Elaboration d'un réseau cartographique de codes CIB à partir de commandes statistiques en ligne

Antonio da Silva, Sandrine Collet, Catherine Bainier-Girard, Luc Quoniam, Hervé Rostaing

ILE ROUSSE' 99

Journées d'études sur les systèmes d'information élaborée
Bibliometrie - Linguistique - Information Stratégique - Veille Technologique
Intelligence Economique

27 septembre au 1er octobre 1999

Élaboration d'un réseau cartographique de codes CIB à partir de commandes statistiques en ligne

*Antonio DA SILVA * ; Sandrine COLLET ** ; Catherine BAINIER-GIRARD * ; Luc QUONIAM *** ; Hervé ROSTAING ****

* ***Snecma***, *Site de Villaroche, 77550 Moissy-Cramayel*
** ***Bureau Van Dijk,*** *57 boulevard Montmorency, 75016 Paris*
** ***Université de Toulon / Saint-Raphaël***, *200 Av, Victor Sergent, 83700 St Raphäel*
*** ***CRRM***, *Université Aix-Marseille III, 13397 Marseille Cedex 20*

Résumé

L'information technologique contenue dans les brevets est une information de grande valeur pour les entreprises. Mais, l'accès aux bases de données brevets en ligne est cher, en particulier lorsqu'on désire réaliser l'étude d'un domaine. Les outils d'analyses classiques qui ont été développés jusqu'à présent, ne fonctionnent en effet qu'à partir des notices au format complet, dont le coût d'acquisition est important. Pour des raisons d'économie, il est alors préférable de procéder d'une autre manière.

L'objet de cette communication est de décrire une méthode faisant appel à la bibliométrie, ainsi qu'aux commandes statistiques en ligne (nettement moins coûteuses que les notices au format complet) disponibles sur les centres serveurs, pour élaborer une cartographie relationnelle de l'activité technologique d'un domaine.

Introduction

Avec l'explosion des Nouvelles Technologies de l'Information et de la Communication, l'information fait aujourd'hui l'objet de toutes les attentions. La pléthore de documents rendus accessibles à tous par la magie du réseau des réseaux est à l'origine de ce phénomène.

Pourtant, l'information professionnelle, et particulièrement l'Information Scientifique et Technique, n'ont pas attendu le succès de l'Internet pour être exploitées à grande échelle

par l'intermédiaire des bases de données commerciales.

Alimentés depuis un trentaine d'années, ces immenses réservoirs d'information représentent aujourd'hui plusieurs dizaines de milliards de documents, à comparer aux quelques centaines de millions de pages présentes sur le Web [VEI 98].

De plus, ce type d'information possède de gros avantages par rapport à l'information en provenance de l'Internet. D'une part, elle est structurée et donc parfaitement adaptée à une analyse statistique de son contenu. D'autre part, cette information est issue de publications soumises à un comité éditorial, ce qui lui confère une valeur et une fiabilité certaines. Ceci est loin d'être le cas de l'information en provenance du Web, où la totale liberté de diffusion permet certains excès qui peuvent aller jusqu'à la désinformation. L'exploitation de l'information contenue sur les bases de données commerciales reste donc incontournable.

Cependant, les méthodes classiques d'analyse de ces ressources nécessitent actuellement l'acquisition des références bibliographiques dans leur format intégral. Le coût de ces analyses est par conséquent relativement élevé, et reste donc réservé aux grands groupes industriels.

L'objectif de cette communication est d'exposer une méthode alternative aux méthodes classiques, qui permet une analyse de contenu des notices sans nécessiter leur télédéchargement.

La première partie de cette communication exposera la théorie de cette méthode alternative, basée sur les fonctions de statistiques en ligne offertes par les grands serveurs. La seconde partie validera par l'exemple le bien fondé de cette nouvelle méthode.

