ETUDES DE CAS

LE TRAITEMENT DE L'IST PAR LES INDICATEURS SCIENTOMÉTRIQUES

Henri Dou, Parina Hassanaly, Albert Latela, Maurice Milon
Centre de recherche rétrospective de Marseille

LES SCIENTIFIQUES se trouvent actuellement confrontés à une masse croissante d'informations. Durant la dernière décennie, les bases et banques de données ont permis de faire face à cette inflation. Elles ne suffisent plus aujourd'hui au traitement correct des informations. De nouvelles techniques, faisant appel au traitement informatisé local, à l'analyse factorielle, au traitement des graphes, etc., vont permettre de mieux analyser la littérature pour prendre des décisions adaptées au contexte de la recherche effectuée. Cet article a pour but de présenter quelques résultats représentatifs obtenus par des méthodes originales développées dans notre laboratoire[1].

Nous prendrons comme exemple des informations issues de la chimie. Mais il faut garder à l'esprit que toute banque de données structurée en champs homogènes peut faire l'objet du même traitement (1). La démarche de départ est la suivante. Si on peut extraire de la littérature un ensemble de références à partir d'un certain nombre de descripteurs, cet ensemble est souvent volontairement restreint au sujet décrit par les mots clés utilisés, mais il est aussi limité par la capacité de lecture du demandeur d'information. Ainsi les banques de données et les puissants progiciels d'interrogation sont actuellement utilisés bien en deçà de leurs capacités, le but final étant souvent de réduire les réponses au strict nécessaire, ce qui élimine toute capacité d'innovation et de découverte en « feuilletant » (*browsing*).

En effet, le volume de la littérature croît, le nombre d'années antérieures accessibles augmente (environ 20 années actuellement), ce qui conduit pour des recherches même fines, à des ensembles de références importants. On aboutit alors à un paradoxe, puisque les outils modernes ne sont pas utilisés pour augmenter la capacité de réflexion et d'innovation, mais pour restreindre à des limites acceptables un sujet de réflexion. C'est pour cela que nous nous sommes engagés, au Centre de recherche rétrospective, dans une réflexion et une conception de produits qui permettront d'apporter un certain nombre de solutions à la situation antérieure.

Le but poursuivi est d'effectuer une recherche large, même très large, puis d'analyser celle-ci avec des programmes de traitement de l'information assez proches des systèmes d'intelligence artificielle. On peut ainsi, sans connaître les descripteurs utilisés, ni les mots, ni les codes employés, dégager les idées et les concepts les plus fréquents, ainsi que les groupes de références qui les contiennent. On aboutit donc tout de même à des ensembles analysables par un individu, mais sans restriction de recherche au départ.

Les méthodes scientométriques utilisent le concept suivant : élargir la recherche au départ, même au-delà des capacités de lecture du demandeur, puis analyser les références en « local », par des programmes de travail spécifiques, afin d'obtenir les informations sous-jacentes, les centres d'intérêt, et les relations entre les travaux plus ou moins liés par des données ou des idées communes.

Les *Chemical abstracts* (2) indexent la littérature chimique, en utilisant plus de 12 000 périodiques. Plus de 450 000 articles sont résumés chaque année. Cet outil bibliographique est diffusé soit sous forme de produit papier, soit sous forme d'une banque de données accessible en ligne. Le format des références est généralement constitué de différents champs qui se prêtent bien à une exploitation informatique. Ils sont caractérisés par des intitulés de champ et peuvent être imprimés ou stockés sur disque soit en format standard, soit selon l'ordre spécifié par l'utilisateur.

Le premier exemple des différentes méthodes de traitement des informations concerne le champ des codes de catégorie *(CATEGORY CODES)*, qui, dans le cas présent, contient la ou les sections des *Chemical abstracts* auxquelles le sujet se rapporte. La chimie est ainsi divisée en 80 sections. Dans une bibliographie d'une centaine de références (ce qui est assez commun en recherche) , il n'est plus possible d'effectuer une analyse manuelle de l'ensemble des thèmes mis en relation par l'ensemble constitué au départ.

