Deuxième partie

La Carte Bibliographique

Chapitre 5

Les données

Le CDS[1] est un laboratoire dont la vocation est de regrouper et organiser les différents types d'information en astronomie (Egret et al., 1995). C'est ainsi que nous pouvons y trouver une base documentaire qui contient à ce jour plus de 108000 références bibliographiques. Parmi ces références, il en existe environ 30000 pour lesquelles nous disposons d'une liste de mots-clés ou d'un résumé. Ces documents sont des articles parus dans les plus importantes revues en astrophysique[2].

Nous allons examiner de plus près le contenu de ces données, et mettre en évidence les éléments susceptibles d'apporter une information sur le contenu des documents. Cette information est primordiale puisqu'elle est ensuite utilisée pour la création de vecteurs directement utilisables par notre système (voir la section 2.3 sur la représentation des documents).

5.1 Le contenu des enregistrements

Les enregistrements des références bibliographiques contenues au CDS contiennent tous au moins (Tab. 5.1) :

- un identificateur qui regroupe de manière unique et condensée[3] : l'année de publication, le journal, le numéro de volume ou le type de publication, un qualificateur ainsi que le numéro de la première page de l'article, et enfin la première lettre du nom du premier auteur ;
- des informations concernant la revue dans laquelle est publié l'article, les numéros des pages qu'il y occupe ;
- la liste des auteurs ;
- le titre de l'article.

Selon les sources de ces enregistrements, d'autres données relatives aux articles sont disponibles dont les plus intéressantes, en raison de l'information sur le contenu de l'article qu'elles apportent, sont une liste de mots-clés ou encore un résumé. Ce sont ces informations

[1] Centre de Données astronomiques de Strasbourg

[2] Astronomy and Astrophysics, Astronomy and Astrophysics Supplement Series, Astronomical Journal, Astrophysical Journal, Astrophysical Journal Supplement Series, New Astronomy et Publications of the Astronomical Society of the Pacific et les Monthly Notices of the Royal Astronomical Society.

[3] Ce code (le *bibcode*) a été défini en collaboration avec une équipe de Pasadena (Californie). L'information est codée sur 19 caractères.

apportées par les mots-clés ou les résumés ainsi que par le titre des articles, que nous pouvons utiliser pour la représentation des documents (voir la section 2.3).

%R	1999ApJ...523..855S
%D	October(I) 1999
%J-859	
%T	Experimentally Derived Dielectronic Recombination Rate Coefficients for Helium-like C V and Hydrogenic O VIII.
%A	Savin, D.W.
%I	Department of Physics and Columbia Astrophysics Laboratory, Columbia University, New York, NY 10027 ; (savin@astro.columbia.edu)
%c	American Astronomical Society 1999
%B	Using published measurements of dielectronic recombination (DR) resonance strengths and energies for C V to C IV and O VIII to O VII, we have calculated [...]
%K	Atomic Data
%K	Atomic Processes

Tab. 5.1: *Exemple d'enregistrement d'un article, dans lequel nous trouvons le bibcode, la date de publication, le numéro de la dernière page, le titre, le nom de l'auteur, l'institution, le résumé et les mots-clés. .*

5.2 La représentation des documents par leur mots-clés

Dans un premier temps, nous nous sommes limités à l'utilisation des termes descriptifs des documents (mots-clés) présents dans certains enregistrements. En procédant ainsi, le volume de données à manipuler reste raisonnable puisque chaque document est finalement associé à peu de termes : 3 en général, 8 au plus (voir Fig. 5.2). De plus, nous pouvons supposer que les termes descriptifs dépeignent bien le contenu des documents puisqu'ils sont choisis soit par les auteurs eux-mêmes, soit par un expert. Cette solution qui fait appel aux mots-clés présents dans les enregistrements nous a donc semblé un bon point de départ pour la représentation des documents en vue d'une utilisation par notre système.

5.2.1 Travaux préliminaires sur les mots-clés

écriture correcte	Stars : Hertzsprung-Russell Diagram
variante 1	stars : HR diagram
variante 2	stars : H-R diagram
variante 3	Hertzsprung-Russell diagram
etc...	