Une méthode alternative

Information brevet

Partant du fait qu'il existe une corrélation entre recherche et application industrielle, on considère le brevet comme un " indicateur de technologie " [JAK 94]. Le brevet demeure la source de renseignement technique par excellence. On estime que 80 % de l'information technique est contenue uniquement dans les brevets [WAG 92]. Cette information est en grande partie exploitée par l'intermédiaire des bases de données commerciales.

D'une manière générale, seuls les grands groupes industriels exploitent efficacement ces données. En effet, l'acquisition d'une telle source de renseignement a un coût difficilement abordable pour les petites et moyennes entreprises. Nous verrons dans cet article comment mettre à leur portée financière ce type d'analyse.

Intérêt des réseaux de relations

L'analyse réseau présente deux avantages non négligeables [BOU 99]. D'une part, elle permet de présenter une vue synthétique d'un corpus de documents. D'autre part, elle met en relief les relations qui existent entre les éléments d'une même catégorie informationnelle. Ainsi, à partir d'un corpus de notices bibliographiques de brevets, on peut dresser des cartographies de sociétés déposantes, d'inventeurs, et faire l'état de l'art d'un domaine technologique en analysant les codes CIB (Classification Internationale des Brevets).

Ces réseaux de relations permettent d'avoir une photographie des actions menées par les acteurs de la technosphère de l'entreprise. Ainsi, identifier les menaces et déceler les opportunités devient plus évident.

Tel qu'un bibliomètre le conçoit, un réseau est composé de sommets, symbolisant les formes, et de segments déterminant la force du lien entre les sommets. Dans ce type de représentation, la co-occurrence de deux

termes dans une même référence se traduit par un lien physique [ROS 96].

Par ailleurs, l'analyse réseau présente un avantage certain : la mise en relief des liaisons entre les formes permet de visualiser de manière graphique l'interconnexion entre les différents domaines liés à la technologie étudiée.

De plus, elle permet d'identifier clairement les sous-groupes d'un graphe, et par là même de faire ressortir les axes majeurs de la réflexion dans ce domaine [BOU 96].

Eléments nécessaires à la réalisation d'un réseau de relations

Pour la représentation d'un réseau de relations, nous nous sommes appuyés sur un logiciel spécialisé dans la construction automatique de réseau : le logiciel Matrisme [MAT 97]. Pour réaliser une cartographie réseau, ce logiciel nécessite en entrée une matrice de fréquence de co-occurrences de formes qui soit carrée et symétrique : les mêmes formes doivent apparaître dans les intitulés de lignes et de colonnes, et ce dans le même ordre.

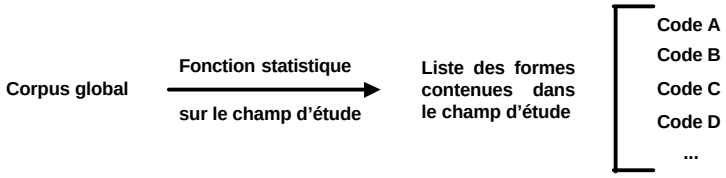
Habituellement, les matrices de ce type sont construites automatiquement par des outils bibliométriques, tels que Dataview [DAT 93], à partir d'un corpus de notices bibliographiques. Notre objectif est de construire la même matrice, non pas à partir des notices intégrales, mais à partir des fonctions de statistiques en ligne [DOU 91] disponibles sur les grands centres serveurs tels que Dialog ou Questel-Orbit.

Les trois étapes de la méthode

Avant même de débuter la construction de la matrice, il est nécessaire de savoir quel type d'analyse réseau doit être réalisé : réseau d'inventeurs, de sociétés, de codes CIB…

Les interrogations statistiques qui suivront seront alors appliquées au champ correspondant. Nous l'intitulerons " champ d'étude ".

La première étape de la méthode consiste à délimiter le corpus des documents sur lequel nous réaliserons l'analyse réseau. Ceci se fait par une requête que nous appellerons " équation de base ", qui fixera en réponse un corpus que nous nommerons " corpus global ".
La deuxième étape de la construction de la matrice consiste à connaître toutes les formes présentes dans le champ d'étude. Ceci est possible en opérant sur le corpus global une



statistique portant sur le champ d'étude.