Dans le champ des termes indexés *(INDEX TERMS)*, il existe des numéros de registre *(Registry num-*

1. Le Centre de recherche rétrospective de Marseille, Université Aix-Marseille III, est engagé dans un ensemble de recherches en scientométrie dans le cadre du programme PARUSI d'aide aux sciences de l'information.

bers) , qui caractérisent de façon univoque, mais sans signification chimique particulière, les composés utilisés (3). Ces numéros de registre sont de l'ordre de 5 à 10 et plus par référence, selon les domaines étudiés.

Une bibliographie de 100 références va conduire à la génération d'environ 500 à 1 000 numéros de registre. Leur analyse n'est plus du ressort de la simple lecture et ne peut pas être réalisée par les progiciels accessibles à travers les serveurs de banques de données. Ce sont des lacunes de cette sorte que comblent les recherches actuelles en science de l'information, principalement lorsqu'elles sont appliquées aux sciences exactes.

La même remarque peut être faite en ce qui concerne les mots contenus dans les termes supplémentaires *(SUPPLEMENTARY TERMS)* en langage libre. Dans ce cas, le simple classement de ces descripteurs est sans objet (ils sont trop nombreux), et il faut recourir à la méthode des co-occurrences des paires de mots (4,5) pour aboutir à une information supplémentaire par rapport aux données de base de la recherche : une paire de mots a un contenu en information bien supérieur à la somme du contenu informatif des deux mots isolés.

Le traitement des sections

Chaque référence est caractérisée par une section principale, et, dans certains cas, par une (ou des) section secondaire si la publication concerne un autre domaine de recherche. Le traitement automatique du champ des codes (CC) permet alors de déterminer :

— les pôles de recherche principaux d'un sujet ;

— la cartographie des liaisons qui existent entre pôles de recherche au niveau du sujet considéré.

Les pôles de recherche d'un sujet

L'analyse en fréquence de la répartition des sections principales dans une recherche bibliographique concernant les brevets traitant de l'acide tartrique (fig. 1.1 et 1.2) et les autres types de documents concernant le même composé (fig. 1.3) mettent en évidence les différences de pôles de recherche (définis par les numéros de sections des *Chemical abstracts*) entre la recherche appliquée (fig.1.1 et 1.2) et la recherche fondamentale (fig. 1.3).

Par exemple, en recherche appliquée, la section 74 *(Radiation chemistry, Photographic and other related process)* est le pôle de recherche principal, alors qu'en recherche fondamentale ce sont les pôles 16 *(Fermentation, Bioindustrial chemistry)* et 17 *(Food and feed chemistry)* qui sont les plus importants. On peut ainsi, par cette technique, effectuer des comparaisons rapides entre domaines, thèmes, auteurs, laboratoires, etc. Ces comparaisons sont possibles dans le temps, puisque pour une période donnée (dans notre cas la 11e collection), la structure et le contenu des sections restent identiques.

La cartographie des liaisons qui existent entre les pôles de recherche

La représentation précédente permet de dégager des pôles de recherche, mais n'explicite pas les relations de ces derniers avec les autres domaines de la chimie. Pour arriver à déterminer celles-ci, nous considérons les sections secondaires qui sont présentes dans les références bibliographiques des sujets analysés. Ainsi le pôle *Electrochemistry* (section principale 72), dans les publications traitant de l'acide tartrique, est associé à un certain nombre de sections secondaires ; on détermine ainsi les premières liaisons du graphe (fig. 2.1). Lorsque ces sections secondaires deviennent des sections principales, celles-ci génèrent à nouveau des liaisons, ce qui complète le graphe du pôle. Celui-ci est terminé lorsqu'il n'y a plus de sections secondaires, ou que les sections considérées deviennent un autre pôle de recherche.

