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`).

2 Un critère de sélection basé sur la couverture de classe par l'entropie (CCDE)

Étant donnée une collection D de documents appartenant à un ensemble de classes disjointes $C = \{c_1, \ldots, c_k, \ldots, c_r\}$, on note $T = \{t_1, \ldots, t_j, \ldots, t_{|T|}\}$, un index de taille $|T|$ contenant la liste des termes (ou descripteurs) figurant dans les documents de D. Un document d_i de D est représenté par un vecteur $\vec{d_i} = (w_{i,1}, \ldots, w_{i,j}, \ldots, w_{i,|T|})$ où $w_{i,j}$ représente le poids du terme t_j dans le document d_i. La formule TF.IDF peut être utilisée pour calculer ce poids $w_{i,j} = tf_{i,j} \times idf_j$ où $tf_{i,j}$ est la fréquence relative du terme t_j dans le document d_i et idf_j est la fréquence inverse de document du terme t_j. $CCDE(t_j, c_k)$ est défini par :

$$CCDE(t_j, c_k) = \frac{(P(t_j, c_k))^2 - (P(t_j, \bar{c_k}))^2}{P(t_j, c_k) + P(t_j, \bar{c_k})} \times \frac{E_{max} - E(t_j)}{E_{max}} \quad (1)$$

où $P(t_j, c_k)$ (resp. $P(t_j, \bar{c_k})$) désigne la probabilité qu'un document contienne le terme t_j sachant qu'il appartient à la classe c_k (resp. aux autres classes), $E(t_j)$ est l'entropie de Shannon du terme t_j et E_{max} , la valeur maximale prise par l'entropie.

$CCDE$ peut être calculé en construisant une table de contingence pour le terme t_j et la classe c_k. Si dans cette table, on note A, le nombre de documents de la collection appartenant à la classe c_k et contenant le terme t_j ; B, le nombre de documents de la collection n'appartenant pas à la classe c_k et contenant le terme t_j ; C, le nombre de documents de la collection appartenant à la classe c_k et ne contenant pas le terme t_j ; D, le nombre de documents de la collection n'appartenant pas à la classe c_k et ne contenant pas le terme t_j (avec $N = A + B + C + D$), l'équation 1 peut être calculée par :

$$CCDE(t_j, c_k) \approx \frac{AD - BC}{(A + C)(B + D)} \times \frac{E_{max} - E(t_j)}{E_{max}}$$

	c_k	$\bar{c_k}$
t_j	A	B
$\bar{j}$	C	D

Références

Gery, M., C. Largeron, et C. Moulin (2009). UJM at INEX 2008 XML Mining Track. In *Proceedings of the INEX Workshop INtitiative for Evaluation of XML Retrieval, S. Geva, J. Kamps, and A. Trotman (Eds.)*, pp. 446–452. Springer-Verlag.

Summary

In the context of text categorization, we propose a novel feature selection criteria, called *Entropy based Category Coverage Difference*(*CCDE*), based on one hand on the distribution in the categories of the documents containing the term and on the other hand on its entropy. *CCDE* compares favorably with usual feature selection methods based on document frequency (*DF*), information gain (*IG*), mutual information (*IM*), χ^2, *odd ratio* and *GSS* on a large collection of XML documents from Wikipédia encyclopedia.