# Chapitre 7

# Construction de la carte bibliographique, architecture et interface

Le chapitre précédent nous a permis de détailler l'apprentissage des cartes auto-organisatrices (SOM), tant dans son rôle que dans son déroulement.

Nous allons maintenant nous intéresser à la manière dont nous utilisons les SOM, et les classifications qui en découlent, pour faciliter la recherche et l'exploration des données textuelles. Nous décrirons ensuite l'architecture des réseaux neuronaux qui constituent la base de notre système, ainsi que l'interface d'utilisation que nous avons construite sur cette base.

## 7.1 Les principes de la carte bibliographique

Nous utilisons les SOM pour classer des documents en fonction des termes d'indexation (ou mots-clés) qu'ils contiennent. Nous sommes donc en présence, après l'apprentissage, d'un tableau dans lequel les documents sont classés. Chaque case de ce tableau contient des documents ayant un maximum de mots-clés en commun, et il en est de même (mais dans une moindre mesure) pour les documents classés dans des cases proches.

Notre système exploite les caractéristiques des classifications obtenues avec les SOM, dans le but de permettre principalement :

- la découverte et l'exploration du contenu d'un grand ensemble de documents ;
- la recherche de documents.

### 7.1.1 L'exploration d'une base documentaire

Le regroupement des documents en classes permet l'examen du contenu d'un grand ensemble de documents.

En effet, chaque classe peut être représentée par un document moyen, dont le vecteur descriptif est la moyenne des vecteurs descriptifs des documents appartenant à cette classe. Afin d'évaluer le contenu d'une base documentaire, un utilisateur pourra examiner le document moyen de chacune des classes, travail beaucoup moins fastidieux que l'examen de chaque document. De plus, l'organisation spatiale des classes sur la carte est telle que les documents relatifs à un thème général se trouvent répartis dans plusieurs classes d'une

même zone. L'exploration des documents correspondants à un même domaine général en est facilitée car elle se fait par l'examen de classes voisines.

### 7.1.2 La recherche de documents

La recherche des documents relatifs à un thème précis peut se faire de deux façons :

- en localisant la classe qui contient un document, connu à l'avance, qui traite du thème choisi ;
- en localisant les classes des documents qui contiennent des termes en rapport avec le thème en question.

Dans ces deux cas, l'apport des classifications obtenues avec les SOM est considérable. En effet, lorsqu'une classe susceptible de contenir les documents recherchés est localisée, un utilisateur peut :

1. à l'aide du document moyen de la classe, évaluer l'importance du thème qui l'intéresse dans cette classe (s'agit-il du thème principal, ou d'un sous-thème ?) ;
2. accéder aux documents de la classe qui, s'ils ne sont pas tous directement liés au sujet qui l'intéresse, peuvent toutefois être complémentaires ;
3. examiner les classes voisines qui peuvent, elles aussi, contenir des documents relatifs au sujet de la recherche, ou complémentaires.

## 7.2 L'architecture

Les fonctionnalités que nous venons de décrire, et que notre système doit proposer, imposent des contraintes sur le réseau (la carte auto-organisatrice) qui se trouve à la base de la classification des documents.

### 7.2.1 L'influence du nombre de classes

La taille de la carte, c'est-à-dire le nombre de neurones de sortie, détermine le nombre de classes de documents. Nous devons choisir ce nombre en sachant que trop de classes peuvent être aussi nuisibles que trop peu.

**Un nombre de classes trop faible** conduit assurément à une classification imprécise, à des classes peu homogènes comportant un grand nombre de documents.

La découverte du contenu d'une base documentaire souffre peu d'un faible nombre de classes, surtout si les utilisateurs désirent avoir une idée générale sur les documents présents : les documents moyens des classes constituent une sorte de lissage ; les sujets peu courants passent alors inaperçus.

C'est plutôt la recherche de documents sur un thème précis qui est ici désavantagée car c'est un lourd travail qui attend les utilisateurs après qu'ils aient repéré la ou les classes qui contiennent les documents qu'ils recherchent. Des classes peu nombreuses contiennent nécessairement un grand nombre de documents, que les utilisateurs devront examiner afin d'obtenir l'information qu'ils désirent. L'interrêt d'utiliser les SOM au lieu d'un système traditionnel (qui renvoie les documents pertinents sous forme de liste, voir la section 2.2.1.2) devient alors incertain.

