Chapitre 8

La mise à jour de la carte bibliographique

Une base de données reste rarement figée dans le temps. En effet, on est souvent amené à y apporter des modifications, que ce soient des corrections, des suppressions ou des ajouts d'éléments : c'est ce qu'on appelle la mise à jour. Dans le cas des bases de données textuelles qui nous intéressent, en particulier regroupant des articles scientifiques parus dans les revues spécialisées en astrophysique, les modifications en question sont dans la grande majorité des cas, l'ajout des articles récemment publiés. Typiquement, pour les journaux A&A[1] et ApJ[2], de nouveaux articles nous parviennent au rythme de 5 par jour en moyenne.

L'algorithme d'apprentissage des cartes auto-organisatrices (SOM) est celui d'une optimisation du classement sur une grille bidimensionnelle d'un ensemble de données (voir 3.4.3.1 et 6.1). Cette classification est optimisée, précisément, pour l'ensemble des données utilisées durant l'apprentissage. La mise à jour de la carte bibliographique, qui consiste à modifier cet ensemble de données, peut être réalisée de différentes façons :

- par la réalisation d'un nouvel apprentissage ;
- par le simple ajout des nouveaux documents, en utilisant le résultat de l'apprentissage initial ;
- en procédant à des ré-arrangements locaux sur la base de la classification initiale.

Nous allons décrire ces différentes techniques, ainsi que le contexte pour lequel elles sont les plus adaptées.

8.1 La réalisation d'un nouvel apprentissage

L'ajout de nouveaux articles, c'est-à-dire la modification de la collection des données à classer sur la grille, devrait, en toute logique, être suivie d'un nouvel apprentissage. C'est en effet l'unique méthode qui permette d'obtenir une classification optimisée pour le nouvel ensemble de données.

[1] Astronomy and Astrophysics
[2] Astrophysical Journal

Le problème de l'instabilité des classifications

Cette technique pose un problème rédhibitoire, dû à l'instabilité des classifications obtenues. En effet, l'algorithme d'apprentissage est un algorithme d'optimisation que l'on applique à une fonction qui comporte un très grand nombre de minima locaux. En raison des processus aléatoires qui apparaissent durant l'apprentissage, ce dernier peut converger vers l'un ou l'autre de ces minima locaux.

Deux minima locaux, même proches, peuvent correspondre à des configurations différentes des classifications. Cela se traduit par l'impossibilité d'obtenir des classifications identiques, d'un apprentissage à l'autre, même en utilisant les mêmes données.

Si la mise à jour d'une carte bibliographique s'accompagnait d'un apprentissage, elle générerait une nouvelle classification des classes de documents organisées différemment lors de chaque ajout de documents dans la base. Ces fréquents changements dans l'organisation des documents auraient surtout pour effet de désorienter les utilisateurs habitués à une carte.

Cette méthode n'est donc pas appropriée aux mises à jour fréquentes.

8.2 L'utilisation du résultat de l'apprentissage initial

Le résultat d'un apprentissage consiste en l'attribution de vecteurs de références à chacune des classes d'une carte. Les données sont alors rangées dans la classe dont le vecteur de référence est le plus proche de son vecteur descriptif (voir 6.1.4). Il est donc possible de classer les nouveaux documents de la même manière que ceux qui ont servi à l'apprentissage, à condition qu'ils soient définis sur la même base de termes d'indexation, afin que leur vecteur descriptif soit comparable aux vecteurs de référence des classes.

8.2.1 L'ajout de documents : influence sur les traceurs $T1$ et $T2$

Les nombres $T1$ et $T2$ sont des traceurs de la qualité d'une classification issue d'un apprentissage (voir 6.2.1). La comparaison des valeurs de ces traceurs avant et après l'ajout de nouveaux documents peut nous donner une indication sur la qualité de la classification résultante.

