ETUDE BIBLIOGRAPHIQUE DE LA RECHERCHE SCIENTIFIQUE AU MAROC

Bernard DOUSSET

dousset@irit.fr

Institut de Recherche en Informatique de Toulouse, IRIT-SIG, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 9 (France),

Mots clefs :
Intelligence économique, Veille scientifique et technique, Bibliométrie, Scientometrie, Analyse de réseaux, Indicateurs, Collaborations internationales, Co-auteurs, Productivité scientifique.
Keywords:
Competitive Intelligence, Science and technology watch, Bibliometry, Scientometry, Network analysis, Indicators, International collaborations, Co-authorsip, Scientific productivity.
Palabras clave :
Inteligencia Competitiva o "Económica", Vigilencia Científica y Tecnologica, Bibliometría, Cienciometría, Análisis de Redes, Indicators, Collaboratio Internationale, Co-autoría, Productividad científica.

Résumé :

Nous allons présenter, lors de cette conférence, un ensemble de méthodes de fouilles de textes qui sont largement utilisées dans notre laboratoire pour dresser un état des lieux de la recherche dans un contexte donné : domaine scientifique spécifique ou, comme ici, zone géographique bien délimitée. Nous avons choisi comme sujet la recherche scientifique marocaine afin d'illustrer ce que peut apporter, en amont, le text mining à l'intelligence économique. Afin de limiter le coût de l'étude et de garantir sa représentativité nous avons choisi de ne travailler que sur la base bibliographique PASCAL de l'INIST-CNRS, qui assure une bonne couverture des publications scientifiques francophones et notamment Marocaines. Une étude plus complète peut être réalisée de la même manière en intégrant des bases comme le Web of Science (SCI), Current Contents, Medline ou Science Direct. L'hétérogénéité de ces bases n'est pas vraiment un problème face aux puissantes fonctions d'analyse morphologique proposées par notre plateforme d'analyse bibliométrique « Tétralogie ». Malgré tout, le temps passé à la constitution d'un corpus multi-sources et à sa préparation grève inévitablement le budget d'une telle étude. Pour les mêmes raisons, nous nous sommes limités aux 8 dernières années soit de 2002 à mi 2009. Le nombre de publications est largement suffisant pour dresser un état des lieux de la recherche et pour en établir son évolution récente. Nous nous sommes attaché à faire ressortir les collaborations internationales, les signaux forts, les nouveaux sujets de recherche, les journaux scientifiques les plus utilisés pour publier, l'évolution des équipes, … Cette liste n'est bien entendu pas exhaustive, nous n'avons pas exploité le texte libre (titres et résumés) ainsi que des données plus technique comme le type de document, les éditeurs, les conférences, les villes, les langues, les codes de classification, … Les différents indicateurs que nous proposons sont déjà très évocateurs de l'état des lieux de la recherche Marocaine, une étude plus complète est toujours possible, mais pour pousser plus loin il convient de se focaliser sur chaque discipline afin de limiter la complexité essentiellement due à la multiplicité des acteurs (auteurs, laboratoires, journaux et congrès) et de la terminologie (thésaurus, mots clés, codes, multi-termes).

1 INTRODUCTION

Le but de cette étude est d'illustrer certaines méthodes de fouille de texte utilisées en bibliométrie en traitant un sujet connexe à la tenue de VSST cette année à Marrakech. En nous intéressant à la production scientifique marocaine, nous avions deux objectifs : montrer la diversité et l'efficacité des outils d'analyse disponibles à l'heure actuelle ainsi que certaines difficultés d'exploitation liées à la forme même des données sources. Nous avons constitué un corpus de publications scientifiques issu de la base bibliographique PASCAL de l'INIST-CNRS sur une période de près de 6 ans (2002 à 2007) et comportant au moins une fois dans le champ adresse un organisme Marocain. Pour réaliser cette étude, nous avons largement utilisé notre plate-forme « Tétralogie » dédiée à la veille stratégique en essayant, chaque fois que possible, d'illustrer notre propos par des approches de traitement et de visualisation différentes. Nous ne pouvons présenter, ici, qu'une faible partie des résultats potentiels de ce type d'étude, le but étant surtout d'informer le public sur l'ensemble des possibilités offertes en extraction de connaissances à partir des données textuelles.

Comme « Tétralogie » permet de se connecter à distance sur les éléments d'une telle analyse, les participants au colloque pourront, à loisir, venir consulter les résultats chiffrées déjà obtenus et éventuellement pousser plus loin les investigations en réalisant des zooms sur leurs centres d'intérêts particuliers (zone géographique, domaine de recherche, équipe, laboratoire, …).

2 CARACTERISTIQUES DU CORPUS ETUDIE

2.1 Description de la Base PASCAL (INIST-CNRS)

Principales caractéristiques de cette base:

- Disponible sur CD/Rom avec abonnement
- Télédéchargeable par groupe de 1000 notices
- Mise à jour régulière
- Sciences exactes et appliquées
- Mots-clés (descripteurs et identifieurs) en français, anglais et espagnol
- Les balises courtes ne permettent pas de distinguer les champs DE: et IN: en anglais, français et espagnol
- Codes de classification
- Titre et résumé d'une dizaine de lignes
- Adresse de tous les laboratoires concernés

Format de l'adresse:

AD: Department of Mathematics, University of Science and Technology of China, Hefei 230026, **China**; Department de Mathematiques, Faculte des Sciences Semlalia, Universite Cadi Ayyad, B.P. 2390, Marrakech, **Morocco**; Department of Mathematics, Morgan State University, 1700 E. Cold Spring Lane, Baltimore, MD 21251, **United States**

- Le champ auteur est assez pollué par les éditeurs, préfaceurs, directeurs, traducteurs, …

Support et périodes télé-déchargées : CD/Rom de janvier 2002 à août 2007.

Volume du corpus : 5140 fiches bibliographiques

2.2 Problème de reformatage de la base

Nous avons été obligé de reformater légèrement cette base pour les raisons suivantes :

- Obligation de prendre l'option balises longues (noms donnés par PASCAL aux différents champs de la base) car les balises courtes ne permettent pas de distinguer « Descripteurs » et « Identifieurs » disponibles en 3 langues (d'où mélange des 3 langues dans le résultat de chaque extraction).
- Caractères accentués saisis dans différentes pages codes, d'où génération inutile de variations morphologiques.
- Absence de certains accents générant de nouvelles variations.
- Nettoyage du champ « Auteur » de tous les parasitages (éditeurs, préfaceurs, …)