On aboutit ainsi à un graphe de relations entre le pôle considéré et les autres domaines de la chimie. On peut donc déterminer les nœuds principaux qui vont constituer des éléments de comparaison. Cette méthode est généralisable selon les besoins et peut être utilisée comme outil de gestion scientifique (exemple de la figure 2.2 qui décrit en détail un pôle obtenu à partir de l'analyse des publications publiées par le *Bulletin de la Société chimique belge*).

Tableau

Description d'une référence issue de la base chemical abstracts

Accession number	CA03-224282(26)
Title	Possibility of photoisomerization of free radicals in solid phase
Authors	Mel'Inikov, M. Ya.; Seropegina, E.N.; Razskzov Yu. V.; Fok, N.V.
Source	Khim. Vys. Energ. (KHVKAO), V19 (5), p. 442-7, 1985, ISSN 00231193
Organizational source	Mosk. Univ., Moscow, USSR
Document type	J (Journal)
Language	Russ
Category codes	SEC74-1
Index terms	78-84-2; 124-40-3, reactions; 9003-05-8;9003-20-7 : (photochem.reactions of free radicals generated by radiolysis of)
Index terms	15337-44-7; 31277-24-4;99435-65-1; 99435-66-2;99435-67-3 : (photochem. reactions of, in solid state)
Index terms	Isomerization, photochem.; Substitution reaction, photochem. : (of free radicals IN solid phase)
Index terms	Photolysis : (of free radicals in solid phase, mechanism of processes in)
Index terms	Radicals, reactions : (photo chem. reactions of generated by, in solid phase)
Index terms	Radiolysis : (photolysis of free radicals)
Supplementary terms	Photolysis; free; radical; solid; state; photoisomerization; radical; solid; state

Le traitement des numéros de registre

Les composés chimiques sont représentés dans les *Chemical abstracts* par un numéro de registre. Sur l'ensemble des 7 000 000 de composés répertoriés depuis 1967, plus de 70 % n'apparaissent qu'une fois dans la littérature. Il est donc intéressant, pour une bibliographie donnée, de traiter auto-

matiquement ces derniers, afin de déterminer l'histogramme de leur apparition dans les travaux analysés. Divers profils, correspondant à des types de documents différents sont représentés dans les figures 3.1, et 3.2, les périodes considérées étant de 1982 à nos jours et de 1977 à 1981. Les ensembles bibliographiques ont été réalisés avec les mots clés suivants, recherchés sur l'index de base : HEADACHE× AND P/DT et HEADACHE× NOT P/DT.

HEADACHE(S)
— pour la période 1982-1985 : *256 références*
— limité à la période 1984-1985 : *94 références*
— se décomposant en : *22 brevets et 72 autres travaux*
— pour la période 1977-1981 : *177 références*
— limité à l'année1980 : *58 références*
— se décomposant en : *17 brevets et 41 autres travaux*

Ces résultats permettent de déterminer pour un sujet donné un profil standard, avec les produits les plus usuels. Ce profil sert alors de témoin dans le temps. Le même sujet est traité à intervalles réguliers et analysé de façon identique. Si le traitement fait apparaître une différence avec le profil standard, cela met en évidence des changements dans les habitudes de recherche, ou l'émergence de nouvelles molécules (6,7,8).

Comme le programme met en évidence les numéros de publications dans la recherche bibliographique en regard des numéros de registre utilisés, il est possible de tracer le graphe des relations existant entre les divers travaux. En ce sens, cette méthode conduit à des résultats différents de ceux obtenus en ligne avec les commandes ZOOM, GET, MEMTRI, utilisables sur différents serveurs[2]. On peut comparer l'ensemble des numéros de registres avec une liste de composés cible organisés en fichier informatique (par exemple des catalyseurs). On aboutit alors à une sélection des travaux sur des critères croisés : la présence d'un ou de plusieurs composés cible ; la déviation des travaux par rapport au profil standard des numéros de registre (par comparaison des deux histogrammes).

Il est évident que les molécules qui ont une structure très proche ont des numéros de registre différents. Ce phénomène limite les applications de la méthode, car on ne pourra pas homogénéiser les composés recherchés par sous-structure. A l'heure actuelle ceci n'est possible que par traitement des tables de connection ou des notations WLN[3] des composés chimiques, car la recherche en ligne par sous-structure ne permet plus ensuite de manipuler les résultats de la façon précédente.

