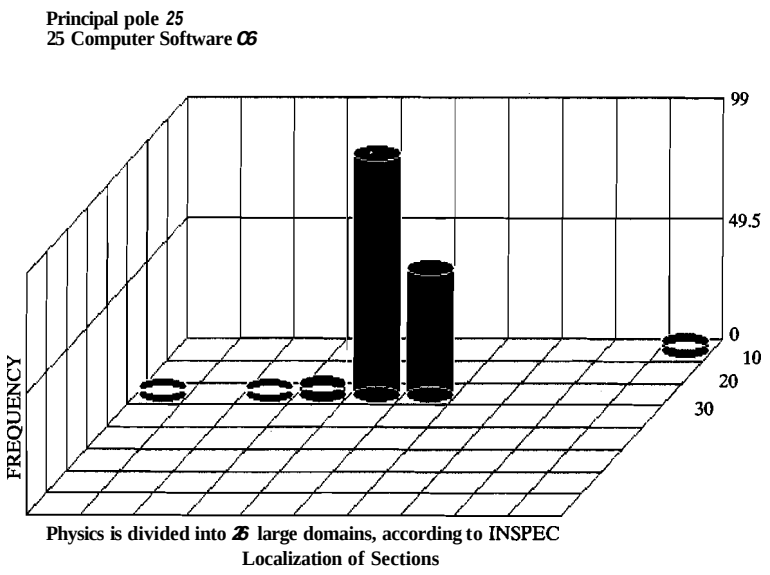
Figures 2, 3 and 4 show a classical representation. The correspondence between code numbers and topics, are given in Table 2.

4. Discussion

The results obtained from the keyword factorial analysis and from the research pole analysis are in good agreement.

The two main research poles deal with codes C7 and *C6:* computer applications and computer software, followed by numerical analysis, communications, astrophysics and geophysics.

To give a fast picture of the new trends in databases (at least in the field of INSPEC which covers physical sciences), we split into two periods the code field



Classical areas of phenomenology A4, freq. 1
Cross-disciplinary physics and related areas of science and technology A8, freq. 1.
Communications B6, freq. 1.
Computer and control. Systems and control theory C1, freq. 3.
Numerical analysis and theoretical computer topics C4, freq. 3.
Geophysics, astronomy and astrophysics A9, freq. 4.
Computer hardware C5, freq. 4.
Computer applications C7, freq. 52.
Computer software C6, freq. 99.

Figure 4. INSPEC — databases France: earlier period (1977–1979).

extracted from the 620 references. These two periods have been analysed separately and the results are presented Figures 3 and 4. They show the emergence of new areas of development in databases in physics.

The first period deals with the 155 older references covering 1977 to 1979, and the last period deals with 155 more recent references dealing mainly with the years 1985 to date.

These two ways of analysing the answers of a database could be incorporated directly into the main commands of a host. This will be a good complement to the statistical commands such as GET, ZOOM, etc., because links between words are visualized by factorial analysis, and also because the main research poles can be compared for different periods of time or for different subjects on a constant base when the MSPEC Classification is truncated to 2 or 3 digits.

5. Conclusion

The constant increase in the amount of literature and the use of CD-ROM will generate the need for new commands that are able to analyse globally the answers

obtained from online databases. ***This problem will remain even if more sophisticated softwares are used to discriminate and choose references,*** because the scientific and technical production is growing.

In our opinion, the combination of modules such as Datalink or Pole will be one of the answers. These modules already work on microcomputers for several hundred references. Our results suggest that they are useful to users as a mean of obtaining a synthetic view of the data. It appears from the results of the tests that they bring to the users new aspects of information.

References

[1] Creyssel, P. (1986). Second International Symposium on Chemical Information. December, Lyon.

[2] Dou, H., P. Hassanaly and A. La Tela (1987). Le développement scientifique et les réseaux d'expertise. Revue Française de Bibliométrie 1, 2–13.

[3] Callon, M., J. Courtial, W. Turner and S. Bauin (1983). From translation to problematic network; an introduction to Co-Word analysis. Information sur les Sciences Sociales, 191.

[4] Armstrong, C.J. and J. Large (1985). In house information retrieval on downloaded data. 9th Online Meeting, December London: Learned Information.

[5] La Tela, A. (1987). Système interactif d'aide à la décision, SIAD. Analyse statistique dynamique des bases de données. Thèse Science Marseille.

[6] Turner, W.A., G. Chartron and B. Michelet (1985). Describing scientific and technological problem networks. Manually and automatically indexed full text databases: some co-word analysis techniques. Paris: OECD.

[7] Le Coadic, Y. and R. Bertrand (1987). Les systèmes d'information scientifique et technique. Colloque DBMIST, Avril, Paris.

[8] Moureau and A. Girard (n.d.). Utilisation des bases de données brevets pour les études statistiques. Internal report 34686, IFP, Rueil Malmaison.

[9] Jakobiak, F. (1985). Utilisation d'outils bibliométriques et de recherche terminologique. Exemple d'utilisation. Les banques de données et le Videotex. Congrès, Palais des Congrès, September, Paris.

[10] Dunlop, S. (1987). A la découverte de la commande GET. UPDATE 45. Pergamon Infoline.

[11] Orbit Information Technologies (1988). Search light. London, 12 Vandy Street. January.

[12] Turner, W.A., B. Chartron, F. Laville and B. Michelet (1988). Packaging information for peer review: new co-word analysis techniques. In: A. van Raan, Handbook on Quantitative Analysis of Science and Technology. Amsterdam: Elsevier.

[13] Smith, M.W.A. (1986). A critical review of word-links as a method for investigating Shakespearean chronology and authorship. Lit. & Linguist. Comput. (GB) 1(4), 202–206.

[14] Benzécri, J.P. (1973). L'Analyse des Données. Vol. II: L'Analyse des Correspondances. Paris: Dunod.

[15] STATITCF, Statistical Analysis Shareware ITCF, Institut Technique des Céréales et des Fourrages, 8 avenue du Président Wilson 75116 Paris, France.

[16] Dou, H., P. Hassanaly, A. La Tela and M. Milon (1987). Le traitement de l'information scientifique et technique par les indicateurs scientométriques. Bull. Bibliothèques de France, September.