2.3 Méta données de Tétralogie pour la Base PASCAL

Sur la plate-forme Tétralogie, il est nécessaire d'établir un descripteur de la structure de la source utilisée (meta données) afin que les outils d'extraction puissent parfaitement s'adapter au traitement d'un format bibliographique particulier. Il faut en outre définir chaque champ (nom, séparateurs, utilité, filtrage) et trouver une balise de synchronisation pour délimiter chaque document. Voici le descripteur Tétralogie pour la base PASCAL (en gras les champs qui ont changé de nom).

```
Notice
# descripteurs des champs de la base Pascal sur CD/Rom #
# nom          abrev  champ      visible Separateurs  #
Notice         NO     Notice     True    de"ORD1"
Titre          TI     TI:        False   "
titre          Ti     T1:        True    ."," ";":"b"(")"["]"sb"s."s,"s:"s;""|"
T-Traduit      TT     TT:        False   "
Auteur_l       AL     AU:        True    ;"
Auteur-c       AC     AU:        True    ;"
Adresse        AD     AD:        True    ;"
Organisme      OR     AD:        True    ,";"1"2"3"4"5"6"7"8"9"0"
Source         SO     SO:        True    "
Conference     CM     CM:        False   ;"
Publisher      PU     PU:        False   "
ISBN           IB     IB:        False   "
ISSN           IS     IS:        False   "
Date           DP     PY:        True    b"
Ville          VI     AD:        True    ,";"
Pays           PA     AD:        True    ,";"
pays           Pa     AD:        False   ,";"
Pays_publi     CP     CP:        False   "
Langue         LA     LA:        False   "
BL             BL     BL:        False   "
LT             LT     LT:        False   "
Resume         AB     AB:        True    ."," ";":"b"(")"["]"sb"s."s,"s:"s;""|"
MTM            MTM    MTM:       False   b"
AI             AI     AI:        False   "
Nb_Ref         NR     NR:        False   "
Codes          CC     CC:        False   "
PC             PC     PC:        False   "
Descripteurs   DE     DE:        True    ;"
Index          IN     IE:        True    ;"
Descripteurs-F DE     DF:        False   ;"
Index-F        IN     IF:        False   ;"
Descripteurs-S DE     DS:        False   ;"
Index-S        IN     IS:        False   ;"
Journal        JN     JN:        True    "
SI             SI     SI:        False   "
LD             LD     LD:        False   "
AN             AN     AN:        False   "
CODEN          CD     CD:        False   "
Conference     CF     CF:        False   "
Corporate      CA     CA:        False   "
FIN            FIN    FIN        FIN     "
```

Le champ MTM a été ajouté, il correspond au résultat du traitement sémantique des champs en texte libre (titre et résumé) afin d'en extraire les mots composés (multi-termes) qui vont permettre une indexation à jour du corpus permettant de détecter l'innovation absente des champs d'indexation proposés par PASCAL (Descripteurs, Indicateurs, Codes, …).
Dans le descripteur de format ci-dessus, la balise True permet de travailler sur le champ associé, la balise False le masque dans tous les menus du logiciel. Pour les séparateurs, certains jockers sont utilisés : ORD_i permet d'extraire uniquement le $i^{ième}$ segment de texte du champ découpé suivant les séparateurs proposés (ORD_0 pour extraire le dernier segment), \n désigne le changement de ligne, \" le double guillemet, b le blanc, …

2.4 Répartition des articles dans le temps

Pour la période étudiée (janvier 2002 à août 2007) nous avons récupéré 5140 notices bibliographiques contenant Morroco dans le champ adresse. Ci-dessous, nous illustrons la répartition de ces documents dans le temps. L'histogramme obtenu ne doit pas être interprété sans prendre en compte le retard d'indexation inhérent à toute base bibliographique et dû au décalage entre la parution d'un article et son entrée dans la base. Pour la base PASCAL, nous estimons ce retard à environ 6 mois mais inégalement répartis entre les publications de premier plan (délai moins long) et les autres (le délai peut alors passer à plus d'un an). Le déficit constaté sur 2006 est en grande partie dû à ce phénomène de retard dans l'indexation. Celui de 2007 à 2 origines : le délai que nous venons de mentionner et le fait que l'année n'est pas complète (seuls les mois de janvier à août étant pris en compte). Une légère tendance à la baisse est à retenir sans toutes fois être alarmante.

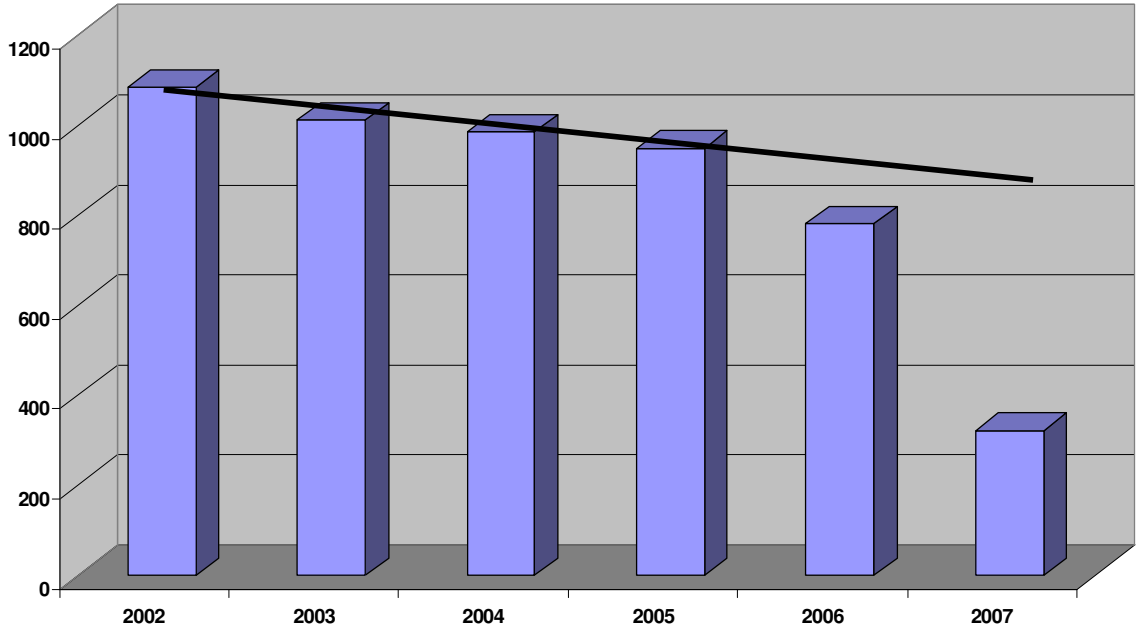


Figure 1 : évolution de la production scientifique marocaine

3 QUELQUES RESULTATS QUANTITATIFS

3.1 Le nombre de publications par auteur sur la période de référence

63 LAKHDAR-HAKIMA
43 HAMMOUTI-BELKHEIR
37 SOULAYMANI-RACHIDA
37 BENCHEKROUN-ABDELLATIF
36 CHERKAOUI-A
35 KZADRI-MOHAMED
33 HASNAOUI-MOHAMMED
32 ALLALI-FADOUA
31 HASSAM-BADREDDINE
30 ZANNOUD-MOHAMED
29 HAJJAJ-HASSOUNI-NAJIA
29 BENAMEUR-M
28 SLASSI-ILHAM
28 EL-MRINI-MOHAMED
28 AMEUR-AHMED
27 TRAFEH-M
27 RHALEM-NAIMA
27 MIRI-ADBELHAMID
27 JIRA-HASSAN
26 JIDDANE-MOHAMED
26 DAFIRI-R
26 CHAOUIR-S
25 RABII-REDOUANE
25 NOUINI-YASSINE