Pour cela, nous avons utilisé les documents de la revue ApJ dont nous disposons, et divisé l'ensemble en cinq groupes que décrit le tableau 8.1.

année(s)	nombre de documents	apprentissage	ajout
1994-1995	3752	X	
1996	1915		X
1996-1997	3926		X
1996-1998	5918		X
1996-1999	7336		X

Tab. 8.1: *Ensembles de documents utilisés pour évaluer l'influence sur la classification de l'ajout de nouvelles données. Ces documents, issus du journal ApJ, sont définis sur 350 termes.*

La figure 8.1 montre d'une part la variation des deux traceurs au cours de l'apprentissage réalisé sur les 3752 articles du premier groupe (voir le tableau 8.1), et d'autre part, les variations de ces paramètres causés par l'ajout des documents des groupes suivants. Nous constatons :

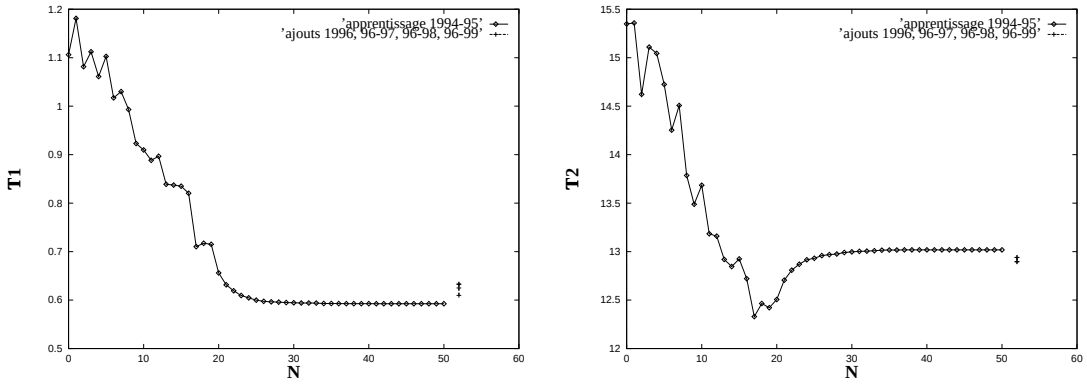


Fig. 8.1: *Variations de $T1$ et $T2$ lors de l'ajout de documents. Les croix correspondant aux valeurs de $T1$ et $T2$ obtenues après l'ajout des nouveaux documents montrent respectivement une augmentation ($T1$) et une diminution ($T2$) de ces valeurs, par rapport aux valeurs obtenues avec l'apprentissage initial (losanges). Remarque : certaines croix sont confondues.*

- un accroissement de $T1$ qui montre une moins bonne homogénéité des classes obtenues après l'ajout des nouveaux documents (voir fig. 8.2) ;
- une diminution de $T2$ qui correspond à une meilleure organisation spatiale (les documents de chaque classe ressemblent davantage aux vecteurs moyens des classes voisines). Cette amélioration apparente est, en fait, due au bruit apporté par les nouveaux documents qui viennent se greffer dans les classes existantes (voir fig. 8.2).

année(s)	nombre de documents	$(\Delta T1)/T1$	$(\Delta T2)/T2$
1994-1995	3752	0	0
1996	1915	3.4%	-0.7%
1996-1997	3926	6.7%	-0.6%
1996-1998	5918	7.3%	-0.9%
1996-1999	7336	7.1%	-0.9%

Tab. 8.2: *Variations de $T1$ et $T2$ lors de l'ajout des nouveaux documents. Nous voyons que l'ajout d'une année supplémentaire de documents entraîne une augmentation de $T1$ et une diminution de $T2$, sauf pour les documents de l'année 1999 qui inversent un peu la tendance, signe que les nouveaux documents apportés sont en majorité semblables aux documents des années précédentes.*

Nous remarquons également que les variations de $T1$ et $T2$ sont faibles devant les valeurs que les deux traceurs prennent au cours de l'apprentissage. En pratique, les classifications résultantes sont tout à fait utilisables. Le détail des variations des deux traceurs en fonction des groupes de documents ajoutés est donné dans le tableau 8.2.