Pour conclure, les banques de données bibliographiques qui sont des outils de recherche documentaire destinés à cerner un problème spécifique, peuvent aussi servir d'instrument de mesure et d'évaluation grâce aux différents champs documentaires. Un traitement adapté de cette information permet de faire de la recherche documentaire un outil d'analyse stratégique (9).

2. ZOOM, GET, MEMTRI sont des commandes de tri utilisées sur les serveurs ESA, INFOLINE, TQ. Elles permettent, à partir d'une réponse, de classer des données issues d'un champ (auteurs, numéros de brevets, codes...) par ordre de fréquences croissantes ou décroissantes.

3. WLN : Wisswesser line notation. Représentation linéaire de la structure d'une molécule en utilisant uniquement une succession de lettres ou de chiffres. Consulter pour plus d'information : W.J. WISSWESSER, « The ABC of the AWLN (Advanced Wisswesser line notation) », *Journal of information science*, 4, mai 1982, p. 69-77.

RÉFÉRENCES

1. **Garfield, E.**, *Citation indexing*, John Willey and sons, NY, 1979.
2. *Orbit technologies*, Pergamon
3. *Comment utiliser les* Chemical abstracts, American chemical society, Audio-course,1978.
Subject coverage and arrangement of Abstracts by sections, ACS, 1975 et 1982.
4. « Information scientifique et technique et Méthodes d'aide à la décision pour les sciences et l'industrie », Ecole d'été organisée par la DBMIST, septembre 1985.
5. CRRM, Programme de traitement statistique version 1.0, Marseille, 1986.
6. **Garfield, E.**, « Bradford's law and related statistical patterns », *Current contents*, mai 1980, p. 5.
7. **Lotka, A.J.**, « The Frequency distribution of scientific productivity », *Journal of Washington academic science*, 1926, p. 317.
8. **Zipf, G.K.**, *Human behavior and the principle of least effort*, Hafner, NY, 1972.
9. **Chaumier, J.**, *Analyse et langage documentaire, le traitement linguistique de l'information documentaire*, Entreprise moderne d'édition, Paris,1985.

Figure 1
Représentation en x, y (tri à plat), des travaux concernant l'acide tartrique
Sections des CA : 1982 à octobre 1985