**Des classes trop nombreuses** contiennent peu de documents, mais ont de bonnes chances d'être homogènes. Néanmoins, elles ne facilitent pas la recherche de documents relatifs à un thème précis, ni même l'exploration de la base documentaire.

Le danger que présente un trop grand nombre de classes, est d'obliger les utilisateurs à examiner une quantité de classes trop importante. Le travail à fournir pour appréhender le contenu d'une base documentaire consiste alors à considérer une collection de classes dont beaucoup sont trop précises, et peut devenir aussi difficile que la consultation de chaque document. Pour ce qui est des recherches sur un thème précis, les utilisateurs devront également se pencher sur le grand nombre de classes qui contiennent les termes en rapport avec le thème de leur choix. Pour finir, il existe un risque important de perdre de l'information en oubliant d'examiner des classes.

### 7.2.2 Les cartes détaillées

Une façon de contourner le problème du choix du nombre de classes pour notre système est de hiérarchiser la classification. Cette technique a déjà été utilisée, notamment pour la compression d'image (Luttrell, 1989; Bhandarkar et al., 1997), et principalement en raison des réductions de temps et de mémoire nécessaires à l'apprentissage qu'elle permet. En ce qui nous concerne, c'est surtout pour son incidence sur les fonctionnalités de notre système, et les façons de l'utiliser, que nous avons opté pour une classification hiérarchique.

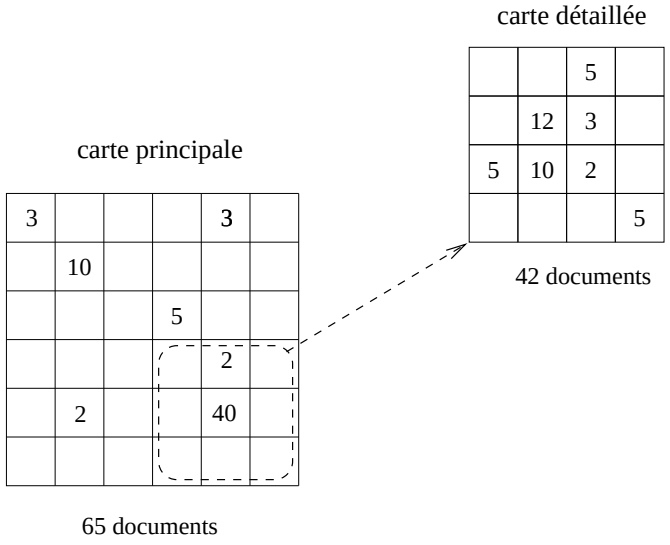


**Fig. 7.1:** *La carte principale et une carte détaillée.*

Les cartes détaillées sont un bon moyen d'accéder à des classes précises et homogènes, tout en proposant aux utilisateurs un nombre raisonnable de classes à examiner. L'idée est d'utiliser :

- une **carte principale**, regroupant la totalité des documents de la base ;
- des **cartes secondaires** qui contiennent, chacune, les documents classés dans une zone de la carte principale. Chacune d'entre elles est une nouvelle SOM utilisée pour la classification d'un sous-ensemble de documents.

Ainsi, les utilisateurs recherchant une vue globale de l'ensemble des documents pourront se limiter à l'exploration de la carte principale. Celle-ci doit comporter un nombre suffisamment faible de classes pour qu'un lissage des documents se produise avec les vecteurs moyens. La carte principale regroupe donc les documents dans des classes générales.