La liste des formes ainsi obtenues, va permettre d'établir la structure de la matrice (nombre de lignes et de colonnes), et d'y inscrire toutes les entrées :

	Code A	Code B	Code C	Code D	…
Code A					
Code B					
Code C					
Code D					
…					

La troisième étape de la méthode consiste à remplir colonne par colonne la matrice.

Procédons par l'exemple : pour le code A, il nous faut savoir à quelle fréquence de co-occurrence il apparaît avec chacun des autres codes dans le corpus global. On posera donc comme requête : " corpus global ET code A ",

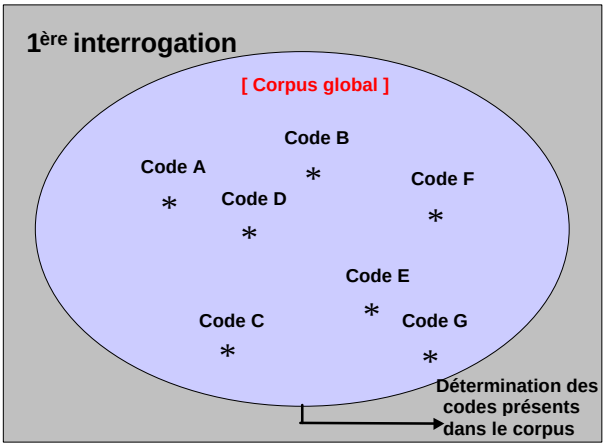
suivi d'une commande statistique sur ce corpus réduit.

La liste des codes obtenus, avec les fréquences de co-occurrences correspondantes, est à reporter dans la matrice. La même opération est à répéter pour chacun des codes présents dans la première liste.

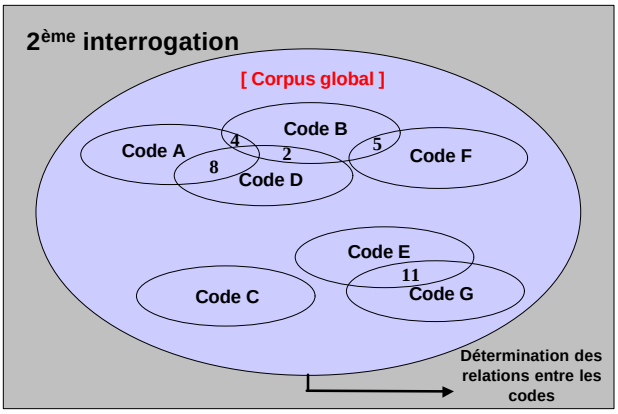
Lorsque cette deuxième série de statistiques est achevée, la matrice est alors intégralement complétée.

Les résultats obtenus par cette méthode peuvent être expliqués par la théorie des ensembles.

La première commande statistique en ligne permet de connaître les éléments du corpus global.



La deuxième série de statistiques permet de savoir, dans ce corpus global, quels sont les liens entre les différents éléments, et dans quelles proportions ils sont liés.



Validation de la méthode

Définition de l'étude

Dans l'exemple qui va suivre, nous nous sommes fixés le sujet d'étude suivant :

Les superalliages monocristallins.

Les mots-clés choisis pour définir le corpus ont été choisis avec l'aide d'un expert du domaine : Alloy(s), Super Alloy(s), Nickel base alloy(s), Single crystal(s), Monocrystal(s). Nous interrogerons la base brevets WPI sur le centre serveur Dialog.

L'analyse réseau portera sur le champ des codes CIB, permettant ainsi de visualiser les thèmes abordés par cette technologie. Seule l'année 1998 sera considérée.