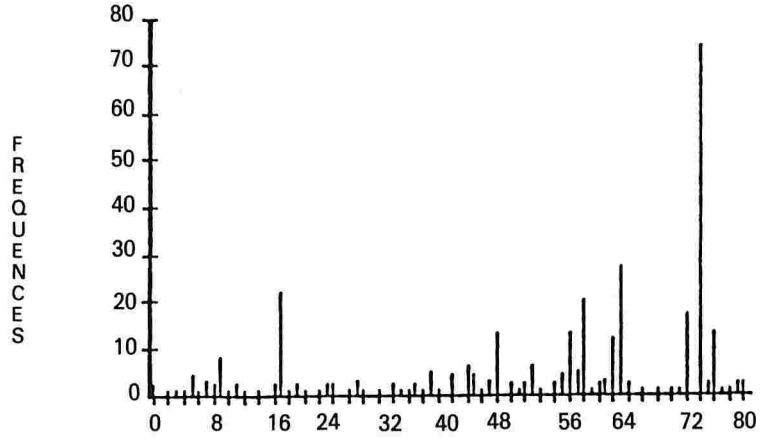


Fig. 1.1
Brevets

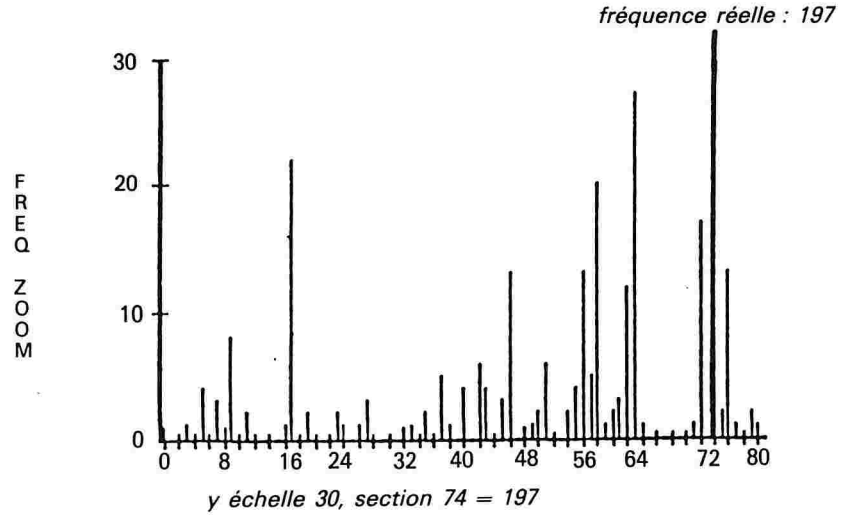


Fig. 1.2
Brevets

expansion d'échelle de la fig. 1.1

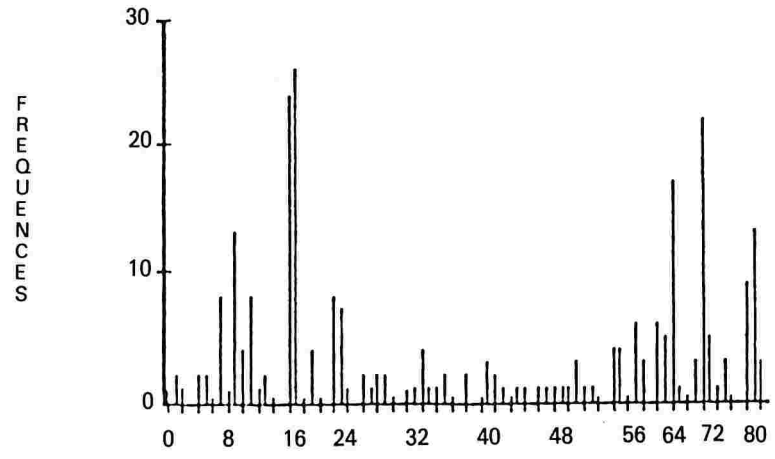


Fig. 1.3
Publications

* On notera la différence entre les brevets et les publications.

R

Fig. 2

1 : Acide tartrique non-brevets. Pôle 72 : electrochemistry
Représentation graphique

