

# THÈSE DE DOCTORAT

présentée par

**Philippe Poinçot**

le 15 décembre 1999

---

**Classification et recherche d'information  
bibliographique par l'utilisation des cartes  
auto-organisatrices, applications en astronomie.**

---

Présidente du jury : A. Acker  
Directeur de thèse : F. Murtagh  
Rapporteurs : C. Chrisment  
F. Genova  
J. Lequeux



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>I</b>	<b>Préliminaires</b>	<b>3</b>
<b>2</b>	<b>La recherche d'information</b>	<b>5</b>
2.1	L'information textuelle . . . . .	5
2.2	Aspects généraux des systèmes de recherche d'information . . . . .	5
2.2.1	L'interface de consultation . . . . .	6
2.2.1.1	Interrogation . . . . .	6
2.2.1.2	Visualisation de l'information . . . . .	7
2.2.1.3	Les visualisations graphiques . . . . .	8
2.2.2	Organisation interne des données . . . . .	9
2.2.2.1	La recherche dans les textes bruts . . . . .	9
2.2.2.2	L'indexation . . . . .	10
2.2.2.3	Les fichiers inverses . . . . .	10
2.2.3	Évaluation des systèmes de recherche d'information . . . . .	11
2.2.3.1	La pertinence des documents . . . . .	11
2.2.3.2	Taux de précision, taux de rappel . . . . .	11
2.2.4	Conclusion . . . . .	13
2.3	La représentation des documents . . . . .	13
2.3.1	Les termes d'indexation . . . . .	13
2.3.2	L'indexation manuelle . . . . .	13
2.3.3	L'indexation automatique . . . . .	13
2.3.3.1	Les anti-dictionnaires . . . . .	14
2.3.3.2	La loi de Zipf et l'élimination des termes peu fréquents . . . . .	14
2.3.3.3	La recherche des radicaux . . . . .	15
2.3.3.4	La pondération des mots-clés . . . . .	15
2.3.4	Les vecteurs documents . . . . .	16
2.3.4.1	Les vecteurs binaires . . . . .	16
2.3.4.2	Les vecteurs numériques . . . . .	16
2.3.5	Conclusion . . . . .	17
2.3.5.1	L'indexation . . . . .	17
2.3.5.2	La conservation de l'information . . . . .	17

<b>3</b>	<b>Les réseaux de neurones</b>	<b>19</b>
3.1	Le neurone . . . . .	20
3.1.1	Le modèle biologique . . . . .	20
3.1.2	Vers une simulation du neurone biologique . . . . .	20
3.1.3	Le modèle formel . . . . .	21
3.2	Les neurones en réseau . . . . .	22
3.2.1	Différentes configurations de réseaux . . . . .	22
3.2.2	L'information dans les réseaux de neurones . . . . .	23
3.2.3	L'apprentissage . . . . .	23
3.3	Les réseaux de neurones à apprentissage supervisé . . . . .	24
3.3.1	Le cas d'un neurone seul . . . . .	24
3.3.1.1	L'apprentissage . . . . .	24
3.3.1.2	Champs d'utilisation . . . . .	25
3.3.2	La règle du delta : descente du gradient . . . . .	25
3.3.3	Les réseaux multicouches : rétro-propagation du gradient . . . . .	26
3.3.3.1	Déroulement de l'apprentissage . . . . .	26
3.3.3.2	Champs d'utilisation . . . . .	28
3.4	Les réseaux de neurones à apprentissage non supervisé . . . . .	28
3.4.1	Principes généraux . . . . .	28
3.4.1.1	L'architecture . . . . .	28
3.4.1.2	Le neurone gagnant . . . . .	28
3.4.1.3	L'ensemble de Voronoi . . . . .	29
3.4.1.4	L'apprentissage . . . . .	30
3.4.2	Réseaux du type "winner takes all" . . . . .	30
3.4.2.1	L'algorithme LBG . . . . .	30
3.4.2.2	Les algorithmes adaptatifs . . . . .	31
3.4.3	Réseaux de type "winner takes most" . . . . .	34
3.4.3.1	Les cartes auto-organisatrices de Kohonen . . . . .	35
3.4.3.2	Autres réseaux de type "winner takes most" . . . . .	38
3.4.4	Applications des réseaux à apprentissage non supervisé . . . . .	39
3.5	La convergence de l'apprentissage . . . . .	40
3.5.1	Le point de départ . . . . .	40
3.5.2	L'ordre de passage des vecteurs d'entrée . . . . .	41
3.5.3	Converger vers un "bon" minimum . . . . .	42
3.6	Conclusion . . . . .	42
<b>4</b>	<b>Les Systèmes de Recherche d'Information</b>	<b>45</b>
4.1	L'approche vectorielle . . . . .	45
4.2	L'approche booléenne . . . . .	47
4.3	L'approche neuronale . . . . .	47
4.3.1	Les systèmes à apprentissage supervisé . . . . .	48
4.3.2	Les systèmes à apprentissage non supervisé . . . . .	48
4.4	L'approche probabiliste . . . . .	49
4.5	Les reformulations des requêtes . . . . .	49
4.5.1	La ré-injection de la pertinence . . . . .	49
4.5.2	L'expansion des requêtes . . . . .	50
4.6	La classification . . . . .	51
4.6.1	Classifications non-hiérarchiques . . . . .	51