8.2.2 Le cas de l'apparition d'un nouveau thème

Les résultats que nous venons de décrire sont issus d'expérimentations sur des données réelles que constituent une partie des articles parus sur une période de six années dans ApJ. Nous avons simulé l'apparition d'un nouveau thème en ajoutant à ces documents, tous les documents décrits par au moins un mot-clé relatif au soleil, que nous avions retirés au préalable. Le nombre de mots-clés concernés est de 26, et les documents de la seconde liste sont au nombre de 1141.

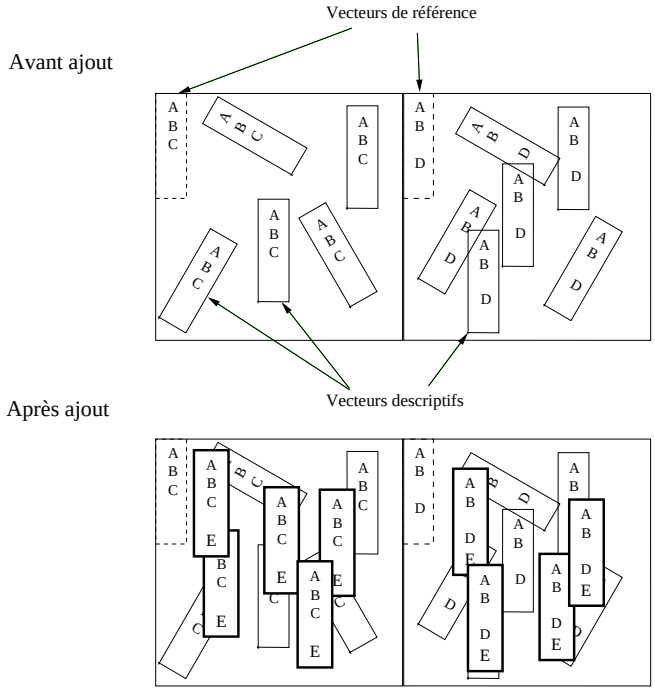


Fig. 8.2: *Deux classes d'une SOM, avant et après l'ajout de nouveaux vecteurs, ne modifiant pas les résultats de l'apprentissage initial. Cet exemple montre comment les nouveaux documents peuvent accroître la fréquence de certains termes (composante E) dans des classes voisines. Il s'en suit une dégradation de l'homogénéité des classes ($T1$ augmente) ainsi qu'une plus grande ressemblance des classes voisines (diminution de $T2$).*

liste de documents	nombre de documents	utilisation	$(\Delta T1)/T1$	$(\Delta T2)/T2$
non relatifs au soleil	11088	apprentissage	0	0
relatifs au soleil	1141	ajout	5.9%	1.8%

Tab. 8.3: *Variations de $T1$ et $T2$ lors de l'ajout des documents décrits par au moins un mot-clé relatif au soleil.*

Le tableau 8.3 montre les variations de $T1$ et $T2$ à la suite de l'ajout des documents qui traitent d'un nouveau thème. Alors que nous ajoutons relativement peu de documents (10% environ de l'ensemble de départ), nous constatons un effet important des nouveaux documents sur la classification :

- l'augmentation de $T1$ est comparable à celle que l'on constate quand les nouveaux documents sont aussi nombreux que les documents de départ, sans qu'un nouveau thème n'apparaisse (voir le tableau 8.2) ;

- la valeur de $T2$ augmente, conjointement à l'augmentation de $T1$. C'est le signe d'une détérioration de la classification, due à la grande différence entre les vecteurs descriptifs des nouveaux documents et les vecteurs de référence les plus proches (voir Fig. 8.3).