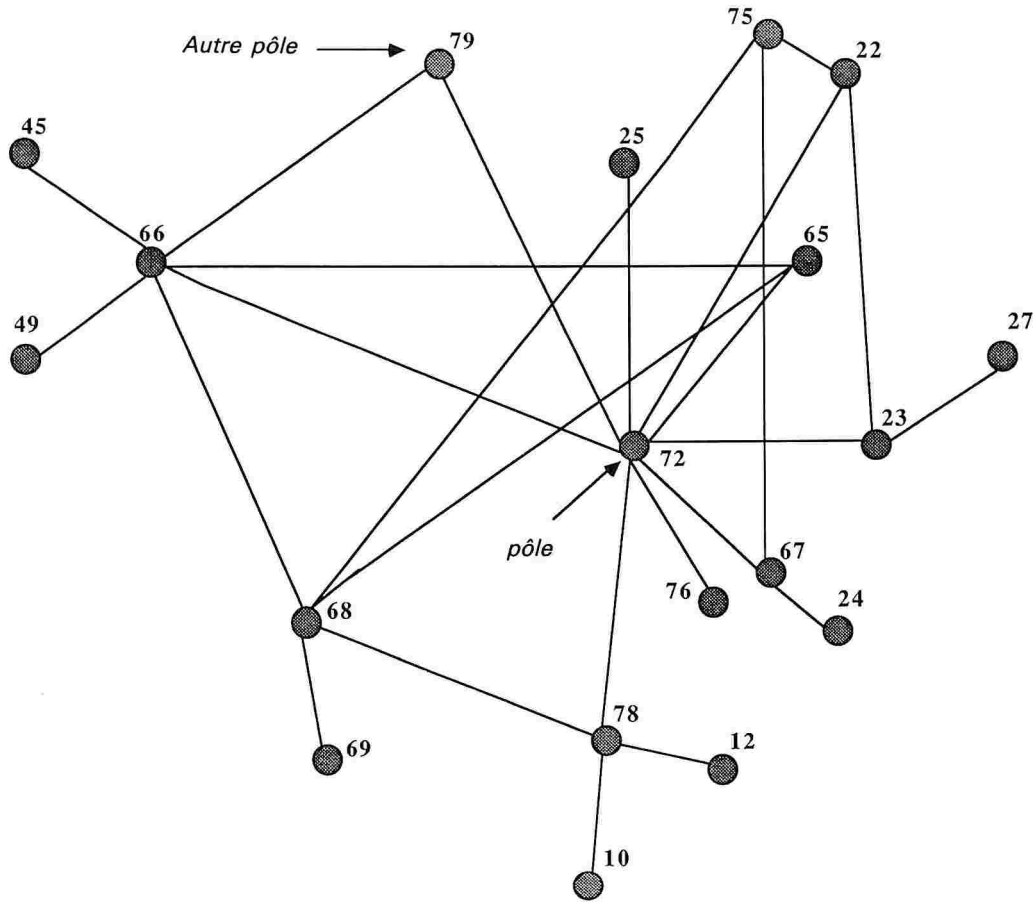


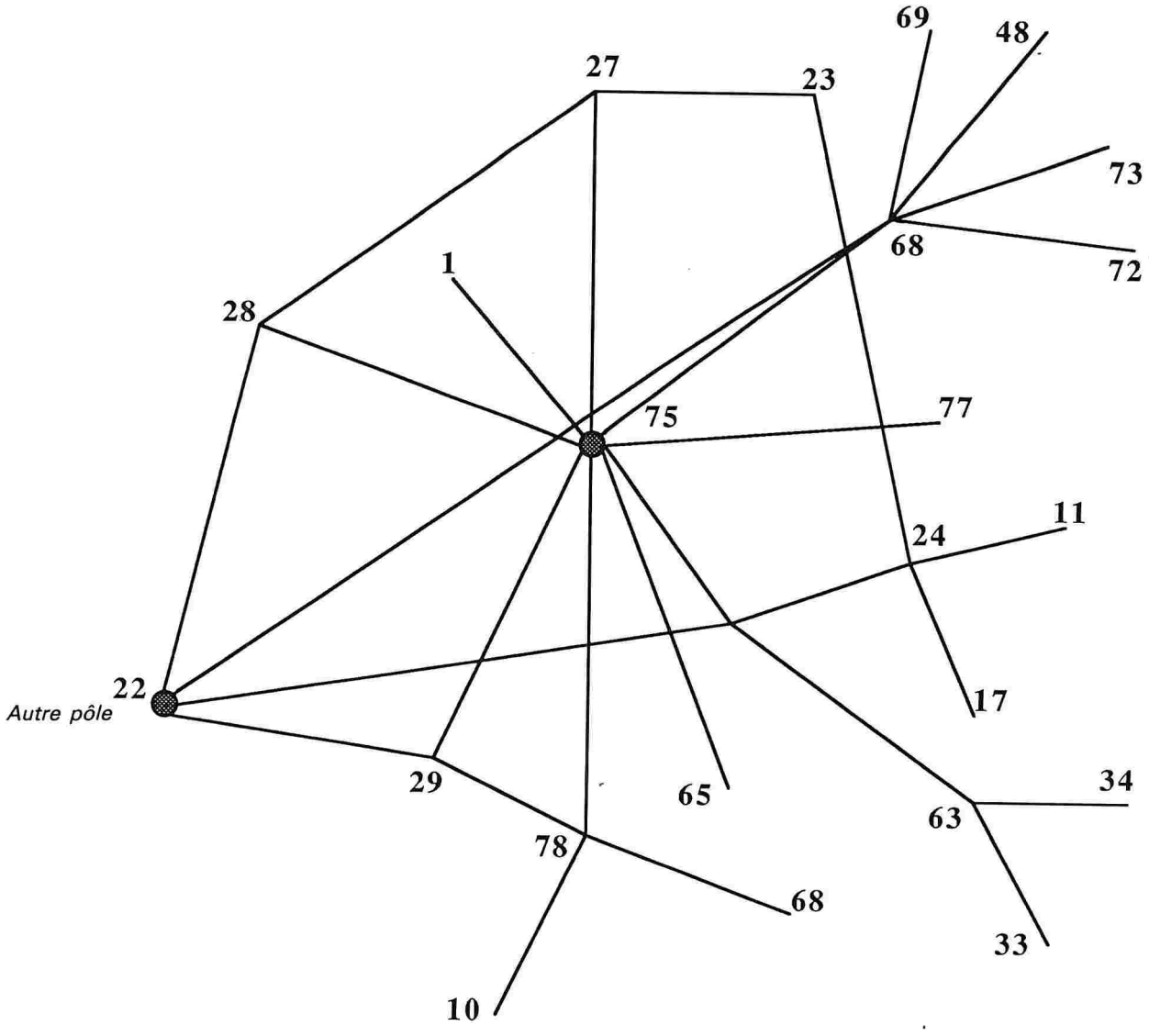
Fig. 2

2 : Représentation d'un des pôles de recherche se dégageant des publications présentes dans le *Bulletin de la Société chimique belge*

Pôle 75 : Crystallography and liquid crystals

Structure des liaisons :

1 Pharmacology
28 Heterocyclic compounds — more than one hetero atom
22 Physical organic chemistry
29 Organometallic — Organometalloidal compounds
78 Inorganic chemical reactions
10 Microbial biochemistry
68 Phase equilibrium — Chemical equilibriul — Solutions
65 General physical chemistry
77 Magnetic phenomena
25 Benzene — Derivatives — Condensed benzenoid compounds
24 Alicyclic compounds
17 Food and feed chemistry
11 Plant biochemistry
63 Pharmaceuticals
33 Carbohydrates
34 Amino Acids — Peptides — Proteins
68 Phase equilibrium — Chemical equilibrium — Solutions
72 Electrochemistry
73 Optical — Electron — Mass spectroscopy — Related properties
48 Unit operations — Processes