4.6.2	Classifications hiérarchiques . . . . .	52
4.6.3	Autres méthodes de classification . . . . .	52
4.6.4	Choix d'une méthode de classification . . . . .	53
4.7	Conclusion . . . . .	54
<b>II La Carte Bibliographique</b>		<b>55</b>
<b>5</b>	<b>Les données</b>	<b>57</b>
5.1	Le contenu des enregistrements . . . . .	57
5.2	La représentation des documents par leur mots-clés . . . . .	58
5.2.1	Travaux préliminaires sur les mots-clés . . . . .	58
5.2.2	Quelques chiffres . . . . .	59
5.2.3	Inconvénients . . . . .	60
5.3	L'utilisation des textes : l'indexation automatique . . . . .	61
5.3.1	Principe . . . . .	61
5.3.2	Résultats . . . . .	61
5.3.2.1	Le nombre de termes d'indexation . . . . .	61
5.3.2.2	Les groupes de mots . . . . .	62
5.3.2.3	Quelques chiffres . . . . .	62
5.4	Conclusion . . . . .	63
<b>6</b>	<b>La carte bibliographique, l'apprentissage</b>	<b>67</b>
6.1	Rappels sur les cartes auto-organisatrices . . . . .	67
6.1.1	Les entrées . . . . .	67
6.1.2	Les neurones de sortie . . . . .	67
6.1.3	Les vecteurs descriptifs et de référence . . . . .	68
6.1.4	L'apprentissage et la classification . . . . .	69
6.2	L'apprentissage . . . . .	70
6.2.1	Le contrôle de l'apprentissage . . . . .	70
6.2.2	Les paramètres de l'apprentissage . . . . .	72
6.2.2.1	Le nombre d'époques . . . . .	72
6.2.2.2	Le coefficient d'apprentissage et la fonction de voisinage . . . . .	73
6.2.2.3	L'apprentissage "batch" . . . . .	77
6.2.3	A propos de notre programme d'apprentissage . . . . .	77
6.2.3.1	La mise en forme "condensée" des données . . . . .	78
6.2.3.2	Les calculs sur les vecteurs . . . . .	78
6.3	Conclusion . . . . .	80
<b>7</b>	<b>Construction de la carte bibliographique, architecture et interface</b>	<b>81</b>
7.1	Les principes de la carte bibliographique . . . . .	81
7.1.1	L'exploration d'une base documentaire . . . . .	81
7.1.2	La recherche de documents . . . . .	82
7.2	L'architecture . . . . .	82
7.2.1	L'influence du nombre de classes . . . . .	82
7.2.2	Les cartes détaillées . . . . .	83
7.2.3	Quelques chiffres . . . . .	84
7.2.4	Les débordements . . . . .	84

7.3	L'interface . . . . .	86
7.3.1	Les cartes de densité . . . . .	86
7.3.2	Utilisation . . . . .	88
7.3.2.1	Les requêtes . . . . .	88
7.3.2.2	La visualisation du contenu des classes . . . . .	89
7.3.2.3	L'accès aux documents . . . . .	89
7.3.3	Implémentation de l'interface . . . . .	89
<b>8</b>	<b>La mise à jour de la carte bibliographique</b>	<b>93</b>
8.1	La réalisation d'un nouvel apprentissage . . . . .	93
8.2	L'utilisation du résultat de l'apprentissage initial . . . . .	94
8.2.1	L'ajout de documents : influence sur les traceurs $T1$ et $T2$ . . . . .	94
8.2.2	Le cas de l'apparition d'un nouveau thème . . . . .	95
8.3	Le ré-arrangement local, ou ré-apprentissage . . . . .	97
8.3.1	Influence sur $T1$ et $T2$ . . . . .	97
8.3.2	Le cas de l'apparition d'un nouveau thème . . . . .	98
8.4	Conclusion . . . . .	99
<b>9</b>	<b>Les résultats d'apprentissages, quelques visualisations</b>	<b>103</b>
9.1	Détail d'une classification : les articles de A&A, de 1994 à 1999 . . . . .	103
9.1.1	La visualisation de thèmes généraux . . . . .	103
9.1.2	La classification globale . . . . .	104
9.1.3	La taille des zones . . . . .	105
9.1.4	Les recouvrements . . . . .	105
9.1.5	L'organisation des documents d'un thème, l'exemple des étoiles . . . . .	106
9.2	Le point de convergence des apprentissages : comparaison de deux classifications . . . . .	107
9.2.1	L'organisation globale . . . . .	108
9.2.2	Les classes de documents . . . . .	108
9.2.2.1	La dispersion des classes . . . . .	108
9.2.2.2	La localisation des classes . . . . .	111
9.3	Conclusion . . . . .	112
<b>10</b>	<b>L'apport de la carte bibliographique : comparaison avec des résultats de l'ADS</b>	<b>113</b>
10.1	L'expérimentation . . . . .	113
10.1.1	Le choix de l'ADS . . . . .	113
10.1.2	La procédure d'interrogation . . . . .	113
10.2	Résultats . . . . .	115
10.3	Analyse des résultats . . . . .	116
10.4	Conclusion . . . . .	117
<b>11</b>	<b>Conclusion</b>	<b>119</b>
<b>A</b>	<b>Trois applications supplémentaires de la carte bibliographique</b>	<b>123</b>
A.1	La visualisation des thèmes abordés par différents journaux . . . . .	123
A.2	La classification de catalogues d'objets astronomiques . . . . .	125
A.3	La classification de sites Internet . . . . .	125

