Clustering de textes et application à la fouille de données expérimentales

Souhil Chakar[1], Marianne Clausel[2], and Brice Olivier[3]

[1]souhil-chakar@imag.fr
[2]marianne.clausel@imag.fr
[3]brice.olivier@inria.fr

11 avril 2016

A partir de données expérimentales, on cherche à comprendre comment un individu qui découvre un texte va le parcourir et en extraire l'information sémantique. Pour cela, nous avons mené une expérience sur une trentaine d'individus qui ont eu à se déterminer le plus rapidement possible sur l'adéquation sémantique du texte au thème qui leur avait été présenté auparavant [3]. Les mouvements oculaires étaient relevés durant la lecture afin d'avoir accès à la séquence des mots fixés pour tous les textes lus. Un des facteurs qui rend cette question très délicate est la très grande variabilité des textes que nous avons considéré. Pour contourner cette difficulté, nous nous proposons de catégoriser les différents documents utilisés en prenant en compte l'évolution de la proximité sémantique entre le thème et le texte en fonction du temps, c'est à dire au fur et à mesure de sa lecture.

Actuellement, la méthode utilisée pour réaliser le clustering de textes consiste à tout d'abord extraire la sémantique des mots des textes via *Latent Semantic Analysis* (LSA)[2]. Puis, une mesure de similarité en cosinus est utilisée pour chaque mot afin de calculer la proximité entre les mots lus et le thème. Cela permet d'obtenir l'évolution de la similarité sémantique d'un texte au cours du temps suivant un indice temporel représenté par le rang du mot dans le texte. A partir de cela, le clustering est réalisée via *Hierarchical Ascendant Classification* (HAC) avec une métrique adaptée, *Dynamical Time Warping* (DTW) [4].

Le travail des étudiants se décomposera en trois parties. Premièrement, la prise en main des techniques actuelles. Deuxièmement, au lieu d'utiliser des similarités en cosinus entre les mots et le thème, nous souhaiterions exploiter des techniques plus fines, comme des ontologies. Celles-ci permettent de représenter un ensemble de concepts sous forme d'arbre. Enfin, les étudiants pourront proposer d'autres méthodes pour le clustering de courbes en s'inspirant d'articles récents [1][5]. Les méthodes développées s'appliqueront dans un premier temps en utilisant systématiquement tous les mots du texte. On testera ensuite la robustesse de la classification en utilisant les séquences effectives de mots fixés par chaque participant.

Mots-Clefs Machine Learning, Clustering de textes, Fouille de données.

Références

[1] Ahlame Douzal-Chouakria and Cécile Amblard. Classification trees for time series. *Pattern Recogn.*, 45(3) :1076–1091, March 2012.

[2] Deerwester S. et al. Indexing by latent semantic analysis. *Journal of the american society for information sciences*, 41(6) :391–407, 1990.

[3] Frey A. et al. Decision-making in information seeking on texts : an eye-fixation-related potentials investigation. *Front. Syst. Neurosci.*, 7(39), July 2013.

[4] Rabiner L. R. Myers C. S. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7) :1389–1409, 1981.

[5] Eric Gaussier Saeid Soheily-Khah, Ahlame Douzal-Chouakria. Generalized k-means-based clustering for temporal data under weighted and kernel time warp. *Pattern Recogn.*, 75 :63–69, May 2016.