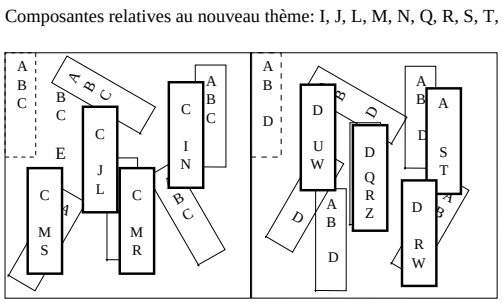


Fig. 8.3: *Ajout de documents qui traitent d'un nouveau thème. Les nouveaux documents sont associés à des classes dont les vecteurs de référence sont très différents de leur vecteur descriptif (une seule composante en commun dans notre exemple). Ainsi, ces documents sont classés uniquement en fonction des quelques mots-clés qu'ils ont en commun avec les autres documents.*

8.3 Le ré-arrangement local, ou ré-apprentissage

Le principe de ce type de mise à jour est le suivant :

- constituer une nouvelle liste de documents, en ajoutant les nouveaux à ceux qui ont servi à l'apprentissage ;
- utiliser cette liste pour effectuer un apprentissage sur les bases du précédent, qui permette un ré-arrangement local des documents, ce que l'on appelle un ré-apprentissage.

Le but du ré-apprentissage est donc d'apporter de petites modifications à la classification initiale, de manière à ce qu'il soit tenu compte des particularités des nouveaux documents. La difficulté est de doser ces modifications afin que la classification résultante ne soit pas très différente de la classification initiale. Pour cela, nous pouvons agir sur les valeurs des paramètres de l'apprentissage, sachant que plus les rayon de voisinage, coefficient d'apprentissage et nombre d'époques sont faibles, plus les modifications apportées sont petites.

Les traceurs $T1$ et $T2$ ne sont pas adaptés à la comparaison directe des classifications. Ils ne permettent pas de différencier deux classifications de qualité équivalente, qui montrent une organisation différente des documents. Nous avons donc procédé à différents essais, en comparant à chaque fois les positions relatives des classes avant et après le ré-apprentissage.

Les meilleurs résultats ont été obtenus en effectuant des ré-apprentissages dont les paramètres sont résumés dans le tableau 8.4.

données	initialisation grille	nb. époques	α	voisinage
initiales + nouvelles	résultat de l'apprentissage initial	15	0.1	R = 1 (unités adjacentes) constant durant l'apprentissage

Tab. 8.4: *Paramètres utilisés pour les ré-apprentissages*

8.3.1 Influence sur $T1$ et $T2$

La figure 8.4 montre les variations de $T1$ et $T2$ au cours de ré-apprentissages effectués sur les données décrites dans le tableau 8.1. En comparaison figurent également les valeurs de $T1$ et $T2$ prises au cours d'un apprentissage sur l'ensemble des données (1994-1999). Nous constatons que :

	$T1$	$T2$
ré-apprentissage	0.61	13.42
apprentissage normal sur tous les documents	0.60	13.26

Tab. 8.5: *Valeurs finales des traceurs $T1$ et $T2$ après diverses opérations.*

- il est possible de réorganiser les classes pour obtenir des degrés d'homogénéité comparables à ceux obtenus avec un apprentissage normal sur l'ensemble des données ($T1_{\text{ré-appr.}} \simeq T1_{\text{normal}}$) ;
- l'organisation globale des cartes obtenues après ré-apprentissage est moins bonne que dans le cas d'un apprentissage normal ($T2_{\text{ré-appr.}} > T2_{\text{normal}}$).
- les résultats sont meilleurs lorsque peu de documents nouveaux sont ajoutés : les valeurs finales de $T1$ et $T2$ s'accroissent simultanément quand le nombre de nouveaux documents augmente.