Les cartes détaillées proposent une classification plus fine des documents. Cela n'est pas dû uniquement au nombre restreint de documents qu'elles renferment, mais aussi (et surtout) au fait que ces documents ont des caractéristiques voisines (car faisant partie d'une même zone de la carte principale). La pondération des termes d'indexation joue ici un rôle considérable, car elle provoque une importante modification des vecteurs descriptifs des documents entre les représentations utilisées pour la carte principale et celles utilisées pour les cartes détaillées. Le tableau 7.1 montre comment les termes rares (dont la fréquence absolue est faible) ont beaucoup plus de poids dans les classifications des cartes détaillées que dans celle de la carte principale.

| | terme 1 : 100 docs. | terme 2 : 4 docs. |
|---|---|---|
| principale : 1000 docs. | poids : 2.3 | poids : 5.5 |
| détaillée : 110 docs. | poids : 0.1 | poids : 3.3 |

**Tab. 7.1:** *Valeurs des poids de deux termes d'indexation dans la représentation des documents utilisée pour la carte principale, et pour une carte secondaire. Nous observons que le rapport poids2/poids1 est plus élevé dans la carte secondaire où le terme 1 est très fréquent, que dans la carte principale. Nous n'avons pas tenu compte, ici, des fréquences relatives des termes dans chaque document.*

Les cartes détaillées seront utilisées par ceux qui désirent accéder à une information plus précise, comme par exemple, les documents similaires à un document donné.

### 7.2.3 Quelques chiffres

Le nombre de documents que nous avons couramment été amenés à traiter est d'environ 5000 à 10000. Nous avons choisi une taille de 15x15 neurones pour la carte principale. Ce nombre nous semble raisonnable puisqu'il correspond à une moyenne d'environ 20 à 45 documents par classe. La répartition des documents dans les classes n'étant pas homogène, les classes très peuplées peuvent contenir jusqu'à 200 documents environ.

Les cartes détaillées sont construites sur des réseaux de 5x5 neurones, et à partir des documents situés dans des zones denses de la carte principale. Nous avons estimé à 30, le nombre de documents qu'un utilisateur peut examiner sans se décourager ; nous construisons donc une carte détaillée pour chacune des classes comprenant plus de 30 documents.

**Remarque :** Les mêmes termes apparaissent souvent dans les documents de classes voisines. C'est donc souvent dans des classes voisines que des documents pertinents (qui contiennent le ou les termes demandés) se situent. C'est pourquoi nous avons, dans les versions les plus récentes de notre système, étendu à 9 classes (une classe centrale et les huit classes adjacentes) les documents classés dans les cartes détaillées. Nous disposons ainsi d'une classification plus précise pour un plus grand nombre de documents, initialement classés dans une zone plus large de la carte principale.

### 7.2.4 Les débordements

Nous avons utilisé une technique qui permet de s'affranchir des zones particulières que présentent les SOM : les bords. Dans une carte "normale", les nœuds situés sur le bord ont un statut à part puisqu'ils n'ont pas de voisins dans toutes les directions. Ceci peut poser des problèmes concernant la stabilité de la configuration des cartes lors de différents

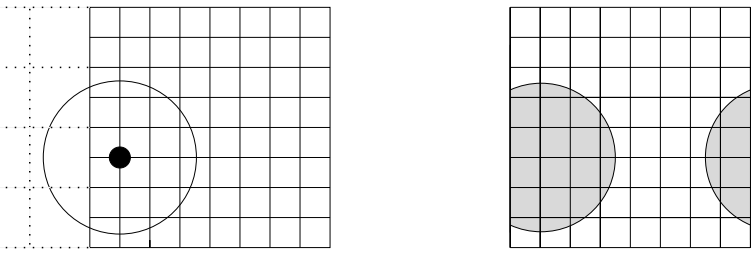

**Fig. 7.2:** *A gauche, le voisinage du nœud gagnant déborde de la grille. A droite, la zone grisée montre les neurones dont les poids sont modifiés durant l'apprentissage.*

apprentissages, soit favoriser la convergence de l'apprentissage dans des minima locaux. En effet, selon l'initialisation de la carte au début de l'apprentissage, une classe de vecteurs d'entrée pourra se voir attribuer une zone se trouvant soit vers un bord de la carte, soit vers le milieu. Pour ces deux cas dissemblables, on voit naître l'impossibilité d'obtenir deux configurations équivalentes. En effet, un neurone ne peut avoir les mêmes voisins lorsqu'il se trouve à une extrémité de la carte et lorsqu'il se trouve au centre.