Tétralogie V7.0 Tableur 2D Fichier : AL-DP

		_2002	_2003	_2004	_2005	_2006	_2007
1	lakhdar1	10	8	15	15	10	5
2	hammout	3	4	10	16	10	
3	soulaym	8	8	9	9	1	2
4	benchek	19	18				
5	cherkao	5	6	7	8	8	2
6	kzadri-	1	7	5	4	10	6
7	hasnaou	5	2	8	7	10	1
8	allali-1	2		3	4	22	1
9	hassam-	6	9	5		6	5
10	zannoud	16	14				
11	benameu	6	1	8	7	5	2
12	hajjaj-1	3	1	1	5	10	1
13	slassi-	1	1	3	4	13	6
14	el-mrin	7	11	7	3		
15	ameur-a	18	7	2	1		
16	rhalem-	2	8	8	6	1	2
17	trafeh-	8	7	5	4	3	
18	miri-ad	11	9	6		1	
19	jira-ha	19	7		1		
20	chaouir	5	1	8	5	3	4
21	dafiri-	3	2	5	7	6	3
22	jiddane	6	5	4	2	7	2
23	rabii-r	7	10	5	2	1	
24	abbar-m	15	9	1			
25	nouini-	13	12				

Figure 2 : production annuelle de chaque auteur dans le tableur

La première colonne représente le nombre de publications de l'auteur, la seconde est l'identifiant de l'auteur qui a été retenu après la phase de synonymie. Pour certains des auteurs, il y a un cumul des occurrences correspondant à plusieurs formes orthographiques rencontrées dans le corpus. N'est toujours pas réglé le problème des vrais homonymes (deux personnes différentes ayant exactement le même identifiant (nom, prénom ou initiale) dans la base Pascal. Heureusement ce type de collision peut être en partie corrigé en cours d'analyse, car l'auteur en question a souvent deux casquettes (deux domaines, deux équipes de collaborateurs, deux origines, deux groupes de journaux) qui n'ont que très peu ou pas de connexion par ailleurs. Il convient alors, soit d'éliminer cet auteur bicéphale pour une analyse macroscopique, soit le différencier dans le corpus en fonction de ses connexions avec son environnement. Ce travail fastidieux représente une des limites de notre approche et il ne pourra être évité que si les bases bibliographiques s'intéressent au problème (différentiation dès la saisie des articles), donc gestion d'une base de données des homonymes détectés.

3.2 Le nombre de publications par journal

Bien que nous soyons dans une base bibliographique, le champ journal demande à être légèrement corrigé de quelques erreurs morphologiques, comme le montre le dictionnaire de synonymes suivant:

COMTES-RENDUS-MECANIQUE COMPTES-RENDUS-MECANIQUE
EUROPEAN-JOURNAL-OF-ORTHOPAEDIC-SURGERY-AND-TRAUMATOLOGIE EUROPEAN-JOURNAL-OF-ORTHOPAEDIC-SURGERY-AND-TRAUMATOLOGY

Les principaux journaux sont les suivants:

528 ESPERANCE-MEDICALE
99 MAGHREB-MEDICAL
92 JOURNAL-DE-PHYSIQUE-IV
89 ANNALES-D'UROLOGIE
72 COMPTES-RENDUS-GEOSCIENCE
56 PROGRES-EN-UROLOGIE-PARIS
53 JOURNAL-DE-RADIOLOGIE-PARIS
51 ANNALES-DE-CHIMIE-PARIS-1914
49 LES-NOUVELLES-DERMATOLOGIQUES
43 JOURNAL-OF-ETHNOPHARMACOLOGY
41 CHIRURGIE-DE-LA-MAIN
40 REVUE-DE-PNEUMOLOGIE-CLINIQUE-PARIS
38 DESALINATION-AMSTERDAM
37 LA-PRESSE-MEDICALE-1983
37 JOURNAL-OF-MATHEMATICAL-ANALYSIS-AND-APPLICATIONS
37 CAHIERS-D'ANESTHESIOLOGIE-PARIS
37 ANNALES-DE-DERMATOLOGIE-ET-DE-VENEREOLOGIE
35 LA-REVUE-DE-MEDECINE-INTERNE-PARIS
35 JOURNAL-OF-MAGNETISM-AND-MAGNETIC-MATERIALS
35 JOURNAL-FRANCAIS-D'OPHTALMOLOGIE
34 REVUE-DES-SCIENCES-DE-L'EAU-PARIS
33 REVUE-DE-STOMATOLOGIE-ET-DE-CHIRURGIE-MAXILLO-FACIALE
32 COMPTES-RENDUS-MATHEMATIQUE
31 JOURNAL-OF-COMPUTATIONAL-AND-APPLIED-MATHEMATICS
31 APPLIED-SURFACE-SCIENCE
30 SECHERESSE-MONTROUGE
30 ANALYTICAL-LETTERS
29 REVUE-NEUROLOGIQUE-PARIS
28 REVUE-DU-RHUMATISME-ED-FRANCAISE
26 MEDECINE-ET-MALADIES-INFECTIEUSES
25 MEDECINE-ET-CHIRURGIE-DU-PIED
25 JOURNAL-OF-SOLID-STATE-CHEMISTRY-PRINT
25 ACTA-BOTANICA-GALLICA
24 COMPTES-RENDUS-BIOLOGIES
23 NEURO-CHIRURGIE-PARIS
23 ACTA-ENDOSCOPICA
22 TUNISIE-MEDICALE
22 ARCHIVES-DE-PEDIATRIE-PARIS
22 ANNALES-FRANCAISES-D'ANESTHESIE-ET-DE-REANIMATION

4 EQUIPES DE RECHERCHE

4.1 Problèmes morphologiques du champ Auteurs

Dans la majorité des bases bibliographiques, le champ Auteur est très souvent source d'erreurs (homonymes, fautes d'orthographe dans les noms, prénoms entiers ou simples initiales, inversions entre le nom et le ou les prénoms, inversions de lettres, doublement de lettres, pollutions de tous ordres : éditeurs, préfaceurs, directeurs, traducteurs, …). Pour toutes ces raisons, il est nécessaire d'envisager un nettoyage puis un traitement morphologique poussé afin de déterminer les correspondances les plus vraisemblables. Un dictionnaire de synonymes est issu de ce traitement, il doit être validé avant d'être utilisé. Voici un exemple des correspondances potentielles détectées par Tétralogie dans le corpus (Maroc/PASCAL).