Définition de la stratégie d'interrogation

L'équation de base est :

(alloy? ? or super?alloy? ? or nickel(w)base(w) alloy?) and (single(w)crystal? or single?crystal? or monocrystal?) and PY = 1998

Le corpus global étant défini, on procède alors à une première analyse statistique qui va permettre de connaître les différents codes CIB contenus dans ce corpus.

?rank ic
Started processing RANK

Completed Ranking 95 records
DIALOG RANK Results

RANK: S4/1-95 Field: IC= File(s): 351
(Rank fields found in 95 records -- 322 unique terms)

RANK	No. Items	Term
1	20	C22C-019/05
2	13	C30B-011/00
3	13	C30B-029/52
4	9	F01D-005/28
5	8	C22C-019/03
6	7	C22F-001/10
7	7	C30B-029/04
8	6	G11B-005/66
9	5	C23C-016/26
10	5	C30B-025/18
11	5	H01L-021/205
12	5	H01L-023/48
13	4	B22D-027/04
14	4	C22C-001/02
15	4	H01L-021/208
16	4	H01L-023/52
17	4	H01L-029/40
18	4	H01L-033/00
19	3	C23C-014/24
20	3	C23C-016/00
21	3	C30B-011/14
22	3	C30B-015/00
23	3	C30B-023/02
24	3	C30B-025/02
25	3	C30B-029/06
26	3	C30B-029/22
27	3	F01D-005/14
28	3	G11B-005/85
29	3	H01F-010/08
30	3	H01L-021/20
31	3	H01L-021/265
32	3	H01L-021/44
33	3	H01L-031/04
34	2	B22C-009/00
35	2	B22C-009/04
36	2	B22C-009/22
37	2	B24D-005/12
38	2	B32B-005/16
39	2	B32B-009/00
40	2	B32B-015/00
41	2	C22C-000/00
42	2	C22C-001/00
43	2	C22C-019/07
44	2	C22C-032/00
45	2	C22C-038/00
46	2	C22C-038/08
47	2	C23C-016/02
48	2	C23C-016/44
49	2	C23C-028/00
50	2	C30B-015/30

…

La liste obtenue contient 322 codes CIB différents, dont 250 codes hapax (fréquence = 1).

Un outil d'analyse réseau classique construirait donc une matrice de 322 × 322. Cependant, pour des raisons de lisibilité évidentes, une analyse réseau ne doit pas comporter plus d'une cinquantaine de sommets. Seule la partie la plus significative de la matrice est donc réellement utilisée.

C'est pourquoi dans cette méthode alternative, nous ne reconstruirons pas la matrice dans son intégralité, puisque cela n'est pas nécessaire. En fixant le bruit statistique (information vide de sens) à une fréquence inférieure à 3, seuls les 33 premiers codes CIB de la liste précédente seront pris en compte.

Afin d'obtenir les liens caractérisant le corpus, chacun des 33 codes est successivement combiné par un opérateur booléen ET avec le corpus global. Chacune de ces combinaisons fait l'objet d'une analyse statistique en ligne, comme c'est le cas dans l'exemple suivant :

?s s4 and C22C-019/05

?rank ic
Started processing RANK
Completed Ranking 20 records
DIALOG RANK Results

RANK: S5/1-20 Field: IC= File(s): 351
(Rank fields found in 20 records -- 59 unique terms)

RANK	No. Items	Term
1	11	C30B-029/52
2	10	C30B-011/00
3	8	F01D-005/28
4	7	C22C-019/03
5	7	C22F-001/10
6	3	C30B-011/14
7	2	C22C-000/00
8	2	C22C-001/02
9	2	C22C-019/07
10	2	C22C-038/08
11	2	F02C-007/00
12	1	B22C-009/22
13	1	B22D-011/06
14	1	B22D-025/06