Dans le but de remédier à ce problème, nous avons utilisé la technique suivante. Nous avons accordé la possibilité au voisinage d'un nœud donné de *déborder* du réseau lorsque le nœud en question est trop proche du bord de la carte, pour que la totalité de son voisinage soit comprise dans cette carte. La méthode consiste à *prolonger* le voisinage d'un point sur le *bord opposé* au débordement (fig . 7.2). Par exemple, les proches voisins situés "au dessus" d'un nœud du bord supérieur de la carte se trouveront sur le bord inférieur de la carte.

### Conséquences

Comme nous venons de le voir, la méthode du débordement rend possible l'équivalence de tous les nœuds du réseau, de manière à ne favoriser aucune zone de la carte durant l'apprentissage. Remarquons que nous n'avons pu mettre en évidence aucune amélioration de la convergence des apprentissages par une diminution des fluctuations des traceurs $T1$ et $T2$ (voir 6.2.1), celle-ci étant peut-être compensée par l'augmentation des contraintes sur les neurones situés aux extrémités de la carte, favorisant les fluctuations (mais peut-être ces effets sont-ils minimes et indétectables avec les traceurs $T1$ et $T2$ que nous utilisons).

Une autre conséquence, très visible contrairement à la précédente, de ce mode d'apprentissage est le fait que la carte résultante présente des similarités sur les bords opposés. C'est-à-dire que deux classes positionnées sur des bords opposés du réseau et à la même latitude sont en fait des voisins directs et regroupent de ce fait des documents de caractéristiques voisines. On peut alors considérer notre carte comme le motif répétitif d'une carte infinie, sur laquelle la continuité des caractéristiques des nœuds serait assurée (chaque nœud ressemblant à ses voisins) dans toutes les directions. Une autre façon de concevoir la carte résultante est de l'assimiler à une carte constituant la surface d'une sphère. On parvient ici au concept de carte sans bords ni limites réels. L'existence des quatre côtés de la carte résulte seulement du mode de représentation choisi, et leur position est en fait arbitraire.

Ainsi, chaque utilisateur aura la possibilité de modifier la carte à son gré en décalant les lignes (ou colonnes) verticalement (horizontalement), et en replaçant la ligne (colonne) ainsi poussée hors de la carte sur le bord opposé à celui qu'elle occupait avant la réorganisation

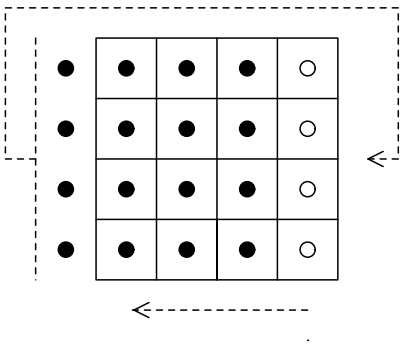

**Fig. 7.3:** *Détail de la reconfiguration d'une grille de neurones.*

de la grille (fig.7.3). Ceci peut être très utile dans le cas où une zone intéressante se situe à une extrémité du réseau. L'exploration de la zone se fait beaucoup plus simplement si la carte est modifiée de telle manière que cette zone soit placée au centre de la carte.

## 7.3 L'interface

L'interface permet aux utilisateurs de communiquer avec notre système, afin d'en exploiter les possibilités. Cette interface doit être capable de :

- présenter une vue globale des classes et des thèmes qui y sont abordés ;
- permettre la sélection de classes, et donner un résumé des documents qu'elles contiennent ;
- donner accès aux documents présents dans les classes ;
- permettre aux utilisateurs de repérer les classes contenant certains documents (connus) ou termes.

### 7.3.1 Les cartes de densité

Une carte de densité est une représentation graphique du tableau qui contient les documents, regroupés en classes spatialement organisées (voir 6.1). Il s'agit d'une image sur laquelle nous faisons apparaître les différentes informations sur la classification, que sont la position des classes, les thèmes généraux qui y sont abordés et la quantité relative de documents qu'elles contiennent (Fig. 7.4).

#### La position des classes

Nous avons vu que les classes de documents sont réparties de manière uniforme, suivant la disposition des neurones de sortie de la carte auto-organisatrice utilisée. Nous avons choisi de représenter les classes par de petits carrés, qui montrent d'une part la position des classes, mais aussi le pas de la grille, c'est-à-dire que ces carrés rappellent qu'il est inutile de chercher un troisième groupe de documents entre deux carrés adjacents.