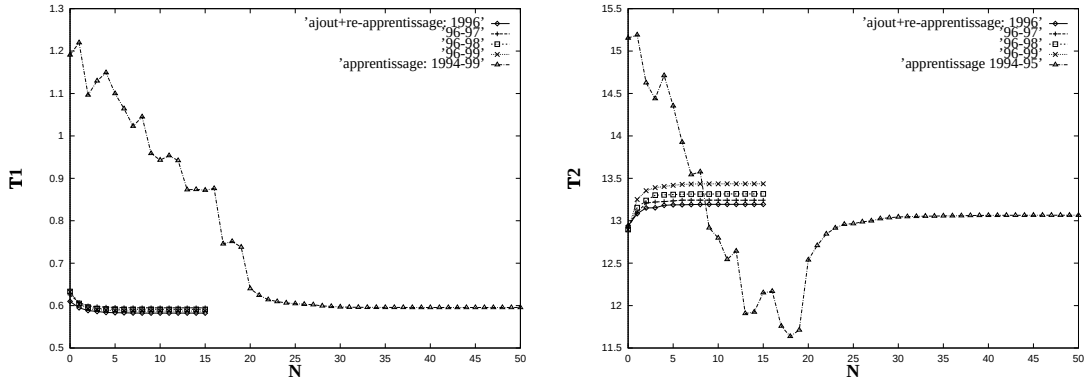


Fig. 8.4: *Variations de $T1$ et $T2$ au cours de ré-apprentissages à la suite de l'ajout des documents parus au cours de l'année 1996, de 1996 à 1997, de 1996 à 1998 et de 1996 à 1999, comparées aux variations au cours de l'apprentissage initial.*

8.3.2 Le cas de l'apparition d'un nouveau thème

Afin de simuler l'apparition d'un nouveau thème, nous avons utilisé les mêmes données que précédemment (voir 8.2.2), à savoir les articles possédant au moins un mot-clé relatif au soleil constituant les nouveaux documents.

Le tableau 8.5 montre les valeurs finales de $T1$ et $T2$ après l'ajout des nouveaux documents suivi d'un ré-apprentissage, ainsi que ces valeurs à la fin d'un apprentissage normal sur l'ensemble des données. On remarque que l'homogénéité des classes est comparable dans les deux cas $T1_{\text{ré-appr.}} \simeq T1_{\text{normal}}$, tandis que l'organisation globale est meilleure dans le cas d'un apprentissage normal ($T2_{\text{ré-appr.}} > T2_{\text{normal}}$).

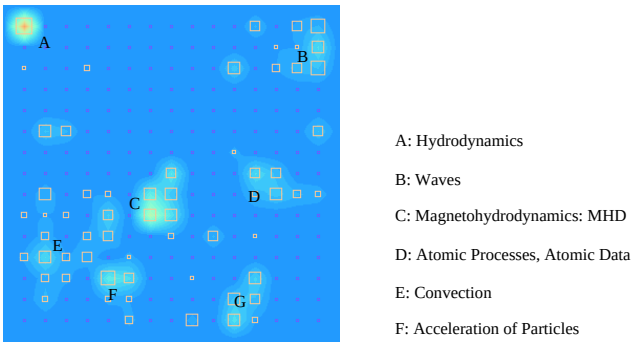
Influence sur la position des documents

La figure 8.5 montre la localisation des nouveaux documents a) avant et b) après un ré-apprentissage. Les différentes zones qui apparaissent contiennent des documents princi-

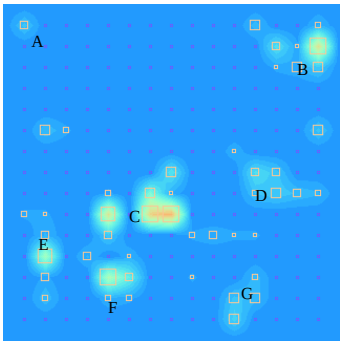
palement attirés par le ou les mots-clés spécifiés sur la figure.

La comparaison de ces deux cartes montre que le ré-apprentissage a eu pour effet de regrouper les nouveaux documents dans des classes qui leur sont plus adaptées. Une étude plus approfondie montre que les anciens documents qui se trouvaient dans ces classes avant le ré-apprentissage ont été déplacés dans des classes voisines, tandis que les autres classes de la carte sont restées quasiment inchangées dans leur composition et leur position.

Nous voyons néanmoins que le ré-apprentissage ne résoud pas les problèmes d'organisation globale. La figure 8.6 montre en effet que, lors d'un apprentissage normal, sur toutes les données, la plupart des articles que nous avons ajoutés pour simuler l'apparition d'un thème nouveau sont en fait situés dans une même zone de la carte. On visualise ici la raison des grandes valeurs que prend $T2$ après le ré-apprentissage.