Tab. 5.2: *Différentes écritures rencontrées pour un même mot-clé : [Stars : Hertzsprung-Russell Diagram].*

Les mots-clés associés aux documents font partie des enregistrements utilisés ici. Ces enregistrements sont donc théoriquement directement utilisables, c'est à dire que, en principe, le travail consiste à :

- effectuer un relevé des différents mots-clés présents dans la base ;
- calculer leur nombre d'occurences (c'est à dire le nombre de documents auxquels ils sont associés) ;

- éliminer les moins fréquents (ceux qui apportent de l'information sur trop peu d'articles) selon la loi de Zipf (voir la section 2.3) ;
- constituer les vecteurs associés aux documents, dont chaque composante est relative à un des mots-clés que nous avons conservés (voir la section 2.3).

Néanmoins, pour les articles provenant du journal "Astronomy and Astrophysics", nous avons constaté des irrégularités dans la forme des mots-clés (voir le tableau 5.2). Rappelons tout d'abord que pour cette revue, les mots-clés sont attribués par les auteurs des articles, qui doivent faire leur choix dans une liste fournie par l'éditeur. Ces irrégularités, qui peuvent aller de la simple faute de frappe, à l'invention de nouveaux mots-clés (chose assez rare, mais compréhensible puisqu'il peut arriver qu'un thème ne soit pas correctement décrit par les mots-clés proposés par l'éditeur), conduit à une **perte d'information**. En effet, les mots-clés les moins fréquents étant éliminés, les mots-clés erronés ont toutes les chances de l'être également, si bien que les documents concernés y perdent un terme descriptif. Dans le pire des cas, certains documents ne sont plus associés à aucun des mots-clés sélectionnés car associés uniquement à des mots-clés peu fréquents ou erronés. Ces documents sont alors éliminés de notre base.

Nous avons été en mesure d'apporter des corrections automatiques sur ces données, en vue d'homogénéiser les différentes écritures rencontrées pour un même mot-clé (une correction individuelle des mots-clés étant difficilement envisageable en raison du volume des données). Ceci a été possible grâce au fait que certaines erreurs, fautes de frappe ou écarts par rapport à la syntaxe pré-définie, étaient assez fréquentes. Ainsi, les corrections apportées ont pu réduire le nombre de documents sans mot-clé (et donc non-utilisables) d'environ 80%. Ces documents représentent alors environ 4% de la totalité des documents.

5.2.2 Quelques chiffres

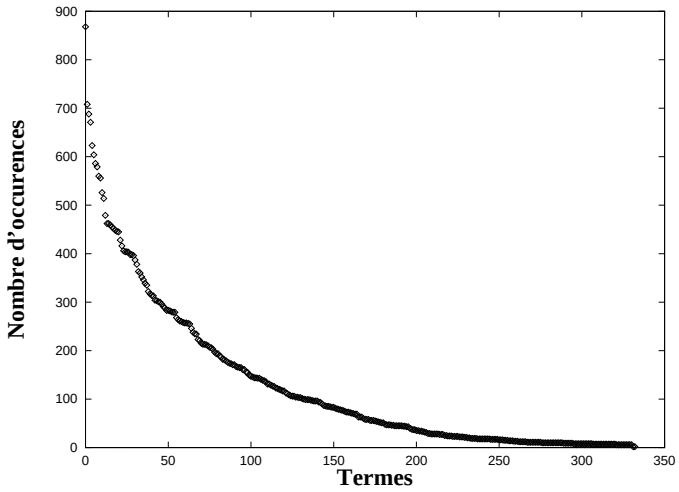

Fig. 5.1: *Fréquence d'apparition des 333 mots-clés (termes) dans environ 12000 documents tirés de ApJ (1994-1999). Nous y voyons que le terme le plus fréquent apparaît dans plus de 850 documents. Les cinq mots-clés les plus fréquents sont : "Accretion, Accretion Disks", "Cosmology : Theory", "X-Rays : Stars", "ISM : Molecules" et "Magnetohydrodynamics : MHD".*

Voici quelques aspects quantitatifs des documents que nous avons été en mesure d'utiliser : les documents issus des journaux "Astronomy and Astrophysics" (A&A), et "Astro-