69 Thermodynamics — Thermochemistry — Thermal properties
23 Aliphatic compounds
27 Heterocyclic compounds — one hetero atom



R

Fig. 3

1 : Représentation de divers histogrammes relatifs aux numéros de registre

Brevets 84-85

Gamme de fréquences : 2-10
Fréquences supprimées : aucune

50-78-2	**	93235-16-6	**
56-12-2	**	93235-17-7	**
135-97-7	*****	93235-18-8	**
2292-08-2	***	93235-19-9	**
2905-62-6	**	93235-20-2	**
6613-44-1	***	93235-21-3	**
14618-78-1	**	93235-22-4	**
19718-92-4	**	93235-23-5	**
21900-23-2	**	93235-24-6	**
72594-40-2	**	93391-33-4	**

Brevets 77-81

Gamme de fréquences : 2-20
Fréquences supprimées : aucune

103-90-2	**	74137-71-6	**
7232-21-5	**	76163-79-6	**
10436-52-9	**	76163-85-4	**
19894-97-4	**	76163-86-5	**
36203-31-3	**	76163-88-7	**
40498-49-5	**	76163-89-8	**
56469-10-4	**	76163-92-3	**
59555-40-7	**	76163-93-4	**
74137-64-7	**	76163-94-5	**
74137-65-8	**	76163-95-6	**
74137-66-9	**	76163-96-7	**
74137-67-0	**	76163-97-8	**
74137-68-1	**	76163-98-9	**
74137-69-2	**	76163-99-0	**
74137-70-5	**		

