Une distance hiérarchique basée sur la sémantique pour la comparaison d'histogrammes nominaux

Camile Kurtz*

*Université de Strasbourg, LSIIT
ckurtz@unistra.fr

Résumé

La plupart des distances entre histogrammes sont définies pour comparer des histogrammes ordonnés (dont les entités représentées sont totalement ordonnées) ou des histogrammes nominaux (dont les entités représentées ne peuvent pas être comparées). Cependant, il n'existe aucune distance qui permette de comparer des histogrammes nominaux dans lesquels il est possible de quantifier des valeurs de proximité sémantique entre les entités considérées. Cet article propose une nouvelle distance permettant de pallier ce problème. Dans un premier temps, une hiérarchie d'histogrammes, obtenue par le biais d'une fusion progressive des entités considérées (prenant en compte leurs proximités sémantiques), est construite. Pour chaque étage de cette hiérarchie, une distance standard de comparaison d'histogrammes nominaux est calculée. Finalement, pour obtenir la distance proposée, ces différentes distances sont fusionnées en prenant en compte la cohérence sémantique associée aux niveaux de chaque étage de la hiérarchie. Cette distance a été validée dans le cadre de la classification de données géographiques. Les résultats obtenus sont encourageants et montrent ainsi l'intérêt et l'utilité de cette dernière pour des processus de fouille de données.

Summary

The usual distances defined for histogram comparison are generally devoted either to ordinal histograms (related to entities equipped with a total ordering) or nominal histograms (related to entities which can not be compared). However, there does not exist any distance for nominal histograms related to entities whose semantic/thematic proximity can be quantified. In this article, we propose a new distance devoted to this issue.