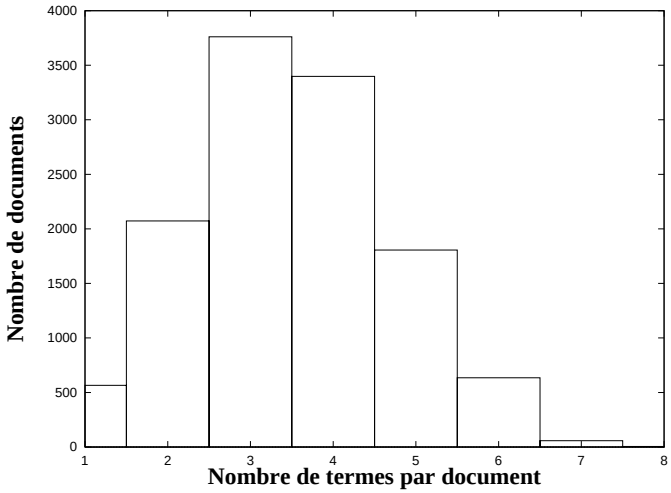

Fig. 5.2: *Histogramme du nombre de mots-clés (termes) par document. Nous y voyons que les documents qui contiennent trois mots-clés sont les plus fréquents.*

physical Journal" (ApJ).

Les graphes 5.1 et 5.2 concernent les documents parus dans "ApJ" entre 1994 et 1999 (environ 12000) pour lesquels nous avons conservé tous les mots-clés présents dans au moins 5 documents. Ces graphes montrent respectivement :

- la fréquence d'apparition des termes retenus (après élimination des moins fréquents) dans les documents . On remarque que la courbe suit bien la "loi" de Zipf (voir la section 2.3) ;
- le nombre de documents qui contiennent un nombre donné de mots-clés. En moyenne, les documents contiennent 3,4 mots-clés, les plus nombreux étant ceux qui en contiennent 3.

Remarque 1 : Nous obtenons des chiffres quasiment semblables pour le journal "A&A", ainsi que des graphes tout-à fait comparables. La seule différence étant dans le nombre inférieur de publications durant la même période : environ 7500 articles (pour environ 12000 dans "ApJ") ;

Remarque 2 : Nous avons traité les articles de ces deux journaux séparément car les mots-clés utilisés n'étaient pas tous semblables, mais il semble maintenant que la plupart des journaux en astronomie/astrophysique commencent à s'accorder sur une liste de mots-clés commune.

5.2.3 Inconvénients

L'utilisation des seuls mots-clés présents dans les enregistrements n'est pas sans poser certains problèmes.

Tout d'abord, il semble regrettable de perdre les informations contenues dans les titres (information disponible dans tous les enregistrements). Il en est de même pour les résumés qui, quand ils sont disponibles, sont évidemment une grande source d'information sur le contenu des articles.

Ensuite, la liste de mots-clés proposée par une revue peut se révéler insuffisante et ne pas contenir de termes pour définir précisément certains articles.

Le plus ennuyeux est certainement que tous les articles ne sont pas encore définis sur la même base de mots-clés (même si les différentes revues commencent à mettre au point une liste commune, cela ne changera rien sur les articles déjà parus), de même qu'un grand nombre d'articles sont enregistrés au CDS sans même être définis par des mots-clés.

L'utilisation exclusive des mots-clés associés aux documents pour créer une représentation utilisable par un système de recherche ne peut aboutir qu'à :

- une segmentation de l'ensemble des documents en fonction de leur revue d'origine, ce qui interdit leur utilisation de manière globale ;
- l'élimination des documents pour lesquels on ne dispose pas des mots-clés (ce qui représente environ 70% des références enregistrées au CDS).

Face aux limitations inhérentes à la représentation des documents basée exclusivement sur les mots-clés qui leur sont attribués, nous nous sommes orientés vers l'utilisation de toute l'information textuelle présente dans les enregistrements.

5.3 L'utilisation des textes : l'indexation automatique

5.3.1 Principe

L'indexation automatique des documents à été décrite dans la section 2.3.3. Néanmoins, nous en rappelons brièvement les principes :

- tous les mots des textes sont passés en revue et comptés ;
- les termes *vides* (dénués de sens) sont éliminés, ainsi que les termes peu fréquents ;

Les termes restant forment la liste de termes descriptifs sur lesquels est ensuite construite la représentation des documents. Pour effectuer ce travail, nous avons écrit notre propre programme d'indexation.

5.3.2 Résultats

Pour effectuer les indexations, nous choisissons d'utiliser le maximum d'information relative au contenu des documents. Ainsi, les textes que nous prenons en compte sont le titre des documents, le résumé (lorsqu'il est disponible), et les mots-clés[4].

