Le Mans Université

lium

Stage : Une approche graphe/réseaux complexes pour modéliser la nouveauté dans des corpus textuels

Détails : Master 2, 6 mois, au LIUM situé au Mans, au sein de l'équipe LST. Encadrement par Nicolas Dugué (contacter nicolas.dugue@univ-lemans.fr).

Mots-clés : Graphes, réseaux complexes, Corpus textuels, Programmation Python.

Contexte : Le projet #neo s'intéresse à la détection de néologismes automatique en exploitant de grands corpus textuels. En particulier, il s'agit notamment de détecter des mots qui changent de sens ou dont un nouveau sens apparaît. Dans ce stage, nous souhaitons fournir à ce projet un moyen d'évaluer les méthodes de détection automatique en créant des *modèles* de corpus artificiellement générés. Ces modèles devront ressembler le plus possible à des corpus réels. Par ailleurs, ils devront nous permettre d'introduire nous mêmes des changements de sens, de façon à tester les méthodes de détection.

Le sujet :

Nous souhaitons développer un générateur de données artificielles basé sur une approche qui modélise les corpus textuels comme des réseaux (ou graphes) de co-occurrence textuels. Ainsi, les noeuds (sommets) du réseau (graphe) sont des mots, et les liens (arcs) entre ces mots, des co-occurrences dans le texte. La topologie de ces réseaux a été étudiée de près ces dernières années, et leurs propriétés ont ainsi été mises en lumière. Il est donc possible de créer des modèles de corpus dont les propriétés correspondent à ceux du réel. Ainsi, par exemple, dans les réseaux de co-occurrence non valués issus des corpus de la langue naturelle, on observe que les degrés des noeuds suivent une loi de puissance [1], parfois à deux vitesses [4] notamment expliquée par le modèle génératif de Dorogovtsev et Mendes [2]. De même, il a également été observé que le degré des noeuds des réseaux de co-occurrence issus de la langue naturelle sont corrélés négativement avec ceux de leurs voisins [5] : on parle de *disassortativité* [3, 6]. Cette propriété ne semble pas présente dans les réseaux générés artificiellement [8]. Enfin, les réseaux de co-occurrence possèdent une structure communautaire sous-jacente [9, 7] dont les communautés représentent des ensembles de concepts liés.

Les objectifs du stage sont ainsi :

- de confirmer les propriétés des réseaux listées ci-dessus sur les corpus du projet #neo ;
- de modéliser les changements dans ces réseaux dans le temps ;
- de se baser sur l'état de l'art et les modèles de génération de graphe pour proposer et développer une approche de génération artificielle de modèles de corpus.

References

[1] Jin Cong and Haitao Liu. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618, 2014.

[2] Sergey N Dorogovtsev and José Fernando F Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1485):2603–2606, 2001.

[3] Hai-Bo Hu and Xiao-Fan Wang. Disassortative mixing in online social networks. *EPL (Europhysics Letters)*, 86(1):18003, 2009.

[4] Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.

[5] AP Masucci and GJ Rodgers. Network properties of written human language. *Physical Review E*, 74(2):026102, 2006.

[6] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[7] Mark EJ Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004.

[8] Günce Orman and Vincent Labatut. The effect of network realism on community detection algorithms. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–305. IEEE Press, 2010.

[9] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.