

UNIVERSITE DE DROIT ET DES SCIENCES D'AIX-MARSEILLE
Faculté des Sciences et Techniques de Saint Jérôme

LA SERIATION DES SIMILARITES SPECIFIQUES : OUTIL POUR LA RECHERCHE DE L'INFORMATION STRATEGIQUE

*Une méthode de classification automatique de
l'information issue des bases de données en Veille
Technologique*

THESE

Présentée et soutenue publiquement par :

BALDIT Patrick

Ingénieur ESCM
Inghieur ESIPSOI

le

pour obtenir le grade de
Docteur Ingénieur

Spécialité :

Sciences de l'information et de la communication

Président de Jury : Marcotorchino F. (Professeur Paris VI)

Autres membres du jury :
Paoli C. (Professeur Marne La Vallée)
Ruiz J.M. (Professeur ENSSPICAM)
Dou H. (Professeur CRRM)
Cornu L. (Professeur Université Aix-Arseille II)
Quoniam L. (Maître de Conférence CRRM)

Rapport concernant le mémoire de doctorat présenté par Patrick Baldit.

La sériation des similarités spécifiques; Outil pour la recherche de l'information stratégique.

Le mémoire présenté comprend trois parties.

Le premier chapitre est consacré à l'analyse du contexte de la réalisation du candidat. Cette analyse du contexte Veille Technologique et Intelligence Economique est elle même replacée dans l'environnement le plus global de la gestion de projets. Le vocabulaire est ici bien défini ainsi que les concepts, et les méthodes. Cette partie replace notamment les activités de Bibliométrie pour le traitement de corpus volumineux afin de valoriser les références bibliographiques dans le cadre d'activités de Veille Stratégiques. Le candidat montre comment l'analyse des informations accessibles sur les réseaux à hauts débits peut aider à la recherche d'innovations indispensables aux activités de Veille Technologiques et Intelligence Economique. Cette partie fait un point sur la production d'information stratégique pour l'aide à la décision.

Le deuxième chapitre rappelle l'environnement technique de la réalisation du candidat. Les lois bibliométriques sont rappelées ainsi que les différentes méthodes d'analyse de données. Il introduit la nécessité d'outils mieux adaptés aux données à traiter. Cette adaptation va dans le sens de plus gros volumes traités et de respects de propriétés spécifiques. Ce chapitre montre également l'intégration avec les outils déjà existant dans au CRRM où le candidat a fait sa thèse.

La troisième partie constitue l'essentiel du mémoire. Elle présente le produit DATABLOC développé au CRRM dans le cadre de ce doctorat. La présentation va du cadre théorique de la méthode de sériation par blocs aux applications en passant par une présentation du produit réalisé. Les capacités de traitement de ce nouveau produit et son interface conviviale ouvre des perspectives importantes en matière de traitement de grandes bases de données bibliographiques. L'application présentée, au niveau de l'exploitation analytique d'enquêtes par interview, montre que le candidat sait interconnecter des domaines apparemment éloignés et comment cet outil peut servir d'aide à la lecture de textes.

L'exposé, de la démarche, est très clairement exprimé et le manuscrit est bien structuré.

Cette thèse a le grand mérite de replacer des outils mathématiques dans un contexte global, mais surtout de les valider par la réalisation d'un logiciel opérationnel et par son application à des cas concrets.

L'ensemble des travaux effectués par Patrick Baldit mérite sans aucun doute d'être présenté en vue de l'obtention d'un doctorat.

Fait à Marseille le 26/11/1994

Lucienne CORNU
Professeur à l'université Aix-Marseille II

**A Maité mon épouse
et Benjamin mon fils**

Je tiens à remercier en tout premier lieu mon laboratoire d'accueil, le Centre de Recherche Rétrospective de Marseille, et plus particulièrement Monsieur **Henri Dou** qui a su mettre à ma disposition un matériel performant et des conditions de travail propice au déroulement optimal de mes travaux.

Je tiens également à saluer ici **Christophe RETOURNA** et **Raimundo DOS SANTOS**, qui m'ont donné à maintes reprises l'occasion d'étayer mon travail notamment dans la phase de mise au point du logiciel DATABLOC.

Je remercie l'ensemble du laboratoire d'ingénierie des projets industriels en la personne de **Jean Michel RUIZ**, qui m'a permis d'entreprendre des développements concernant des applications hypertextes en enseignement assisté par ordinateur

Je remercie également les **membres du jury** et les **rapporteurs** de ma thèse de bien avoir voulu consacrer leur temps à la lecture de ce mémoire

Je ne pourrais terminer sans remercier **Luc Quoniam**, sans lequel ce travail ne serait pas ce qu'il est, et en qui j'ai trouvé une disponibilité de **tous les instants**. J'ai pu apprécier en lui une compétence et un soutien moral qui m'ont été précieux. Qu'il en soit ici chaleureusement remercié.