5.3.2.1 Le nombre de termes d'indexation

Il résulte de l'indexation des textes une grande quantité de termes dont beaucoup peuvent être éliminés. Nous avons utilisé pour cela différents filtres :

1. l'élimination des mots de longueur inférieure à 3 lettres ;
2. l'élimination des termes apparaissant dans peu de documents : nous conservons typiquement les termes dont le nombre d'occurences dans des documents différents est supérieur à 20 ou 10, suivant le volume des documents à indexer ;

[4]Lors de l'indexation des documents, les termes présents dans les champs "mot-clé" présents dans les enregistrements sont traités de la même manière que les termes des champs "titre" et "résumé"

3. pour la recherche des radicaux, deux méthodes sont possibles avec notre programme : la plus simple étant l'élimination systématique de tous les "s" à la fin des mots, la seconde est celle proposée par Porter (Porter, 1980) ;
4. si les termes restant sont encore trop nombreux, nous laissons la possibilité d'éliminer tous les termes qui n'apparaissent jamais deux fois dans au moins un document (ceux dont la fréquence relative[5] n'excède jamais 1).

A titre d'exemple, le tableau 5.3 montre le nombre de termes conservés après l'application de chaque filtre sur les enregistrements des articles parus dans "ApJ" entre 1994 et 1999.

filtre	nombre de termes
1	21815
1 + 2 + 3 (élimination des "s" en fin de mot)	2232
1 + 2 + 3 (Porter)	2100
1 + 2 + 3 (Porter) + 4	1083

Tab. 5.3: *Effet des différents filtres sur le nombre de termes retenus.*

5.3.2.2 Les groupes de mots

Notre programme d'indexation nous donne la possibilité de repérer les groupes de mots qui se retrouvent dans les textes. Nos différents essais montrent que les groupes de 2 mots sont porteurs d'une grande information sémantique (tableau 5.4). Cette particularité est beaucoup moins prononcée pour les groupes de mots plus longs (3 mots et plus).

Nous avons choisi de prendre en compte les groupes de deux mots fréquemment retrouvés dans les textes. Ainsi, lors de la classification, les documents qui contiennent chacun des termes constitutifs d'un groupe sans que ceux-ci apparaissent côte à côte, sont différenciés des documents qui dans lequel le groupe de mots est présent. Les groupes de deux mots, ainsi que chacun des termes simples qui les composent, font alors partie de l'ensemble des termes d'indexation.

5.3.2.3 Quelques chiffres

Nous présentons maintenant, à titre d'exemple, quelques aspects quantitatifs sur les données provenant des articles publiés dans ApJ entre 1994 et 1999 (environ 12000). Ces termes sont ceux issus des titres, des résumés (quand ils existent) et des mots-clés présents dans les enregistrements. Les résultats qui suivent sont ceux correspondant à la liste des termes d'indexation ayant subi les quatre filtres de sélection, celle qui comprend 1083 termes. Ces résultats sont présentés sur deux graphes :

- le graphe 5.3 représente la fréquence d'apparition des termes dans les documents. Comme dans le cas de l'utilisation des mots-clés présents dans les enregistrements, nous pouvons voir une courbe qui suit la loi de Zipf ;
- le graphe 5.4 représente le nombre de documents qui contiennent un nombre donné de mots-clés. On peut y voir deux pics :

[5] Voir la section 2.3.3.4 pour plus de détails sur les fréquences relative et absolue.

Termes simples	Groupes de 2 mots
abell	abell cluster
absence	absorption line
absorbed	absorption system
absorber	abundance ratio
absorbing	accreting neutron
absorption	accretion disk
absorption-line	accretion flow
abundance	accretion rate
accelerated	active galactic
acceleration	active galaxie
accreting	active region
accretion	advection-dominated accretion
accuracy	alfven wave
accurate	all-sky survey
acoustic	ambipolar diffusion
across	angular momentum
active	angular resolution
activity	angular resolution
adaptive	angular scale
additional	aromatic hydrocarbon

Tab. 5.4: *Liste des 20 premiers termes simples en comparaison avec la liste des 20 premiers groupes de 2 mots. Les termes simples apportent moins d'information sémantique que les groupes de 2 mots. Remarque : les mots séparés par un "-" sont compris somme un mot unique par le logiciel.*

⇒ le premier est dû aux documents pour lesquels nous ne possédons pas de résumé (environ 10000). C'est le pic le plus élevé car dû au plus grand nombre de documents dont la majeure partie comprend 12 termes d'indexation ;

⇒ le second pic (le moins marqué) est dû aux documents pour lesquels on possède un résumé. Ces documents sont moins nombreux (2000 environ), le pic est donc plus faible. La majorité de ces documents sont associés à environ 40 termes d'indexation.