ADIL-N	AADIL-NADIA
AADIL-N	AADIL-NADIA
AALLOULA-EH	AALLOULA-EL-H
AALLOULA-E-H	AALLOULA-EL-H
AMAROUCH-N	AAMAROUCH-N
ASSIF-E	AASSIF-E
AASSIF-E-H	AASSIF-EL-HOUCEIN
AATIQ-A	AATIQ-ABDERRAHIM
ABABOU-A	ABABOU-ADIL
ABBAD-A	ABBAD-ABDELAZIZ
ABBAD-M	ABBAD-MOHAMMED
ABAD-M	ABBAD-MOHAMMED

ABBAR-MOHAMED	**ABBAR-MOHAMMED**
ABBAR-M	**ABBAR-MOHAMMED**
ABBASSI-M	ABBASSI-MOHAMED
ABASSI-O	ABBASSI-OMAR
ABBASSI-O	ABBASSI-OMAR
ABOUD-Y	ABBOUD-Y
ABBOUDI-M	ABBOUDI-MOSTAFA
ABDALLAOUI-F	ABDALLAOUI-FAIZA
ABDELGHAFFAR-H	ABDELGHAFFAR-HOURIA
ABDELHAK-M	ABDELHAK-MBAREK
ABDELIJALIL-EL-KHOLTI	ABDELJALIL-EL-KHOLTI
ABDELLAH-L	ABDELLAH-LAMINE

Les synonymies conservées sont alors prises en compte dans le calcul des fréquences de publication et dans toute matrice de croisement sur les auteurs. Ci-dessous, la liste des auteurs les plus prolifiques de ces 6 dernières années.

63	LAKHDAR-HAKIMA	29	HAJJAJ-HASSOUNI-NAJIA	26	DAFIRI-R
43	HAMMOUTI-BELKHEIR	29	BENAMEUR-M	26	CHAOUIR-S
37	SOULAYMANI-RACHIDA	28	SLASSI-ILHAM	25	RABII-REDOUANE
37	BENCHEKROUN-ABDELLATIF	28	EL-MRINI-MOHAMED	25	NOUINI-YASSINE
36	CHERKAOUI-A	28	AMEUR-AHMED	25	**ABBAR-MOHAMMED**
35	KZADRI-MOHAMED	27	TRAFEH-M	24	KANJAA-NABIL
33	HASNAOUI-MOHAMMED	27	RHALEM-NAIMA	23	SENOUCI-KARIMA
32	ALLALI-FADOUA	27	MIRI-ADBELHAMID	23	MOUSSAOUI-DRISS
31	HASSAM-BADREDDINE	27	JIRA-HASSAN	23	FAIK-M
30	ZANNOUD-MOHAMED	26	JIDDANE-MOHAMED		

4.2 Détection des équipes et de leurs relations

Pour cela, nous croisons les auteurs entre eux afin d'obtenir une matrice de cooccurrences qui sera filtrée, décomposée en classes connexes, elles mêmes triées par blocs diagonaux afin de faire apparaître la colonne vertébrale (équipes fortement structurées) de chaque classe. Le graphe global (plusieurs milliers d'auteurs) n'est pas manipulable, ni réellement utile, puisqu'il cumule des disciplines souvent très éloignées. Par contre, des clusters très marqués apparaissent sur le zoom de cette matrice correctement triée par blocs, en voici un exemple parmi d'autres.

Afin de montrer tout de même la struture de la recherche marocaine dans sa globalité, nous avons appliqué une simplification au graphe initial : nous n'avons gardé que les auteurs ayant 5 publications ou plus (soit une par an au moins) et nous avons négligé les liens à 1 (une seule publication en commun avec un autre auteur). Le graphe comporte alors 1000 sommets et il peut être dessiné sans difficulté. On y remarque des équipes bien structurées (celle se trouvant sous forme de blocs diagonaux dans le zoom de la matrice), par contre les liens sont au moins de 2 publications (sinon le graphe n'est plus lisible).

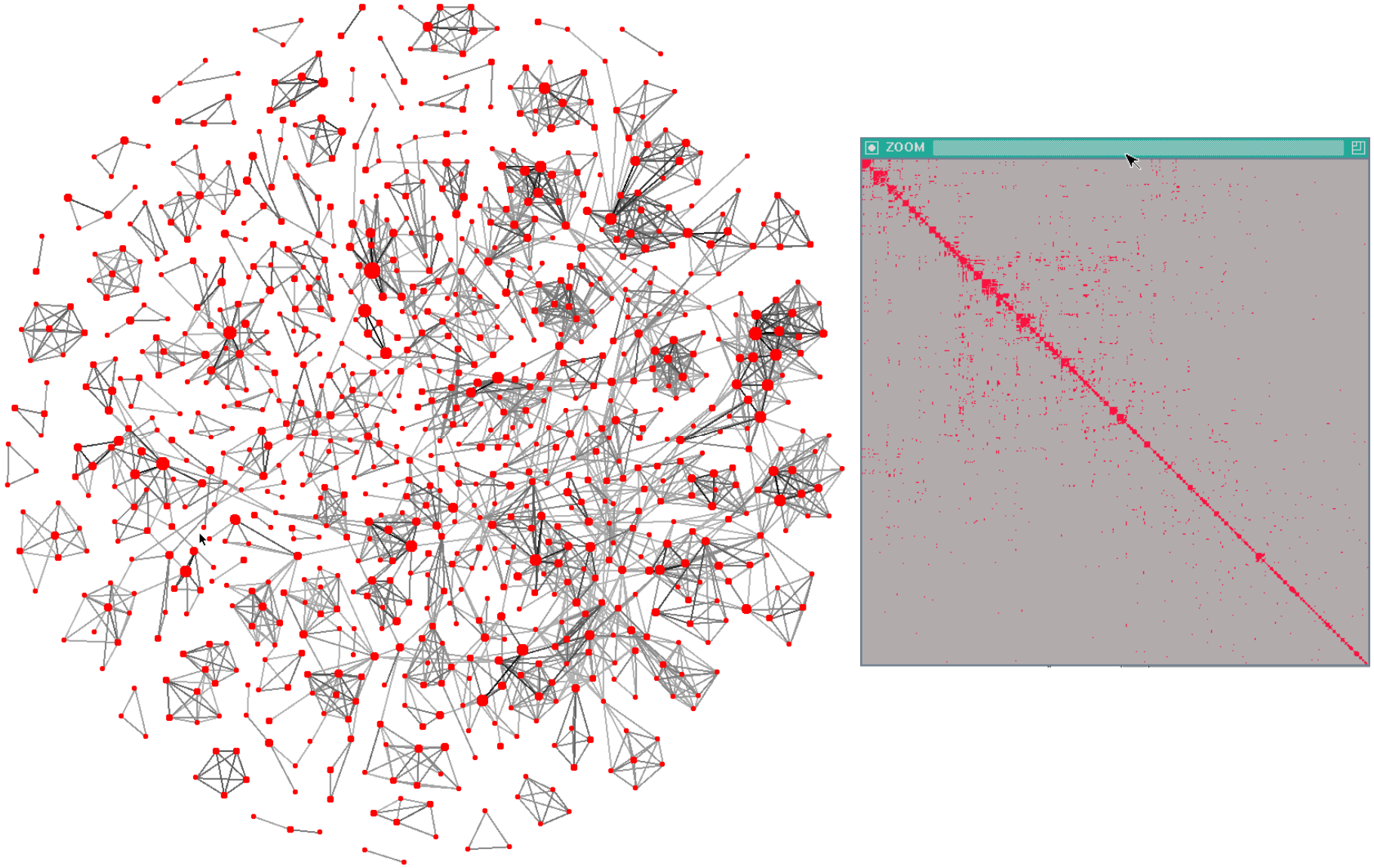


Figure 3 : Graphe et zoom de la matrice AU X AU traduisant les structures et les connexions des équipes de recherche

