CASOM : Carte auto-organisée pour l'analyse exploratoire des tableaux de contingence

Rodolphe Priam
IRISA - Projet TexMex
263 av Gén Leclerc F-35000 Rennes
rpriam@gmail.com

Résumé. La visualisation des connaissances par des méthodes d'extraction de l'information pour un corpus de données multimédia est une question pertinente aussi bien en recherche d'information où l'on cherche les meilleurs documents répondant à une requête, qu'en analyse des données où l'on cherche à comprendre quantitativement le contenu du corpus. En effet, en recherche d'information, les corrélations inter-variables permettent d'enrichir la requête de la même façon qu'elles renseignent sur les liaisons interprétables en analyse des données. Une manière générale de répondre à l'objectif posé est l'emploi de méthodes de réduction efficaces qui permettent de mettre en évidence les différentes caractéristiques principales et locales du corpus. Les méthodes de carte auto-organisatrice entrent dans cette optique tout en apportant la dimension supplémentaire d'une carte projective de la distribution et partitionnant le plan en diverses thématiques adjacentes. Ces méthodes rendent appréhendable par l'humain un nuage de points plongé dans un espace de grande dimension par la construction d'une surface discrète qui épouse la forme de sa distribution. Elles offrent ainsi une propriété de cartographie apte à montrer une structure sous-jacente. Dans ce cadre, nous développons une représentation originale des cartes de Kohonen pour des vecteurs textuels bruts. Nous fournissons des indicateurs numériques interprétables et aboutissons à la définition d'une méthode de visualisation synthétique et globale d'un corpus : un *biplot* sous la forme d'un graphe de mots révélant leurs liaisons statistiques, superposable à la projection des documents. La méthode est illustrée par le graphe du vocabulaire extrait d'un corpus de données réelles.

1 Introduction

Dans ce papier, nous introduisons les algorithmes de carte auto-organisatrice et décrivons les principales représentations de cartes auto-organisatrices présentes dans la littérature. Ce cadre posé, nous développons notre méthode CASOM adaptée aux matrices de comptage et mettons en évidence ses diverses propriétés particulières en terme de critère optimisé et métrique. La méthode est illustrée sur un corpus de résumés textuels en construisant des graphes qui montrent les liaisons statistiques entre les termes ou mots du vocabulaire sélectionné dans le corpus ; le graphe de mots a la propriété de se superposer à celui des documents. Cette représentation est également l'occasion d'une réflexion sur la qualité des graphiques résultants. La conclusion dresse le