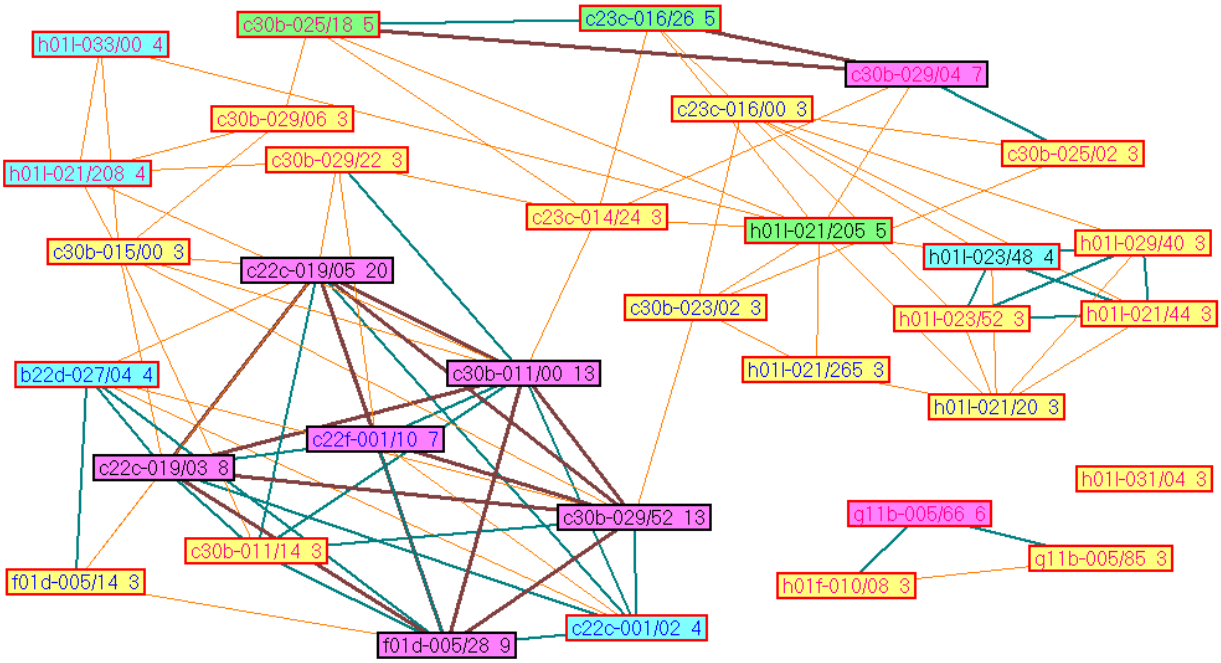
15 1 B22D-027/04
16 1 B22D-027/20
17 1 B22F-005/04
18 1 B23K-010/02
19 1 B23K-026/00
20 1 B23K-035/30
21 1 B23P-006/00
22 1 B23P-006/04
23 1 B24D-011/02
24 1 B32B-015/00
25 1 C03B-000/00
26 1 C03B-011/00
27 1 C03B-037/04
28 1 C03B-037/08
29 1 C03B-037/095
30 1 C21D-001/00
31 1 C22C-001/00
32 1 C22C-001/10
33 1 C22C-009/05
34 1 C22C-013/03
35 1 C22C-019/00
36 1 C22C-019/04
37 1 C22C-019/15
38 1 C22C-032/00
39 1 C22F-001/00
40 1 C30B-000/00
41 1 C30B-007/08
42 1 C30B-013/04
43 1 C30B-015/00
44 1 C30B-021/02
45 1 C30B-029/22
46 1 C30B-033/00
47 1 D01D-004/02
48 1 D01D-005/08
49 1 D01F-009/08
50 1 F01D-005/00
51 1 F01D-005/12
52 1 F01D-005/14
53 1 F01D-005/30
54 1 F01D-009/02
55 1 F01K-023/10
56 1 F02C-003/00
57 1 F02C-006/00
58 1 F02C-007/18
59 1 F02D-043/00
---end of results---

Réalisation de la matrice

Une matrice de 33×33 est alors reconstituée à l'aide d'un tableur quelconque.