5 ANALYSE DES COLLABORATIONS INTERNATIONALES

5.1 Difficultés de l'étude sur les pays

Les pays sont présents dans le champ Adresse (AD:) mais ils doivent être détectés parfois de façon indirecte
Par exemple :

- Guadeloupe et Martinique apparaissent sans la France
- Un état américain apparaît sans Etats-Unis
- Moroco pour Maroc et non pas Morroco, …
- On ne peut pas « couper » au point car il fait partie de certains Pays : U.S.A, U.K., …

Ces erreurs ou omissions sont corrigées par Tétralogie grâce, à nouveau, à des dictionnaires de synonymes :

AFGANISTAN	AFGHANISTAN
AFRIQUE DU SUD	SOUTH-AFRICA
ALABAMA.	USA
ALABAMA	USA
ALBANIA.	ALBANIA
ALBANIE	ALBANIA
ALBANIE.	ALBANIA
ALEMANIA	GERMANY
ALABAMA.	USA

ALABAMA	USA
ALBANIA.	ALBANIA
ALBANIE	ALBANIA
ALBANIE.	ALBANIA
ALEMANIA	GERMANY
ALGERIA.	ALGERIA
ALGERIE	ALGERIA
ALLEMAGNE	GERMANY
AL USA.	USA

AL USA	USA
ANGLETERRE	UK
ANTIGUA AND BARBUDA	
ANTIGUA-BARBUDA	
ARABIA	SAUDI-ARABIA
ARABIA SAUDITA	SAUDI-ARABIA
ARABIE-SAOUDITE	SAUDI-ARABIA
ARG	ARGENTINA
ARG.	ARGENTINA

Le champ « Adresse » ainsi synonymé est ensuite filtré pour ne garder que des noms de pays valides. Pour cela nous utilisons un dictionnaire de pays préétabli qui permettra ensuite de dresser des cartes géostratégiques, nous en donnons ci-dessous le début :

AFGHANISTAN
ALBANIA
ALGERIA
ANGOLA
ANTIGUA-BARBUDA
ARGENTINA
ARMENIA
AUSTRALIA
AUSTRIA
AZERBAIJAN
BAHRAIN
BANGLADESH
BARBADOS
BELARUS
BELGIUM
BELIZE
BENIN
BHOUTAN
BOLIVIA
BOSNIA
BOTSWANA
BRAZIL
BRUNEI
BULGARIA
BURKINA-FASO
BURUNDI
CAMBODGE
CAMEROON
CANADA
CTRL-AFRICAN-REP
CHAD
CHILE
CHINA
COLOMBIA
CONGO
CONGO-PEOPL-REP

5.2 Répartition des pays pour l'ensemble du corpus

La France représente plus de la moitié des collaborations internationales du Maroc. Il faut tempérer cette affirmation car la base Pascal utilisée est d'origine Française (INIST-CNRS), d'où un biais qui suivant les disciplines peut être plus ou moins important. Par expérience, nous pouvons dire qu'il s'agit malgré tout d'une tendance lourde et la place de premier collaborateur ne peut en aucun cas être remise en cause. Nous voyons apparaître dans le top 12 deux pays du Maghreb : la Tunisie puis l'Algérie.

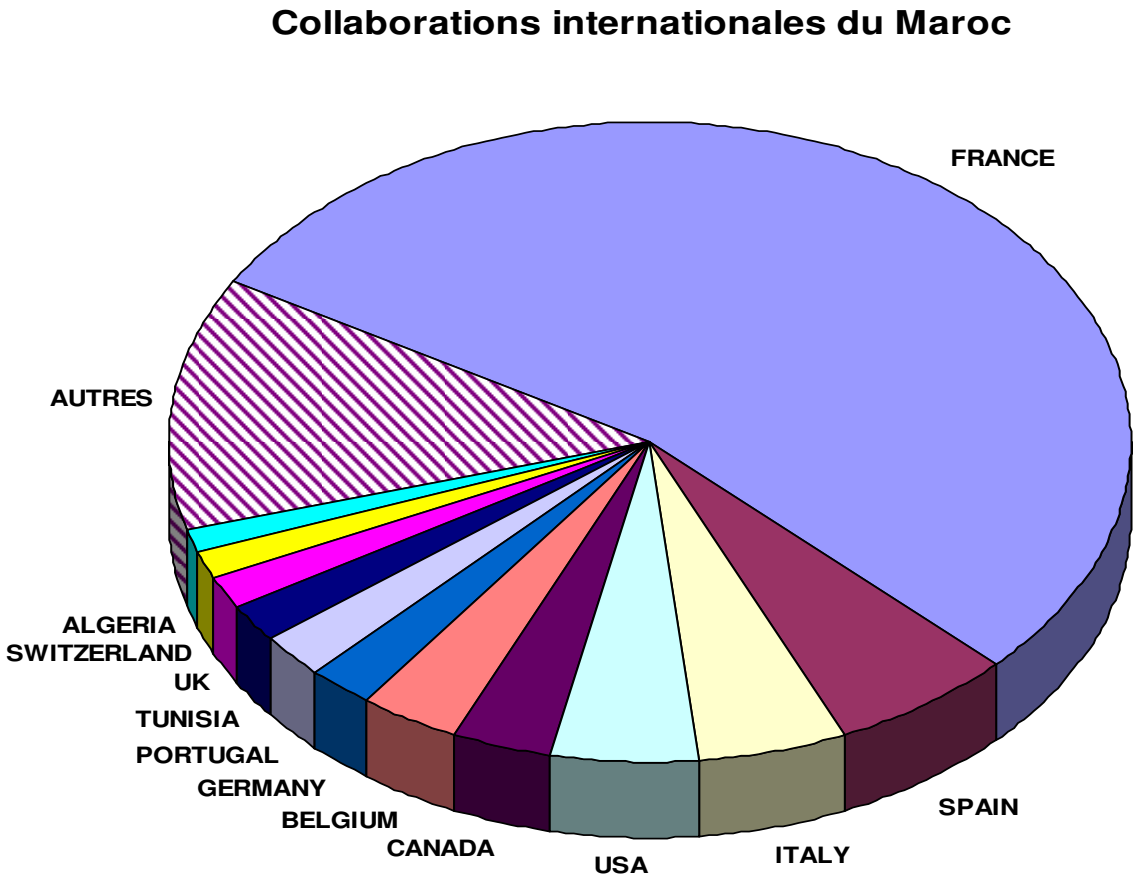


Figure 4 : répartition des collaborations internationales du Maroc pour les 6 dernières années

