SoftJaccard : une mesure de similarité entre ensembles de chaînes de caractères pour l'unification d'entités nommées

Christine Largeron, Bernard Kaddour, Maria Fernandez

Université de Saint Etienne, F-42000, Saint-Etienne, France
Laboratoire Hubert Curien, UMR CNRS 5516
Christine.Largeron | Bernard.Kadour | Maria.Fernandez@univ-st-etienne.fr

Résumé. Parmi les mesures de similarité classiques utilisables sur des ensembles figure l'indice de Jaccard. Dans le cadre de cet article, nous en proposons une extension pour comparer des ensembles de chaînes de caractères. Cette mesure hybride permet de combiner une distance entre chaînes de caractères, telle que la distance de Levenstein, et l'indice de Jaccard. Elle est particulièrement adaptée pour mettre en correspondance des champs composés de plusieurs chaînes de caractères, comme par exemple, lorsqu'on se propose d'unifier des noms d'entités nommées.

1 Mesures entre ensembles de chaînes de caractères

Différentes mesures peuvent être employées pour comparer deux ensembles de chaînes de caractères S et T selon qu'on les traite comme des chaînes de caractères, des ensembles d'éléments ou réellement comme des ensembles de chaînes de caractères.

Si on les assimile à deux chaînes de caractères, alors, on peut avoir recours à la distance de Levenstein (Levenstein (1966)). Mais cette approche s'avère inappropriée si T et S correspondent à des noms composés de plusieurs mots, puisqu'il serait souhaitable alors de ne pas respecter l'ordre de ces mots.

Pour ce faire, on peut mesurer la similarité entre les ensembles T et S, à l'aide de l'indice de Jaccard défini comme le rapport entre le nombre de mots communs à S et T et le nombre total de mots figurant dans S et T (Jaccard (1901)). On peut aussi assimiler S et T à deux ensembles de mots (*bags of word*) et faire appel à la mesure TF-IDF, issue de la fouille de texte et de la recherche d'information (Salton et McGill (1983)). L'inconvénient des mesures de Jaccard et TF-IDF est qu'elles exigent une correspondance parfaite entre chaque chaîne figurant dans S et T. Pour pallier ce défaut, des distances hybrides ont été introduites visant à concilier distance entre chaînes de caractères et mesure entre ensembles de mots. SoftTF-IDF, introduite par Bilenko et *al.* (Bilenko et al. (2003)), en est un exemple. Mais, un des inconvénients de cette mesure, comme d'ailleurs TF-IDF, dont elle est dérivée est qu'elle nécessite le prétraitement du corpus pour déterminer le pouvoir discriminant de chaque mot. Or ce prétraitement n'est pas toujours réalisable ou peut s'avérer coûteux en temps de traitement. C'est ce qui nous a conduit à proposer la mesure SoftJaccard.