Je garderais enfin une pensée pour mon épouse Maité et mon fils Benjamin qui m'ont permis par leur collaboration de pouvoir achever cette thèse dans de bonnes conditions.

Résumé de la thèse

Ces dernières années ont vu l'émergence du concept d'intelligence économique sensibilisant les décideurs publics ou privés à se préoccuper des modifications de **leur** environnement. Depuis une trentaine d'années, au niveau de la planète, le stockage informatique de données couvrant un vaste domaine de connaissance s'est généralisé. La problématique s'est orientée vers une exploitation rationnelle de ce gisement mondial dans le but d'en retirer des informations pertinentes et stratégiques. L'utilisation de la bibliométrie comme outil de traitement dans le cadre de la veille technologique s'est imposée pour permettre d'établir des grilles de lecture de documents primaires pour des experts du domaine. Les distributions spécifiques de ce type d'information rendent l'utilisation des techniques statistiques usuelles difficilement exploitables et nous ont conduit à la création d'un algorithme de traitement adapté, permettant une classification non hiérarchique par optimisation d'un critère global, que nous avons appelé la sériation des similarités spécifiques. L'objectif final du traitement statistique étant de fournir aux experts des représentations cartographiques de l'information recueillie, le développement de visualisation hypertexte s'est avéré primordial pour une exploitable conviviale.

Mots Clefs

Information Stratégique, Analyse de Donnée, Veille Technologique, Base de Données, Indices d'association, Brevets, Classification Automatique, Bibliométrie, Bloc Sériation, Classification Non Hiérarchique.

Sommaire

Sommaire

I. Introduction	12
II. L'information stratégique	18
A. Le monde industriel en quête d'innovation	19
1. Les constantes de temps diminuées	19
2. Les différents types d'innovation	21
a. La nature de l'innovation	21
b. Les degrés de l'innovation	22
B. L'innovation par l'information stratégique	26
1. Les tendances du marché	26
2. L'information scientifique et technique	27
3. L'importance du brevet	28
C. Le marché de l'information	30
1. Les différents types d'information	30
a. L'information de type firme	30
b. L'information de type consultant	31
c. L'information de type foire et salon	31
d. L'information de type texte	31
2. De l'information formelle aux bases de données informatiques	32
3. Les réseaux de communication	34

D. La gestion de projet et la veille technologique :	37
Vers une méthodologie commune	
1. La veille technologique : Un domaine récent	38
2. La Gestion de Projet (GP) et la Veille Technologique (VT)	42
3. Le cadre logique : un outil de détermination	41
des facteurs critiques de succès	
4. La structure de surveillance sectorielle systématique :	53
une structure de projet hybride	
E. La bibliométrie : une aide à la détection d'information stratégique	58
1. Traitement d'une grande masse de références	58
2. Structure des références bibliographiques	60
3. Valorisation de l'information par la bibliométrie	62
III. La bibliométrie et l'analyse de données	63
A. La bibliométrie : Pour une veille technologique efficace	64
1. Historique	65
2. Pré-travail sur le corpus	67
3. Production d'indicateurs	69
4. Avantages des références brevets pour les traitements	73
bibliométriques	

6. Modélisations des distributions bibliométriques	74
1. Loi de Bradford	74
2. Loi de Lotka	77
3. Loi de Zipf	78
4. Propriétés	81
C. Les profils de présence/absence	84
1. Le profil de formes	84
2. Les indices d'association	87
3. De la bibliométrie à la classification	95
IV. La Sériation des Similarités Spécifiques	100
A. Les méthodes d'analyse de données en classification	101
1. Les méthodes factorielles	102
a. L'Analyse en Composantes Principales	102
b. L'Analyse Factorielle des Correspondances	103
2. Les classifications hiérarchiques	104
3. Les classifications non hiérarchiques	107
4. Les Méthodes de blocs sériations	108
a. Les méthodes de sériations unidimensionnelles	108
b. Les méthodes de sériation par bloc	109

B. La sériation des similarités spécifiques	112
1. Position générale du problème	112
2. Rappel de l'algorithme de Coupet Marcotorchino et Parisot (CMP)	118
3. Algorithme 3S	122
4. Synthèse de la classification en quadrant.	133
C. Présentation du logiciel	136
1. Les options	137
2. Les résultats	139
a. Fichiers compatibles SGBD et GED	140
b. Fichiers compatibles tableur	142
c. Sortie hypertexte	142
V Cas étudiés	149
1. "Mise en évidence de concepts dans un discours oral par une nouvelle méthode d'indexation basée sur des critères statistiques." Les organisations aux risques de l'information, 11ème journée des IUT de la recherche en Sciences Sociales, Toulouse, 17-18 Mars 1993	151
2. "Bibliometric analysis of patent documents for R&D management" Research Evaluation, vol 3, n°1, Avril 1993, p 13-18	172
VI Conclusion	179
VII Bibliographie	182