5.3 Cartes cumulant puis comparant les trois périodes retenues

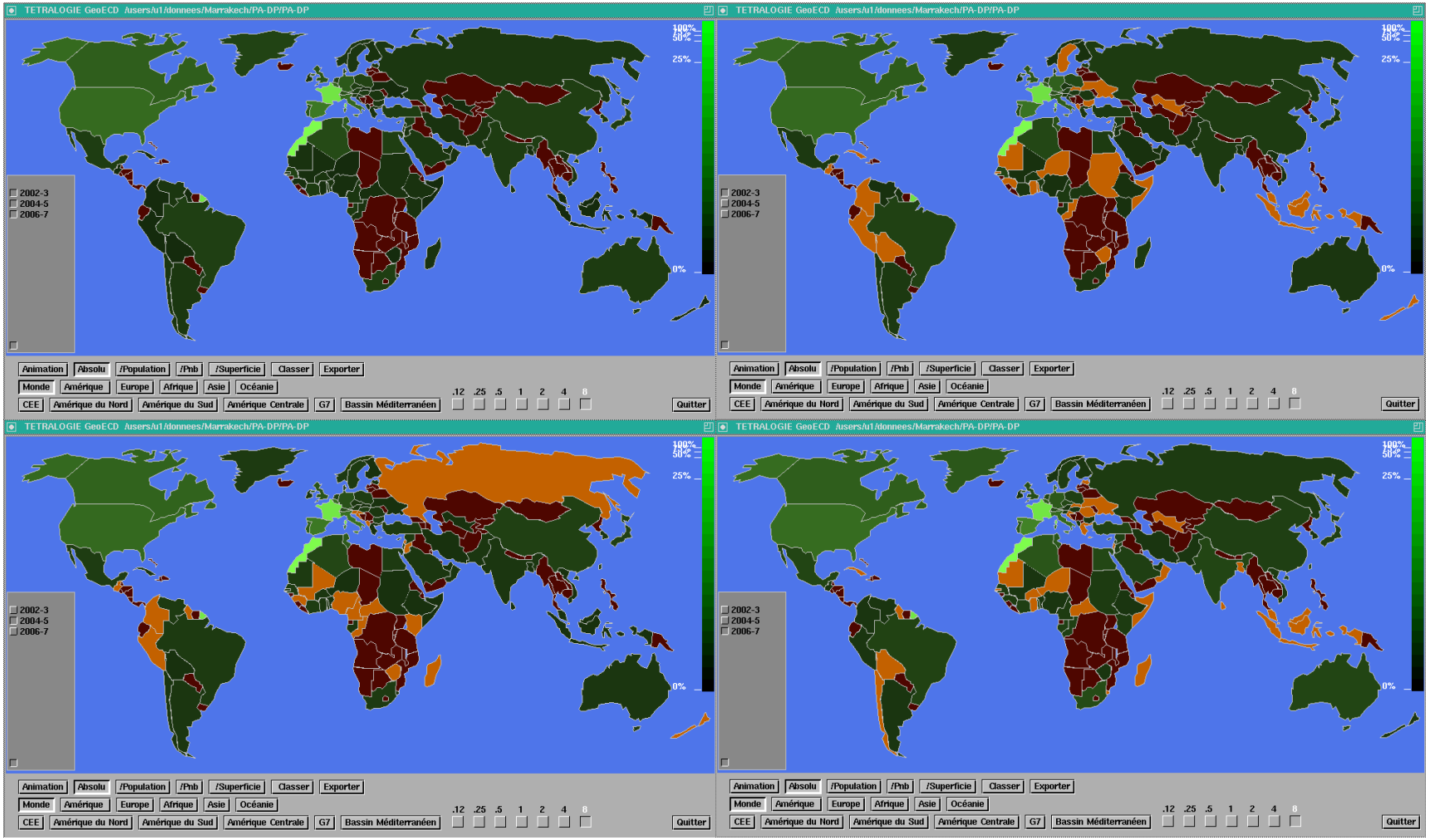


Figure 5 : Cartes illustrant l'évolution des collaborations internationales du Maroc sur les 6 dernières années.

Nous utilisons, ici, le logiciel graphique GéoECD [S. KAROUACH] qui permet de générer automatiquement des cartes géographiques interactives à partir des tableaux de données produits par Tétralogie. Ces cartes sont entièrement manipulables par tout utilisateur local ou distant et permettent donc, dans des temps très

courts, de communiquer des informations géostratégiques via le réseau avec la possibilité de les retravailler à distance. Nous avons ici décomposé le temps en 3 périodes de 2 ans (2002-3, 2004-5, 2006-7). Nous présentons, sous forme de 4 cartes, les résultats obtenus globalement et ensuite suivant les 3 périodes précédemment définies.
Il ressort de cette analyse que les collaborations internationales sont assez fluctuantes, notamment avec le reste de l'Afrique, l'Europe centrale, l'Amérique centrale et l'Amérique du sud ainsi que l'Indonésie et la Nouvelle Zélande. Pour les autres pays, il y a une certaine stabilité comme avec la France, les USA, l'Espagne ou la Chine.

6 LES SIGNAUX FORTS DE LA RECHERCHE MAROCAINE

6.1 Exploitation du champ descripteurs

Le champ descripteur est édité en 3 langues (Anglais, Français et Espagnol), mais les descripteurs anglais sont sans contestation les plus nombreux. Nous avons donc décidé de ne travailler que sur les descripteurs anglais. Voici les plus fréquemment rencontrés dans le corpus :