5.4 Conclusion

Nous sommes en mesure de construire une représentation des documents basée

- soit sur les mots-clés présents dans les enregistrements, ce qui permet la manipulation de données peu volumineuses, mais nous limite aux articles décrits par des mots-clés. De plus, nous ne pouvons travailler de manière globale que sur des articles décrits par un même ensemble de mots-clés ;
- soit sur les mots présents dans les titres, les résumés (lorsqu'ils existent) en plus des mots-clés (lorsqu'ils existent). Cette seconde approche élimine les limitations de l'approche précédente, mais oblige à travailler avec un volume de données important. De plus, le processus d'indexation étant automatisé, le risque de voir associer des documents à des termes peu représentatifs de leur contenu est important (le contexte

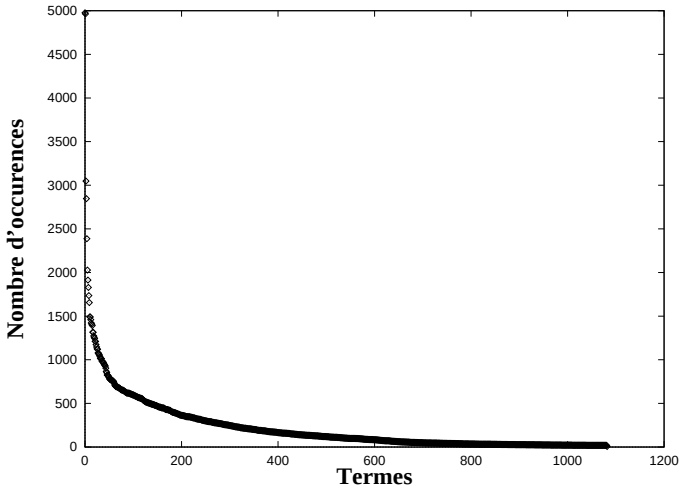

Fig. 5.3: *Fréquence d'apparition des 1083 termes d'indexation dans environ 12000 documents tirés de ApJ (1994-1999). Nous voyons que le mot-clé le plus fréquent apparaît dans presque 5000 documents. Les cinq mots-clés les plus fréquents sont : "stars", "galaxies", "ism", "observations" et "cosmology".*

dans lequel les termes sont employés n'est pas pris en compte). Cela peut se traduire par du bruit dans la classification des documents par exemple.

Ces deux méthodes nous fournissent une liste de termes (ou expressions) auxquels correspondent les composantes des vecteurs descriptifs des documents. Les composantes de chaque vecteur sont ensuite pondérées suivant les formules décrites en 2.3.3.4, puis tous les vecteurs sont normés. Ces vecteurs sont alors directement utilisables par notre système.

Remarque : il est intéressant de noter que les vecteurs descriptifs des documents comportent tous au moins 80% de composantes nulles. Ceci est d'ailleurs vrai que l'on utilise les mots-clés présents dans les enregistrements, ou que l'on effectue une indexation automatique. Nous avons mis cette particularité à profit pour réduire à la fois la quantité de mémoire occupée par les données, et les temps de calcul pour le traitement de ces données.

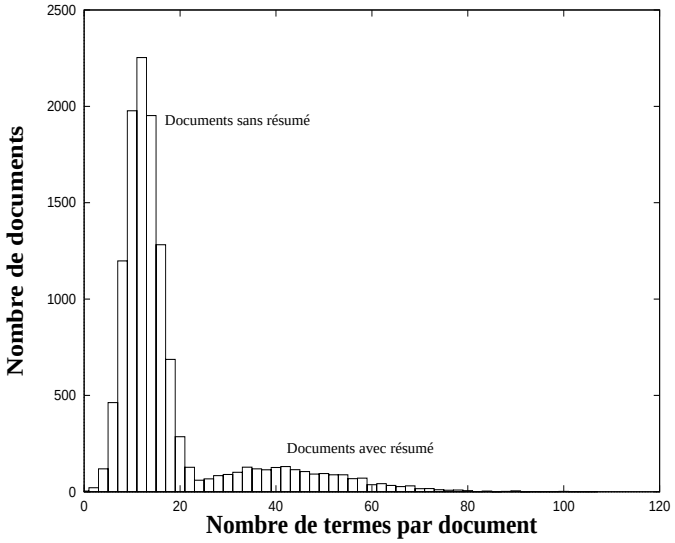

Fig. 5.4: *Histogramme du nombre de termes d'indexation par document.*