La taille de ces carrés peut varier, en fonction des requêtes des utilisateurs. Pour chaque classe, la taille du carré est fonction de la pertinence maximale observée parmi tous les documents qui s'y trouvent. Ainsi, les documents correspondant au mieux à une requête sont facilement accessibles.

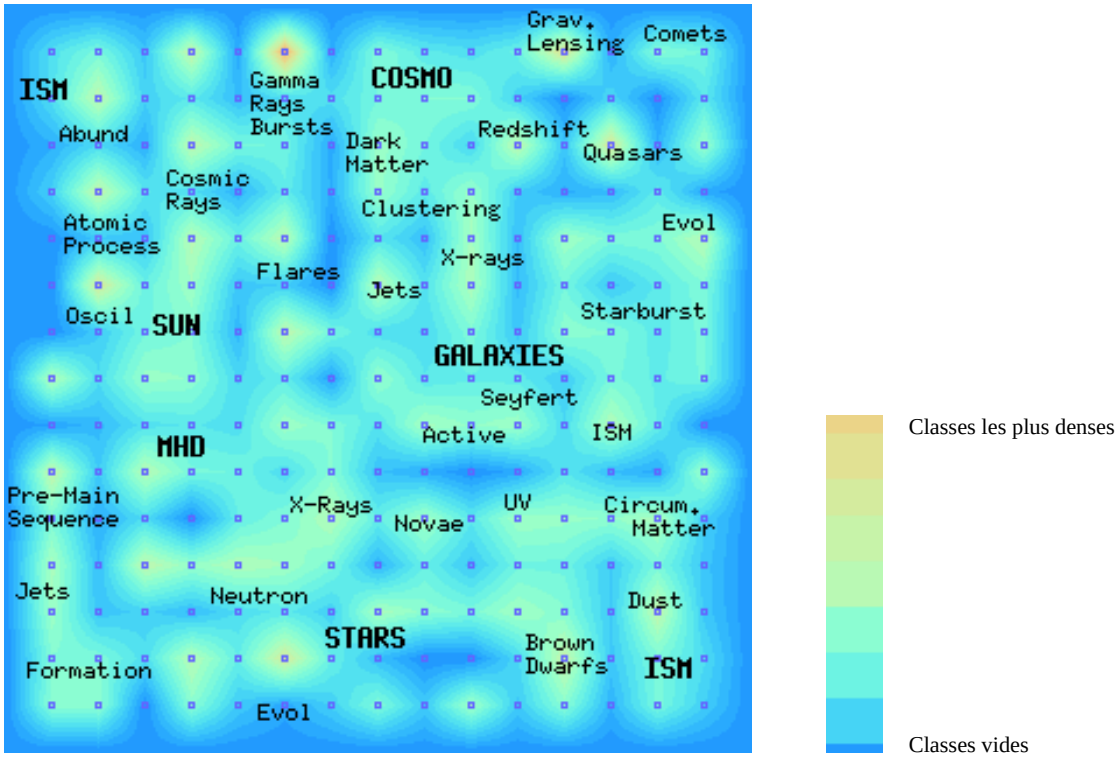


**Fig. 7.4:** *Carte de densité des documents parus dans ApJ de 1994 à 1999.*

### L'indication des thèmes

Afin de permettre le repérage rapide les sujets traités dans les différentes zones d'une carte, nous y inscrivons des termes génériques. Cette tâche est faite manuellement, bien qu'une automatisation soit à l'étude. Plusieurs difficultés rendent ce travail difficile à automatiser :

- le choix des termes à inscrire pour résumer un thème n'est pas toujours aisé ; on peut toutefois imaginer la sélection d'un terme d'indexation, sur des critères statistiques (le terme le plus fréquent par exemple), mais celui-ci ne sera pas forcément le plus générique.
- la position des termes sur la carte pose également des difficultés, car un thème est souvent abordé dans plusieurs classes voisines. Il nous semble cependant raisonnable de placer ces termes à proximité des classes les plus peuplées.
- des détails pratiques viennent compliquer la tâche, comme par exemple, la place limitée dont nous pouvons disposer sur une carte qui limite la longueur des termes que nous pouvons y inscrire.