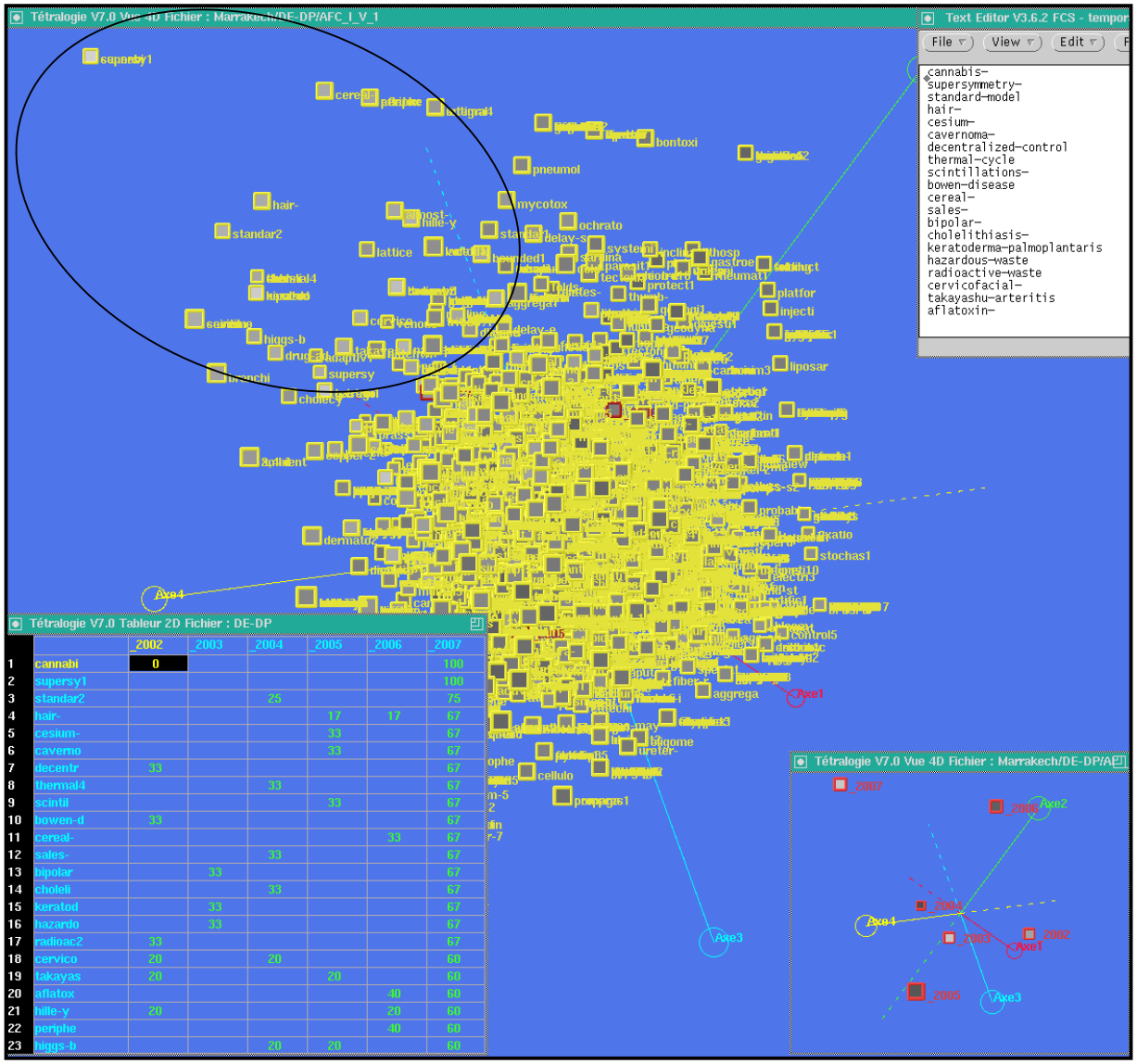
1154 HUMAN-
1047 MOROCCO-
861 CASE-STUDY
667 TREATMENT-
436 EXPERIMENTAL-STUDY
423 DIAGNOSIS-
362 REVIEW-
243 TROPICAL-MEDICINE
225 CHILD-
172 ADULT-
168 MODELING-
159 COMPLICATION-
158 FEMALE-
145 SURGERY-
141 XRD-
136 SYMPTOMATOLOGY-
133 THEORETICAL-STUDY
121 PROGNOSIS-
118 COMPUTERIZED-AXIAL-TOMOGRAPHY
113 EPIDEMIOLOGY-
109 MALE-
103 CRYSTAL-STRUCTURE
98 CHEMOTHERAPY-
90 PLANT-ORIGIN
88 CHEMICAL-COMPOSITION
85 MEDICINAL-PLANT
85 ANIMAL-
82 AQUEOUS-SOLUTION
81 TOXICITY-
80 TUBERCULOSIS-
80 COMPARATIVE-STUDY
78 PUBLIC-HEALTH
75 BONE-
74 RAT-
74 PHARMACOGNOSY-
74 MODELS-
74 ECHOGRAPHY-
73 CLINICAL-MANAGEMENT
73 CHRONIC-
70 BIBLIOGRAPHIC-REVIEW

6.2 Exploitation du champ Indentifieurs

Ce champ est aussi présent dans les 3 langues, pour les mêmes raisons que précédemment, nous ne l'exploiterons que dans sa version anglaise. Afin d'illustrer une autre méthode d'analyse, nous allons réaliser un croisement entre les indentifieurs, ceci nous conduit à une matrice carrée IE x IE. Cette matrice, après avoir été transformée en équivalence par division de chaque élément par la racine carrée de ses éléments diagonaux (diagonale unitaire), est triée par blocs afin de faire apparaître des clusters sémantiques correspondant aux axes de recherche du Maroc. Ceux-ci sont ensuite isolés et listés afin d'en connaître la teneur.

