CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 25/11/2016

Offre de stage

Sujet : Embedding de graphes et détection de communautés

Possibilité de poursuivre sur une thèse

Encadrement

Thomas Bonald

Lieu et dates du stage

LINCS, 23 avenue d'Italie, 75013 Paris

Date de début du stage : début 2017

Équipe(s) d'accueil de la thèse

LINCS (INRIA, Institut Telecom, UPMC, Alcatel-Lucent, SystemX)

Mots clés

Analyse de graphe, détection de communautés, *embedding* de graphe, clustering, machine learning, analyse de réseaux sociaux.

Sujet détaillé

L'*embedding* de graphes (littéralement le plongement de graphe) vise à représenter dans un espace vectoriel de faible dimension l'information souvent complexe contenue dans un graphe. Il s'agit d'un champ de recherche en plein essor [1][13][10], qui trouve des applications dans des domaines aussi variés que les réseaux sociaux (Facebook, Twitter), les bases de données (structure du Web ou de Wikipedia) ou la biologie (graphe d'interaction de protéines). L'*embedding* de graphe est un outil puissant dans de nombreux problèmes d'analyse de graphes, dont le principal est la détection de communautés, qui consiste à *clusteriser* l'ensemble des nœuds d'un graphe en des sous-ensembles denses [9][2]. En effet, en partitionnant les vecteurs dans l'espace d'*embedding*, on peut parvenir à détecter les communautés dans le graphe d'origine. Cependant, les communautés dans les graphes réels sont en général complexes. On observe très souvent des communautés se recouvrant partiellement ou des communautés imbriquées [8][5]. Ainsi, la détection de communautés à partir de l'information contenue dans l'espace d'*embedding* est généralement un problème difficile.

Récemment, plusieurs techniques d'*embedding* de graphe ont été proposées [3][10][4], dont les algorithmes DeepWalk [10] et node2vec [3] qui reposent sur l'algorithme Word2Vec, populaire en traitement naturel du langage. D'autres approches utilisent directement des marches aléatoires dans les graphes [11][12]. Dans le cas semi-supervisé par exemple, l'algorithme Personalized PageRank s'avère très efficace [7][14].

Le sujet du stage porte sur l'étude de propriétés de ces *embeddings* reposant sur des marches aléatoires, et sur leur application à la détection de communautés. Il s'agira d'étudier l'influence du type d'*embedding* et des méthodes de *clustering* utilisées dans l'espace de plongement sur la qualité du résultat, à partir des algorithmes présentés dans [4]. Le travail comportera une étude théorique sur des modèles de communautés dans les graphes [6][9], ainsi qu'une étude comparative sur des données réelles issues de réseaux sociaux (Youtube, DBLP [14]), ou sur des données complexes pouvant être naturellement représentées sous forme de graphes (Amazon, Wikipédia).

Le stage pourra se baser sur les implémentations suivantes d'*embedding* de graphe :

- https://github.com/aditya-grover/node2vec
- https://github.com/ahollocou/walkscan

La Chaire Machine Learning for Big Data

Le traitement statistique des masses de données convoque à la fois mathématiques appliquées et informatique, à travers une discipline en plein essor : le Machine Learning ou apprentissage statistique.

Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée à hauteur de près de 2 M€ par quatre entreprises partenaires : Criteo, PSA Peugeot Citroën, Safran et BNP Paribas, la Chaire Machine Learning for Big Data est portée par le mathématicien Stéphan Clémençon, Enseignant-Chercheur, Professeur au sein du Département du Traitement du Signal et des Images à Télécom ParisTech.

Proposant cinq axes de recherche méthodologiques, enrichis par des applications industrielles concrètes, cette Chaire a pour objectif d'animer, en interaction avec ses partenaires, une activité de recherche de pointe en Machine Learning, ainsi que de proposer des programmes de formation.