### La densité des documents dans les classes

Les carrés représentatifs des classes, et les termes descriptifs des thèmes sont placés sur un fond de couleur variable. La couleur est liée à la quantité relative de documents contenus dans chaque classe.

En réponse à des requêtes, ces couleurs sont modifiées, de manière à ne représenter que

les documents pertinents. Ainsi, on visualise non seulement les classes qui contiennent des documents pertinents, mais aussi la densité de ces documents (rappelons que la taille des carrés associés aux classes donne une indication sur le degré de pertinence des documents).

Ces images sont créées au vol, en fonctions des requêtes des utilisateurs. Nous utilisons pour cela la librairie "gd[1]", qui regroupe des fonctions de création d'images en C.

**Remarque :** la représentation graphique de la classification des documents, sous la forme d'une carte de densité, ressemble à bien des égards au système Websom, proposé par l'équipe de Kohonen (Kohonen et al., 1996), dont nous nous sommes inspirés. Toutefois, d'importantes différences subsistent, principalement en ce qui concerne l'architecture de la carte.

Nous proposons une classification à deux niveaux (une carte principale, puis, des cartes détaillées). Ceci nous permet de proposer des cartes montrant un nombre de classes limité afin que les utilisateurs puissent les examiner dans des temps raisonnables. A l'opposé, la carte Websom peut comporter de 50 à 400 fois plus de classes que notre carte principale (Kohonen, 1997; Kaski, 1999), et les utilisateurs ont la possibilité d'agrandir les zones de leur choix afin d'explorer la carte. Les cartes détaillées ont l'avantage de proposer une classification plus précise, dans laquelle les termes rares sont davantage pris en compte (voir 7.2.2).

### 7.3.2 Utilisation

#### 7.3.2.1 Les requêtes

**La recherche d'un document connu**

Elle permet la localisation de la classe à laquelle le document appartient. Elle se fait en spécifiant le *"bibcode"* (voir 5.1) du document en question ; c'est pourquoi nous parlons parfois de "requête par bibcode".

Il est également possible de rechercher plusieurs articles en même temps, ou encore les articles parus une même année, dans une revue particulière (dans le cas d'une carte regroupant des articles de provenances diverses).

Dans tous les cas, les couleurs de la carte sont modifiées afin de refléter la densité des documents correspondant aux requêtes.

**La recherche de termes d'indexation**

Cette recherche, que nous appelons également *requête par mot-clé*, permet de localiser les classes qui contiennent des documents décrits par les termes choisis. Plusieurs termes peuvent être combinés de deux façons : a) avec ET (recherche de documents contenant tous les termes) ; b) avec OU (recherche des documents qui contiennent au moins un des termes choisis). Des recherches utilisant des combinaisons plus élaborées sont envisageables, mais n'ont pas encore été implémentées.

Les termes à rechercher doivent impérativement faire partie des termes descriptifs des documents (qui forment la base de définition des vecteurs descriptifs). Au lieu de simplement donner la liste (qui contient de 300 à 3000 termes en général) des termes d'indexation aux utilisateurs pour qu'ils effectuent leur choix, nous avons opté pour la mise en place

[1] gd est développé par Thomas Boutell. Les programmes sont accessibles à l'adresse suivante : http ://www.boutell.com/gd/

d'un système qui effectue une présélection des termes. Ainsi, les utilisateurs doivent indiquer une séquence de lettres que contient le terme qu'ils désirent sélectionner, puis tous les termes contenant cette séquence lui sont retournés.

Comme nous l'avons déjà dit, la réponse du système est double :

- les couleurs de fond de la carte indiquent la densité des documents *pertinents* présents dans les classes ;
- la taille des carrés qui localisent les classes est fonction de la pertinence du document le plus pertinent de chaque classe.