	B22D-027/04	C22C-001/02	C22C-019/03
B22D-027/04	4	1	2
C22C-001/02	1	4	2
C22C-019/03	2	2	8
C22C-019/05	1	2	7
...	...	...	...

Cette matrice est ensuite exploitée par le logiciel Matrisme. L'exemple suivant, concernant les super alliages monocristallins sur l'année 1998, montre le résultat que l'on obtient alors.



Réseau des codes CIB des superalliages monocristallins sur 1998 (fréquence > 2)

En conclusion

Ci-dessous, voici un tableau synthétisant le coût global de chacune des deux méthodes d'analyse réseau, réalisées avec le serveur Dialog : la méthode classique versus la méthode alternative.

Type(s) d'analyse(s)	Méthode classique	Méthode alternative	(MA × 100) / MC
CIB	2946,31 FF	333,92 FF	11,3 %
Inventeurs (IN)	2946,31 FF	222,62 FF	7,5 %
Sociétés (SD)	2946,31 FF	203,24 FF	6,9 %
CIB + IN + SD	2946,31 FF	390,60 FF	13,2 %

MC : Méthode Classique
MA : Métode Alternative

Tarifs Dialog :
1 rank obtenu coûte 0,03 $
1 notice au format complet coûte 4,72 $

1 $ = 6,1530 FF au jour de l'interrogation

Avantages de la méthode alternative

L'avantage majeur de cette nouvelle méthode d'analyse est, bien entendu, son faible coût (environ 10% du coût de la méthode classique), ce qui autorise une exploitation plus large de l'analyse réseau.

De nouvelles perspectives s'ouvrent à deux différents types d'acteurs. D'une part, les PME / PMI, qui jusqu'à présent n'avaient pas accès à ce type d'analyse, pourront s'ouvrir à l'information élaborée.

D'autre part, les grands industriels, qui exploitent déjà l'analyse réseau, pourront utiliser cette nouvelle méthode à grande échelle pour pouvoir surveiller systématiquement tous leurs domaines d'activités. La finalité de l'analyse réseau devient alors une simple surveillance permettant de détecter une information digne d'intérêt - information qui donnera lieu à des analyses plus poussées dont les budgets seront plus conséquents - .

L'utilisation de cette méthode en tant que " pré-analyse " permet donc de n'engager des moyens que lorsque cela s'avèrera nécessaire.

Références

[BOU 96]. *Traitement de l'information : analyse de données classiques versus analyse de réseau. Un cas d'application : la bibliométrie*
Eric BOUTIN, Luc QUONIAM, Hervé ROSTAING, Philippe DUMAS
Dixième Congrès National des Sciences de l'Information et de la Communication, 1996

[BOU 99]. *Le traitement d'une information massive par l'analyse réseau : méthode, outils et applications*
Eric BOUTIN
Thèse, N° 99AIX0003, 1999

[DAT 93]. *Le logiciel bibliométrique DATAVIEW et son application comme outil d'aide à l'évaluation de la concurrence*
Hervé ROSTAING, William NIVOL, Luc QUONIAM, Albert LATELA
1993

[DOU 91]. *Automatic Generation of Strategic Matrices from Online Databases*
Henri DOU, Parina HASSANALY, Shirley SNEE
World Patent Information, Pergamon Press plc, 1991

[JAK 94]. *Le brevet source d'information*
François JAKOBIAK
Editions DUNOD, 1994

[MAT 97]. *Logiciel de construction automatique de réseaux*
Réalisé par Eric BOUTIN
Réalisation conjointe des laboratoires CRRM et Le Pont (Université de Toulon), 1997

[ROS 96]. *La bibliométrie et ses techniques*
Hervé ROSTAING
Co-édition Sciences de la société et CRRM, 1996

[VEI 98]. *Les banques de données classiques : Cent fois le volume du Web*
François LIBMANN, Directeur de FLA Consultants
VEILLE, N°13, 1998

[WAG 92]. *Le brevet d'invention*
Jean-Michel WAGRET
Presses Universitaires de France, 1992