Non-Brevets 84-85

Gamme de fréquences : 2-10
Fréquences supprimées : aucune

50-48-6	***	298-57-7	**
50-67-9	*********	361-37-5	****
51-41-2	***	525-66-6	***
51-45-6	**	4205-90-7	**
51-67-2	**	7440-02-0	**
56-65-5	**	7440-70-2	*******
58-00-4	**	9000-81-1	**
60-92-4	**	9002-62-4	**
64-17-5	**	15574-96-6	***
73-22-3	***	52468-60-7	*****
113-15-5	***	57808-66-9	***
129-03-3	***	62571-86-2	**

Non-Brevets 80-80

Gamme de fréquences : 2-20
Fréquences supprimées : aucune

50-67-9	****	379-79-3	***
113-15-5	**	4205-91-8	**

Figure 3

2 : Exemple de traitement, références non-brevets 80-80

Liste des fréquences, RN et Réfs.

Fréquence : 12 RN : Pas de RN Réfs : 6;8;14;16;18;21;24; 26;28;30;32;35;
Fréquence : 4 RN : 50-67-9 Réfs : 33;21;32;39;
Fréquence : 1 RN : 50-78-2 Réfs : 38;
Fréquence : 1 RN : 51-45-6 Réfs : 1;
Fréquence : 1 RN : 51-61-6 Réfs : 37;
Fréquence : 1 RN : 56-65-5 Réfs : 4;
Fréquence : 1 RN : 56-92-8 Réfs : 19;
Fréquence : 1 RN : 58-61-7 Réfs : 4;
Fréquence : 1 RN : 59-33-6 Réfs : 19;
Fréquence : 1 RN : 61-19-8 Réfs : 4;
Fréquence : 1 RN : 79-01-6 Réfs : 2;
Fréquence : 1 RN : 88-29-9 Réfs : 30;
Fréquence : 2 RN : 113-15-5 Réfs : 40;12;
Fréquence : 1 RN : 129-49-7 Réfs : 35;
Fréquence : 1 RN : 151-67-7 Réfs : 3;
Fréquence : 1 RN : 363-24-6 Réfs : 8;
Fréquence : 3 RN : 379-79-3 Réfs : 8;38;35;
Fréquence : 1 RN : 511-12-6 Réfs : 21;
Fréquence : 1 RN : 525-66-6 Réfs : 30;
Fréquence : 1 RN : 551-11-1 Réfs : 8;
Fréquence : 2 RN : 4205-91-8 Réfs : 24;22;
Fréquence : 1 RN : 6190-39-2 Réfs : 35;
Fréquence : 1 RN : 7439-93-2 Réfs : 11;
Fréquence : 1 RN : 9002-62-4 Réfs : 37;
Fréquence : 1 RN : 10024-97-2 Réfs : 3;
Fréquence : 1 RN : 13345-50-1 Réfs : 8;
Fréquence : 1 RN : 13710-19-5 Réfs : 38;
Fréquence : 1 RN : 13838-16-9 Réfs : 3;
Fréquence : 1 RN : 25655-41-8 Réfs : 3;
Fréquence : 1 RN : 29605-96-7 Réfs : 36;
Fréquence : 1 RN : 30484-77-6 Réfs : 14;
Fréquence : 1 RN : 32061-14-6 Réfs : 9;
Fréquence : 1 RN : 33507-63-0 Réfs : 10;
Fréquence : 1 RN : 34839-70-8 Réfs : 19;
Fréquence : 1 RN : 41598-07-6 Réfs : 8;
Fréquence : 1 RN : 51481-61-9 Réfs : 19;
Fréquence : 1 RN : 58962-34-8 Réfs : 8;
Fréquence : 1 RN : 61370-87-4 Réfs : 36;
Fréquence : 1 RN : 63758-79-2 Réfs : 39;
Fréquence : 1 RN : 77503-19-6 Réfs : 16;

Nombre total de RN : 58
Nombre réel de RN : 40