#### 7.3.2.2 La visualisation du contenu des classes

La carte de densité est une image réactive qui permet la sélection des classes (une à la fois). La sélection d'une classe permet aux utilisateurs d'accéder à un résumé de son contenu (voir fig. 7.6) :

- le nombre de documents appartenant à cette classe est indiqué ;
- les termes les plus fréquents rencontrés dans les documents de la classe sont affichés, ainsi que leur fréquence d'apparition. Les utilisateurs ont donc, en quelque sorte, accès aux composantes les plus importantes des vecteurs moyens des classes : il nous a semblé plus parlant d'indiquer la fréquence d'apparition des termes, ainsi que le nombre de documents, plutôt que la valeur des composantes correspondantes ;
- dans le cas d'une requête par mot-clé, les informations sont séparées en deux groupes : celles relatives aux documents pertinents, et celles relatives aux autres. Ainsi, il est facile de repérer les termes communs aux deux groupes de documents, et qui ne font pas partie des termes recherchés. Les utilisateurs peuvent alors juger si les documents qui ne correspondent pas à leur requête, mais qui se trouvent dans une classe qui contient des documents pertinents, peuvent les intéresser ou non.

**Remarque :** les cartes secondaires sont accessibles dès qu'une classe suffisamment peuplée est sélectionnée. Les fonctionnalités de ces cartes sont identiques à celles de la carte principale, que nous avons décrite dans cette partie.

#### 7.3.2.3 L'accès aux documents

Dès qu'une classe a été sélectionnée, il est possible d'obtenir la liste des documents qui s'y trouvent. On utilise pour cela des programmes développés dans le service bibliographique du CDS, qui permettent, outre l'accès aux titre, aux noms des auteurs et aux résumés (s'ils existent) des documents, de faire le lien avec les bases de données maintenues au CDS.

Ainsi, il est possible d'accéder aux données observationnelles concernant les objets cités dans les documents, ou encore à une image de ces objets et aux catalogues issus de missions d'observation où ces objets ont été étudiés.

Il est aussi possible d'accéder au document complet (quand celui-ci est disponible), ainsi qu'aux services proposés par ADS.

### 7.3.3 Implémentation de l'interface

Notre interface (Fig. 7.5 7.6 et 7.7) fait largement appel aux fonctionnalités du langage HTML, elle s'utilise donc via un logiciel qui reconnaît ce langage (Netscape, Internet Ex-

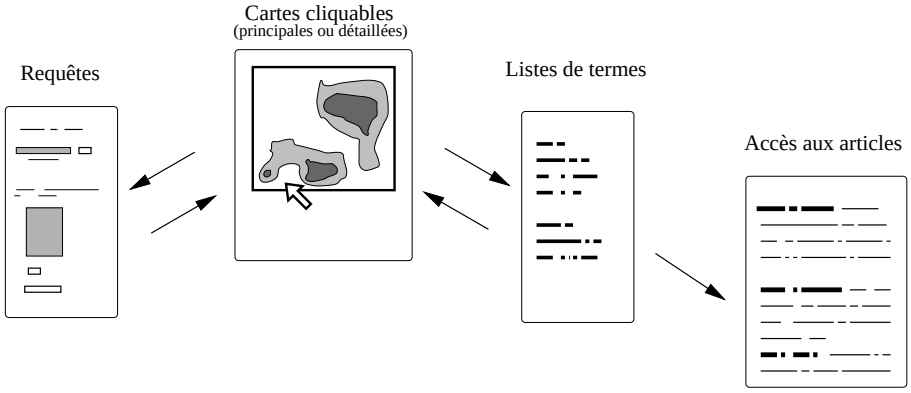


**Fig. 7.5:** *L'utilisation de l'interface. La sélection d'une classe à l'aide de l'image réactive permet d'accéder au résumé de son contenu. Il est alors possible de sélectionner une nouvelle classe, d'accéder aux documents de cette classe ou encore à une carte détaillée si elle existe. Des requêtes par documents ou par termes sont aussi possibles, sur la carte principale, comme sur les cartes détaillées.*