---

A.3.1	La provenance des données . . . . .	125
A.3.2	Le choix du moteur de recherche . . . . .	125
A.3.3	Les résultats . . . . .	126
<b>B</b>	<b>Exemple d'utilisation de la Carte Bibliographique</b>	<b>131</b>
<b>C</b>	<b>Les algorithmes de descente de gradient</b>	<b>137</b>
<b>D</b>	<b>Généralisation de la règle du delta</b>	<b>141</b>
<b>E</b>	<b>Les paramètres de l'apprentissage : quelques résultats</b>	<b>143</b>
E.1	Expressions des fonctions $\alpha(t)$ et $h(r, t)$ . . . . .	143
E.2	Variations de $T1$ et $T2$ en fonction de $\alpha(t)$ et $h(r, t)$ . . . . .	144





# Chapitre 1

## Introduction

L'astronomie tend à devenir la première science "tout-numérique" (Heck et Murtagh, 1993). À la source se trouve l'imagerie astronomique qui fait maintenant appel aux détecteurs numériques, que l'on retrouve aussi bien embarqués dans les véhicules spatiaux (sondes et satellites) que montés au foyer des télescopes au sol. Les données obtenues subissent alors des traitements numériques (élimination de bruits parasites, corrections d'artefacts instrumentaux) que les astronomes effectuent avec les logiciels adaptés, puis sont archivées et mises à la disposition de la communauté astronomique. Depuis 1993, année de l'apparition du navigateur Mosaic, un nombre croissant de ces données est accessible en ligne, via Internet.

Depuis sa création en 1972, le Centre de Données astronomiques de Strasbourg (CDS) développe des services qui permettent de regrouper, organiser et diffuser cette information. La base de données SIMBAD<sup>1</sup> qui y est maintenue, contient actuellement des données observationnelles sur plus de 2,7 millions d'objets astronomiques, les références de plus de 108.000 publications ainsi que plus de 3 millions de citations d'objets dans ces publications. Le CDS s'intéresse en particulier à l'information bibliographique : citation des objets dans les articles pour SIMBAD, mise en ligne de tables publiées dans le service catalogues et VIZIER<sup>2</sup>.

Le volume des données est bibliographique semble toujours s'accroître exponentiellement (Abt, 1998), et il est important de chercher à mettre en œuvre des méthodes innovantes d'accès à cette information. C'est l'objectif de cette thèse. Nous proposons un système qui permet la visualisation globale d'un ensemble de données bibliographiques pour l'exploration de son contenu, tout en permettant la recherche de documents sur des sujets précis. Ce système, la *carte bibliographique*, fait appel à un réseau de neurones pour effectuer une classification des données. Une interface graphique, consultable via Internet<sup>3</sup>, est construite autour de cette classification pour permettre différents types d'interrogation et de visualisation.

Dans la première partie de cet ouvrage, nous rappellerons les fondements de la recherche d'information. Ensuite, nous détaillerons le fonctionnement des réseaux de neurones en distinguant les deux types que constituent les réseaux à apprentissage supervisé

---

<sup>1</sup>Set of Identifications, Measurements and Bibliography for Astronomical Data.

<sup>2</sup>

<sup>3</sup><http://simbad.u-strasbg.fr/A+A/map.pl>, <http://simbad.u-strasbg.fr/ApJ/map.pl>

et ceux à apprentissage non supervisé dont font partie les cartes auto-organisatrices que nous utilisons. Nous terminerons cette première partie par une revue des principaux types de systèmes de recherche d'information.

La seconde partie de cet ouvrage est consacrée à la carte bibliographique. Après une description des données dont nous disposons, nous nous intéresserons à l'apprentissage de notre réseau, ainsi qu'au contrôle du déroulement de celui-ci. Nous tenterons alors d'en déterminer les paramètres optimaux. Ensuite, nous ferons une description détaillée de l'architecture de notre système et de son interface de consultation, avec ses différentes fonctionnalités. Nous détaillerons ensuite la mise à jour de la carte bibliographique ainsi que les différentes techniques utilisables pour la prise en compte des nouveaux documents de la base. Quelques visualisations de résultats de classification seront ensuite commentés, de même que les résultats de deux apprentissages similaires pour illustrer les aspects stables et instables des classifications obtenues avec les cartes auto-organisatrices. Le dernier chapitre mettra en avant les apports de la carte bibliographique par rapport aux systèmes habituels de recherche d'information documentaire en astronomie, par une comparaison avec le système de recherche de l'Astrophysics Data System (ADS).