Classification non supervisée et visualisation 3D de documents

Nicolas Bonnel[*,**], Annie Morin[*], Alexandre Cotarmanac'h[**]

[*]IRISA, Campus Universitaire de Baulieu
Avenue du Général Leclerc, 35042 Rennes Cedex - France
{nicolas.bonnel,annie.morin}@irisa.fr
[**] France Telecom R&D, 35512 Cesson-Sévigné Cedex

Résumé. Le nombre de documents issus d'une requête sur le Web devient de plus en plus important. Cela nous amène à chercher des solutions pour aider l'utilisateur qui est confronté à cette masse de données. Une alternative possible à un affichage linéaire d'une liste triée selon un critère, consiste à effectuer une classification des résultats. C'est dans ce but que l'on s'intéresse aux cartes auto-organisatrices de Kohonen qui sont issues d'un algorithme de classification non supervisée. Cependant il faut ajouter des contraintes à cet algorithme afin qu'il soit adapté à la classification des résultats d'une requête. Par exemple, il doit être déterministe. De plus, la classification obtenue dépend fortement de la distance utilisée pour comparer deux documents. On évalue alors l'impact de différentes distances ou dissimilarités, afin de trouver la plus adaptée à notre problème. Un compromis doit également être trouvé entre le temps d'exécution de l'algorithme et la qualité de la classification obtenue. Pour cela, l'utilisation d'un échantillonnage est envisagée. Enfin, ces travaux sont intégrés dans un prototype qui permet de visualiser les résultats en trois dimensions et d'interagir avec eux.

1 Introduction

Avec l'augmentation constante des données disponibles sur le World Wide Web, il devient de plus en plus difficile d'extraire l'information pertinente pour une recherche donnée. Les moteurs de recherche, qui sont un moyen de représentation du Web pour les utilisateurs, retournent un nombre si important de résultats qu'il faut chercher de nouvelles méthodes de gestion de ces résultats. En effet, il devient nécessaire de trouver une alternative au simple affichage de listes ordonnées selon un seul critère (généralement un rang représentant la "pertinence").

Les résultats ou documents que l'on cherche à classer sont des pages Web. Seule la partie textuelle de ces documents est utilisée. Elle permet d'obtenir une représentation vectorielle (vecteurs de mots) qui est largement utilisée dans le domaine de la recherche d'informations. Ce sont ces vecteurs qui servent de données d'entrées pour la classification. On s'oriente vers des méthodes de classification automatique et plus particulièrement vers les cartes auto-organisatrices. La classification obtenue est ensuite