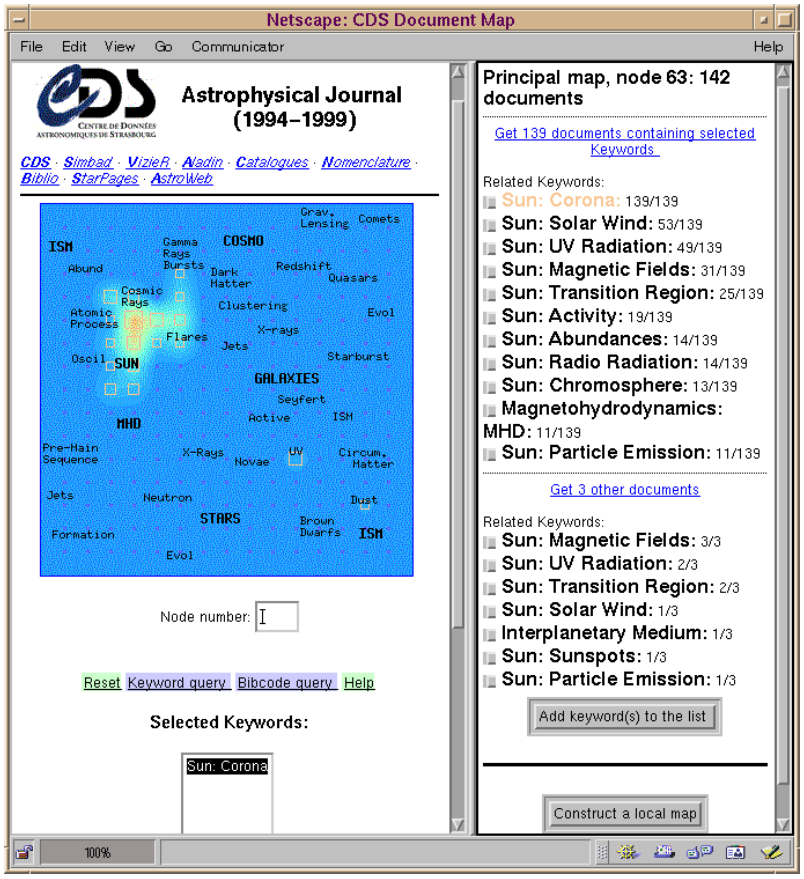


**Fig. 7.6:** *Vue de l'interface. Les documents qui contiennent le terme d'indexation : "sun : corona" apparaissent sur la carte principale. On remarque, à droite : le nombre de documents de la classe sélectionnée, les termes les plus fréquents qui apparaissent dans les documents de la classe. Ces termes sont divisés en deux groupes : ceux relatifs aux documents pertinents, et les autres. Les nombres situés à la suite des termes sont : a) la fréquence d'apparition ; b) le nombre de documents qui constituent le groupe (pertinents, ou non pertinents).*

plorer, etc...). Toute la partie interactive est gérée par des scripts CGI en langage PERL, qui commandent eux-mêmes des programmes en langage C. Nous utilisons ces différents langages pour leurs qualités respectives, qui sont une relative simplicité de mise en œuvre pour le PERL (que l'on utilise donc pour la gestion de l'interface), et une plus grande rapidité d'exécution pour le C (que nous réservons aux tâches qui nécessitent davantage de calculs, comme la création d'images par exemple) (Fig. 7.7).

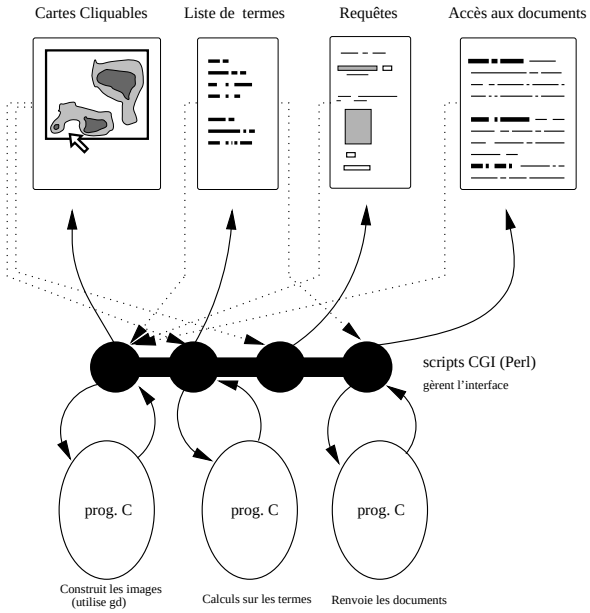


**Fig. 7.7:** *Implémentation de l'interface de la carte bibliographique.*