La variété des données aujourd'hui disponibles (nombres, images, textes, signaux), leur grande dimension et leur volumétrie rendent souvent inopérantes les méthodes statistiques traditionnelles reposant sur le prétraitement humain et un long travail de modélisation. Le Machine Learning vise donc à élaborer et étudier des algorithmes, à vocation prédictive le plus souvent, permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon performante.

Les avancées technologiques, l'omniprésence des capteurs (systèmes embarqués, objets connectés, Internet…) et l'explosion des réseaux sociaux s'accompagnent d'un véritable déluge de données, propulsant les sciences de l'information au centre du processus de valorisation des masses de donnés. En plus de la collecte et du stockage, l'enjeu est de pouvoir analyser ces données afin d'optimiser les décisions et mettre au point de nouvelles applications.

Au-delà du buzz médiatique dont il fait l'objet, le Big Data est donc un sujet stratégique majeur, au cœur d'enjeux économiques et sociétaux considérables. Son impact est désormais perçu dans presque tous les secteurs de l'activité humaine : de la recherche scientifique à la médecine en passant, entre autres, par la finance, le bâtiment, l'e-commerce, la défense ou les transports.

En savoir plus sur la Chaire, ses axes de recherche, ses activités, ses publications :

http://machinelearningforbigdata.telecom-paristech.fr

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique
- Analyse de graphes / analyse de réseaux sociaux
- Bon niveau en programmation (Python, et éventuellement Julia et C++)
- Bon niveau d'anglais

Candidatures

à envoyer à thomas.bonald@telecom-paristech.fr :

- Curriculum Vitae
- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)
- Relevés de notes des années précédentes
- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Référence

[1] AHMED, Amr, SHERVASHIDZE, Nino, NARAYANAMURTHY, Shravan, *et al.*Distributed large-scale natural graph factorization. In : *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013. p. 37-48.

[2] BLONDEL, Vincent D., GUILLAUME, Jean-Loup, LAMBIOTTE, Renaud, *et al.*Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, vol. 2008, no 10, p. P10008.

[3] GROVER, Aditya et LESKOVEC, Jure. node2vec: Scalable Feature Learning for Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[4] HOLLOCOU, Alexandre, BONALD Thomas, LELARGE Marc. Improving PageRank for local community detection. Preprint. arXiv : 1610.08722. 2016.

[5] JEUB, Lucas GS, BALACHANDRAN, Prakash, PORTER, Mason A., *et al.*Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Physical Review E*, 2015, vol. 91, no 1, p. 012821.

[6] KARRER, Brian et NEWMAN, Mark EJ. Stochastic blockmodels and community structure in networks. *Physical Review E*, 2011, vol. 83, no 1, p. 016107.

[7] KLOUMANN, Isabel M. et KLEINBERG, Jon M. Community membership identification from small seed sets. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014. p. 1366-1375.

[8] LANCICHINETTI, Andrea, FORTUNATO, Santo, et KERTÉSZ, János. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 2009, vol. 11, no 3, p. 033015.

[9] LANCICHINETTI, Andrea, RADICCHI, Filippo, RAMASCO, José J., *et al.*Finding statistically significant communities in networks. *PloS one*, 2011, vol. 6, no 4, p. e18961.

[10] PEROZZI, Bryan, AL-RFOU, Rami, et SKIENA, Steven. Deepwalk: Online learning of social representations. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014. p. 701-710.

[11] PONS, Pascal et LATAPY, Matthieu. Computing communities in large networks using random walks. In : *International Symposium on Computer and Information Sciences*. Springer Berlin Heidelberg, 2005. p. 284-293.

[12] ROSVALL, Martin et BERGSTROM, Carl T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 2008, vol. 105, no 4, p. 1118-1123.

[13] TANG, Jian, QU, Meng, WANG, Mingzhe, *et al.* Line: Large-scale information network embedding. In : *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015. p. 1067-1077.

[14] YANG, Jaewon et LESKOVEC, Jure. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 2015, vol. 42, no 1, p. 181-213.