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.

2 SoftJaccard

Etant donnés S et T deux ensembles de chaînes de caractères tels que $S = s_1..s_{|S|}$ avec s_i la $i$ème chaîne de caractères de S, de longueur $|s_i|$, avec $1 \leq i \leq |S|$, la mesure SoftJaccard, que nous proposons pour comparer deux ensembles de chaînes de caractères, est définie par :

$$SoftJaccard(S,T,\alpha,\beta) = \frac{|S \cap T| + \sum_{\{s \in S-(S \cap T)/|s|>\alpha\}} max_{\{t \in T/|t|>alpha, delta(s,t)>\beta\}} delta(s,t)}{|S \cup T|}$$

avec : $delta(s,t) = 1 - delta'(s,t)$ et $delta'(s,t) = d(s,t)/max(|s|,|t|)$ où d est la distance de Levenstein. Cette mesure de similarité prend la valeur 1 si deux affiliations sont identiques, 0 si elles sont totalement différentes. A la différence d'autres mesures hybrides, du même type telles que celles basées sur TF-IDF qui requièrent le calcul préalable du pouvoir discriminant de chaque mot figurant dans les données, SoftJaccard ne nécessite pas de disposer de l'ensemble des données. Cette mesure est bien adaptée pour l'unification de noms d'entités nommées. Il s'agit alors de détecter automatiquement les noms (ou appellations) qui ne sont pas nécessairement identiques textuellement bien qu'ils se rapportent à une même entité et de les unifier. Dans ce cadre, SoftJaccard présente l'avantage d'identifier deux ensembles de chaînes de caractères comme étant les noms d'une même entité même si les mots composant ces noms ne sont pas dans le même ordre et si certains mots ne sont pas orthographiés exactement de la même façon. Comparée à des mesures classiques telles que TF.IDF, SoftTF.IDF et Jaccard pour l'unification d'entités sur les bases test $Cora$ et $Restaurant$ disponibles sur Internet [1] et sur une base de données de réponse à des appels d'offre européen $PCRD$, cette mesure ne s'est révélée ni meilleure ni moins bonne que les autres. Par contre, à performance sensiblement identique, les mesures basées sur l'indice de Jaccard se sont avérées moins coûteuses en termes de temps de traitement. Les résultats de ces expérimentations seront détaillés lors de la conférence.

Références

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, et S. Fienberg (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems 18*(5), 16–23.

Jaccard, P. (1901). Bulletin de la société vaudoise des sciences naturelles. *37*, 241–272.

Levenstein, A. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady 10*, 707–710.

Salton, G. et M. J. McGill (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Summary

In this paper, we propose an extension of the Jaccard index for sets composed of strings of characters. This hybrid measure, called SoftJaccard, allows to combine distance between strings of characters, like Levenstein distance, and Jaccard index. This measure is suited for field matching problems.

[1] http ://www.cs.utexas.edu/users/ml/riddle/data.html