Sélection par entropie de descripteurs textuels pour la catégorisation de documents

Christophe Moulin, Christine Largeron

Université de Lyon, F-69003, Lyon, France
Université de Saint-Étienne, F-42000, Saint-Étienne, France
CNRS UMR5516, Laboratoire Hubert Curien
{christophe.moulin, christine.largeron}@univ-st-etienne.fr

1 La sélection de descripteurs

Dans le contexte de la catégorisation de documents, la sélection des descripteurs est une étape de pré-traitement importante qui permet non seulement de réduire la taille de l'index, mais aussi d'améliorer les performances des classifieurs. Parmi les approches utilisées pour construire un sous-ensemble de l'index, on peut distinguer d'une part, les méthodes de réduction de dimensions qui génèrent un nombre limité de nouveaux descripteurs en regroupant les descripteurs initiaux par affinité sous forme de concepts, comme par exemple la méthode LSA et, d'autre part, les méthodes de sélection de descripteurs qui visent à choisir un sous-ensemble des attributs initiaux à l'aide de critères tels que le critère de couverture de classe (*CC*) que nous avons défini dans (Gery et al. (2009)). Cependant, comme la plupart des critères de sélection de descripteurs, qu'il s'agisse de la fréquence d'apparition du terme (DF), du gain d'information (IG), du χ^2 (CHI2) ou encore de l'information mutuelle (IM), la couverture de classe *CC* exploite la distribution dans les classes des documents contenant ou ne contenant pas chaque terme. Or, un terme caractéristique d'une classe devrait non seulement apparaître dans un plus grand nombre de documents de la classe que des autres classes mais il devrait aussi y figurer plus fréquemment. Dans cet article [1], pour tenir compte, non seulement de la distribution entre les classes des documents contenant un terme, mais aussi de son nombre d'occurrences, nous proposons une extension du critère de couverture de classe (*CC*), appelée *Entropy based Category Coverage Difference* (*CCDE*), qui intègre l'entropie du terme. Evalué sur une large collection de documents XML extraits de l'encyclopédie Wikipédia, ce critère fournit de meilleurs résultats que des techniques classiques de sélection d'attributs basées sur la fréquence des documents contenant le terme, comme le gain d'information, l'information mutuelle ou le χ^2 et ses dérivés.

[1] Ce travail a été réalisé dans le cadre du projet Web Intelligence de la région Rhône-Alpes (`http://www.web-intelligence-rhone-alpes.org`).