Liste des figures

Chapitre II : L'information stratégique

Figure II. 1	: Les différents types d'innovation (Selon Broustail et Fréry)	24
Figure 11.2	: Les différents types d'information (Selon Hunt)	32
Figure II.3	: La chaîne de l'information automatisée	33
Figure II.4	: Les réseaux internationaux (Selon J.Chaumier)	36
Figure II.5	: Le cadre logique	SI
Figure II.6	: Le cadre logique pour un projet de veille technologique	52
Figure II.7	: Les réseaux de spécialistes (Selon Jakobiak)	55
Figure II.8	: Les différentes structures d'un projet	56
Figure II.9	: Les champs d'une référence d'un brevet de la base WPIL	61

Chapitre III : La bibliométrie et l'analyse de données

Figure III.1	: Formes liées à un champ pour le délimiteur “;”	67
Figure III.2	: Formes liées à un champs pour les délimiteurs ” ” et “;”	68
Figure III.3	: Classification des indicateurs en bibliométrie (Selon Vinkler)	72
Figure III.4	: Données expérimentales collectées par Bradford	75
Figure III.5	: Graphe construit lors de l'étude de Bradford	76
Figure II 1.6	: Distribution des auteurs sous forme d'histogramme en fonction du nombre de leurs publications	77
Figure III.7	: Valeurs étudiées par Zipf pour le livre Ulysse de Joyce	79
Figure III.8	: Représentation graphique de la loi de Zipf	80
Figure III.9	: Représentation gaussienne de la répartition des tailles dans une population adulte	82
Figure III. 10	: Distribution Zipfienne	83
Figure III.1 1	: Profil de forme j et de référence i	84
Figure III.12	: Totaux marginaux de lignes et de colonnes	85
Figure III.13	: Profil d'indexation	86
Figure III.14	: Distribution des totaux marginaux suivant les deux espaces	87

.Figure III.15 : Matrice de contingence des co-occurrences de 1 pour deux formes i et i'	32
Figure III.16 : Matrice des valeurs de l'indice d'association utilisé pour établir des proximités entre formes	35
Figure III.17 : Matrice de comptabilisation d'occurrence de mots clefs	36
Figure III.18 : Matrice de présence de formes de plusieurs champs	38
Figure III.19 : Classification des mots clefs et des auteurs simultanément	38
 Chapitre IV : La Sériation des Similarités Spécifiques	 100
Figure IV. 1 : Matrice de présence/absence ou tableau disjonctif complet	101
Figure IV.2 : Une classification hiérarchique	105
Figure IV.3 : Distribution du nombre de formes d'un descripteur ayant une fréquence X	116
Figure IV.4 : Sériation par bloc associé à la distribution de la figure IV.3	117
Figure IV.5 : Détermination de la classe dominante	126
Figure IV.6 : Zone de calcul des paramètres PkR et PkF	127
Figure IV.7 : Cas d'une forme triviale à faible fréquence	128
Figure IV.8 : Résultat schématique d'une classification des similarités spécifiques	129
Figure IV.9 : Matrice de départ	130
Figure IV.10 : Matrice sériée par la méthode 3S au moyen de DATABLOC	131
Figure IV.11 : Représentation schématique de la sériation des formes triviales	132
Figure IV.12 : Représentation en quadrant	433
Figure IV.13 : Place de DATABLOC dans la chaîne de traitement bibliométrique	A36
Figure IV.14 : Ecran de démarrage du logiciel DATABLOC	137
Figure IV. 15 : Choix de la matrice binaire à analyser	138
Figure IV. 16 : Paramétrage du calcul de la sériation	138
Figure IV.17 : Présentation de l'analyse statistique	139
Figure IV.? 8 : Exemple de références enrichies de champs supplémentaires	141
Figure IV.19 : Exemple de présentation de matrice sous EXCEL	143

Figure IV.20 : Constituant des classes sous forme hypertexte en mode index	144
Figure IV.21 : Constituant des classes sous forme hypertexte en mode sommaire	145
Figure IV.22 : Référence bibliographique sous forme hypertexte	146
Figure IV.23 : Recherche de formes par le bouton Rechercher	147
Figure IV.24 : Historique de la consultation	147