Avant ré-apprentissage



Après ré-apprentissage

Fig. 8.5: *Localisation des nouveaux documents correspondant aux classifications avant et après ré-apprentissage. La dispersion des nouveaux documents est plus faible après le ré-apprentissage.*

8.4 Conclusion

Les différentes techniques de mise à jour que nous venons de décrire présentent des caractéristiques telles que, selon les cas, chacune d'entre elles pourra être utilisée.

L'ajout seul des nouveaux documents

Cette technique est bien adaptée à l'ajout de documents qui restent assez similaires dans leurs caractéristiques aux documents qui ont servi à l'apprentissage initial. Dans ces conditions, il est possible d'ajouter ainsi un grand nombre de documents : nous avons

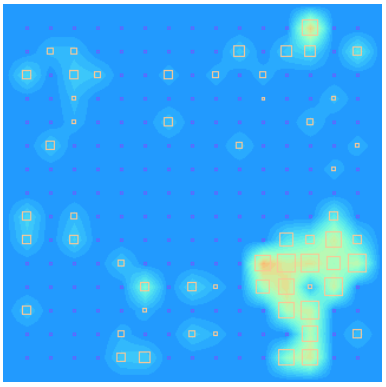


Fig. 8.6: *Localisation des articles relatifs au soleil. Ces articles, lorsqu'ils font partie de l'ensemble de documents utilisés lors de l'apprentissage, sont pour la plupart localisés dans une zone bien définie (à comparer à leur localisation Fig. 8.5).*

ajouté à la liste initiale jusqu'à deux fois plus de documents qu'elle n'en contenait, sans constater une diminution importante de la qualité de la classification

C'est cette méthode que nous utilisons pour la mise à jour des cartes en ligne du CDS, dont la quantité de documents a été multipliée par un facteur 1.5 environ pour la carte des articles de ApJ, et par 2 environ pour celle des articles de A&A.

Le ré-apprentissage

Nous avons vu les effets intéressants de cette technique sur la classification des documents relatifs à un nouveau thème. Ces derniers se retrouvent mieux classés entre eux, sans que la configuration des autres classes soit bouleversée. Nous pouvons également effectuer des ré-apprentissages alors qu'aucun thème nouveau n'est apparut, afin de rendre les classes plus homogènes.

Remarque : pour la mise en évidence de l'apparition d'un nouveau thème, on peut songer à la mise en place d'un système de contrôle régulier des fréquences d'apparition de tous les termes d'indexation dans la base. Un nouveau thème pourrait alors être mis en évidence par une augmentation rapide de la fréquence des termes qui lui sont reliés.

Le nouvel apprentissage sur toutes les données

C'est la méthode la plus radicale, qui conduit, bien sûr, à une classification optimale pour les données réactualisées. Mais, comme nous l'avons souligné au début de ce chapitre, les importantes modifications de la configuration de la carte qu'elle implique, ainsi que le désagrément occasionné aux utilisateurs qui en découle, rendent cette méthode difficilement utilisable.

La mise à jour d'une carte à long terme

Nous n'avons pas encore le recul suffisant pour avancer un procédé éprouvé de mise à jour à long terme. Nous pouvons toutefois proposer des éléments de réponse à ce problème.

La mise à jour fréquente d'une base peut être assurée par l'ajout seul des nouvelles données. Comme nous l'avons vu, le doublement de la taille de la base (par rapport à sa taille lors de l'apprentissage initial) ne pose pas de problème.

On peut ensuite imaginer que l'on puisse améliorer la qualité des classes en pratiquant un ré-apprentissage, par exemple à chaque fois qu'une quantité d'articles égale à celle qui a été utilisée durant l'apprentissage initial a été ajoutée.

Enfin, de nouveaux apprentissages utilisant toutes les données actuelles peuvent être effectués si ceux-ci ne sont pas trop rapprochés dans le temps. Une réorganisation complète d'une carte ne devrait pas poser de déconvenues si elle est pratiquée tous les ans, par exemple.