6.3 SIGNAUX FAIBLES

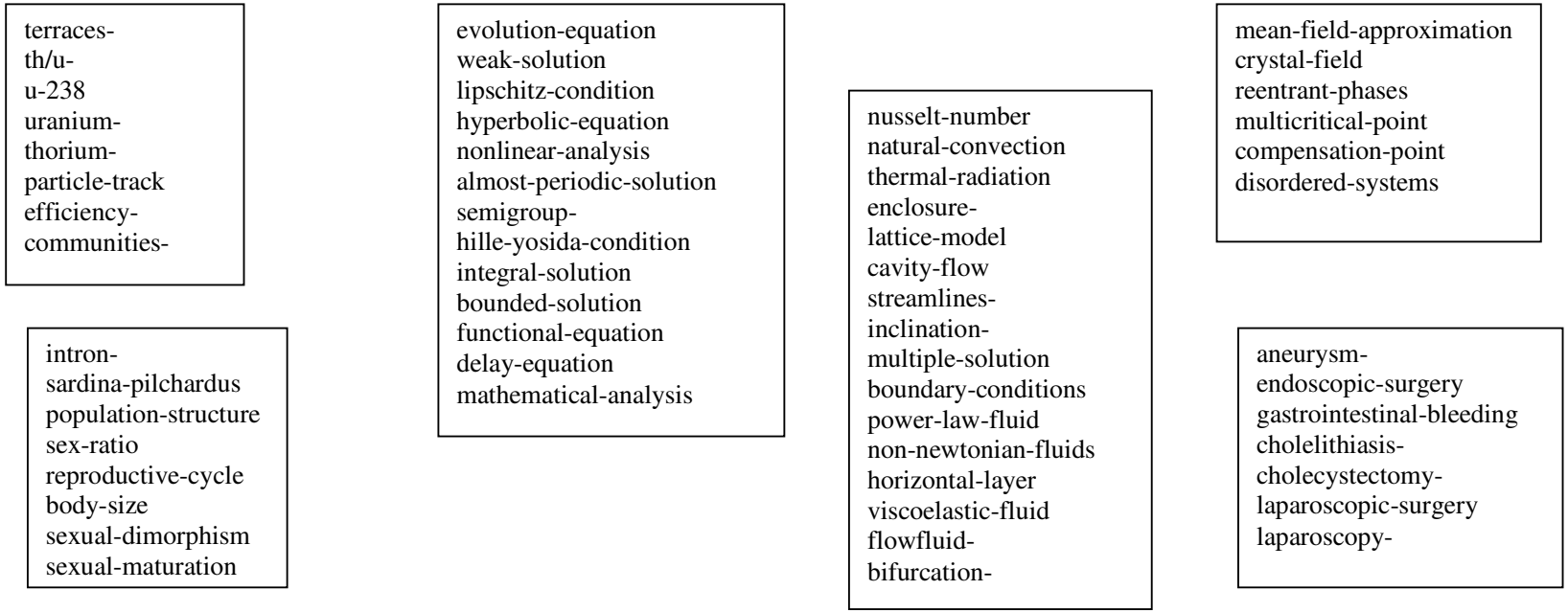
Afin de faire apparaître les signaux faibles, nous réalisons, dans un premier temps, une analyse factorielle des correspondances (AFC) sur la matrice qui croise les descripteurs anglais (les plus nombreux) et le temps exprimé en années : DE x DP. Une visualisation de la carte factorielle des années (DP) permet d'isoler 2007 dans un coin de la carte (ici en haut à gauche), une exportation de l'azimut ainsi déterminé vers la carte des descripteurs anglais (DE) permet de détecter tous ceux qui sont typiques de 2007 et ce, quelque soit leur nombre. Une capture à la souris permet ensuite d'en extraire la liste qui est récupérée dans un filtre. Dans un second temps, il est possible de croiser ces termes émergents entre eux afin de regarder, comme précédemment, s'ils s'organisent en clusters significatifs. Si c'est le cas, on a détecté les signaux faibles du moment. Chaque cluster est ensuite listé et croisé avec les autres variables de la base (auteurs, laboratoires, pays, descripteurs non émergents, journaux, …) afin d'en valider la pertinence. En effet, si un laboratoire prestigieux ou une équipe de renom est lié à un signal faible, celui-ci est digne d'intérêt. Idem s'il s'agit de journaux de premier plan ou de pays donc la recherche est réputée. Bien souvent les experts sont déstabilisés par ce type d'information, car ils ne sont, bien entendu, ni à l'origine de cette recherche, ni dans le petit cercle d'initiés qui, éventuellement, est au courant de son existance. La seule façon de les convaincre est de leur montrer en montrer l'origine.



		2002	2003	2004	2005	2006	2007
1	cannabi	0					100
2	supersy1						100
3	standar2			25			75
4	hair-				17	17	67
5	cesium-				33		67
6	caverno				33		67
7	decentr	33					67
8	thermal4			33			67
9	scintil				33		67
10	bowen-d	33					67
11	cereal-					33	67
12	sales-			33			67
13	bipolar		33				67
14	choleli			33			67
15	keratod		33				67
16	hazardo		33				67
17	radioac2	33					67
18	cervico	20		20			60
19	takayas	20			20		60
20	aflatox					40	60
21	hille-y	20				20	60
22	periphe					40	60
23	higgs-b			20	20		60

Figure 6 : AFC sur l'évolution des descripteurs et présence relative par année

Voici un extrait des signaux faibles obtenus en croisant les descripteurs en émergence (période 2006-07). La matrice obtenue est triée par blocs diagonaux (en mode relatif) et chaque cluster apparaissant nettement sur la diagonale est validé, s'il intervient dans plusieurs documents (cooccurrence > 1).



CONCLUSION

Nous avons volontairement limité cette étude pour pouvoir en présenter les méthodes et résultats dans un temps très limité. Le corpus constitué à partir de la base PASCAL peut encore livrer de nombreuses informations utiles, soit synthétiques, soit plus ciblées comme par exemple sur un laboratoire, une équipe ou un domaine de recherche précis. Comme ces données sont en ligne sur le serveur tetralogie.irit.fr, il est possible d'affiner à distance (par une connexion ssh) l'ensemble des analyses déjà produites et d'en réaliser d'autres notamment à partir des champs non encore utilisés. Ce type d'analyse nous est régulièrement demandé par des grands laboratoires de recherche afin de caller au mieux leur politique à long terme : collaborations internationales, suivi des sujets porteurs, détection de nouveaux axes (signaux faible), évaluation de la recherche, gestion des abonnement aux revues, recherche de partenaires, mise à jour de l'indexation, facteur d'impact, cartographie des connaissances, aide à la mise au point d'ontologies du domaine, recherche d'information à partir des cartes sémantiques. Les applications sont nombreuses, mais ce qui compte c'est que sous un même formalisme et avec des traitements bien maîtriser, il est possible d'envisager l'étude de pratiquement toutes les sources d'information électronique et même de les combiner pour s'approcher de l'exhaustivité et éviter certains biais des études mono source.

7 BIBLIOGRAPHIE

[1] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret, *"Information mining: use of the document dimensions to analyse interactively a document set. "* 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

[2] J.-L. Multon, G. Lacombe, B. Dousset, *"Analyse bibliométrique des collaborations internationales de l'INRA"*. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

[3] B. Dousset, S. Karouach, *"Collaboration interactive entre classifications et cartes thématiques ou géographiques. "* 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.

[4] J.-L. Multon, G. Lacombe, B. Dousset, *"Analyse bibliométrique des collaborations internationales de l'INRA"*. 9ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, 14-18 octobre 2002.

[5] J. Mothe, C. Chrisment, B. Dousset, S. Karouach, *"Représentation des documents textuels : étude d'un domaine à travers des publications"*. 5ième Congrès de la société française de recherche opérationnelle et d'aide à la décision ROADEF. (Avignon France), pp 130-131, 26-28 février 2003.

[6] S. Karouach, B. Dousset, *"Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grande taille"*. Workshop sur la recherche d'information : un nouveau passage à l'échelle, associé à INFORSID'2003, (Nancy France), 3-6 juin 2003.

[7] J. Mothe, C. Chrisment, B. Dousset, J. Alaux, *"DocCube : multi-dimensional visualization and exploration of large document sets*". Journal of the American Society for Information Science and Technologiy JASIST, Special topic section: "Web Retrieval and Mining". Guest Editor: Hsinchun Chen, 54(7), pp 650-659, March 2003.

[8] S. Karouach, B. Dousset, *"Analyse d'information relationnelle par des graphes interactifs de grandes tailles*". 4ièmes journées d'EGC (Extration et Gestion de Connaissances), Clermont Ferrand, 20-23 janvier 2004.

[9] C. Chrisment, B. Dousset, S. Karouach, J. Mothe, *"Information mining : extracting, exploring and visualising geo-referenced information*". Workshop on Geographic Information Retrieval, SIGIR 2004.