Get Twitter for science - A la recherche du Twitter perdu

Contexte Avec l'explosion de l'utilisation des réseaux sociaux, ceux-ci sont de plus en plus étudiés par le monde académique. De nombreuses questions se posent par exemple quant au réseau Twitter :

- Comment s'organise le réseau Twitter ? Taille/Densité/Communautés [6]
- Quels en sont les utilisateurs influents ? [4, 2, 10]
- Quels sont les facteurs qui influencent le Retweet ? [8]
- Comment détecter le contenu de type spam et les spammeurs ? [3, 9]
- Quels utilisateurs participent à une course aux followers ? [5]

Pour répondre à ces questions, le monde académique a besoin de données utilisables. Deux snapshots du réseau Twitter ont été effectués : le premier par [4], le second par [7]. Ces deux snapshots datent de 2009 et ne sont donc plus d'actualité, notamment au vu des chiffres révélés sur la croissance de Twitter [1]. Certaines expérimentations tendent à montrer que le réseau/graphe se serait également fortement densifié (augmentation du nombre de followers par utilisateur).

Afin de répondre à toutes ces questions, il est donc temps de partir à la conquête des données de Twitter à nouveau ! Néanmoins, le géant ayant changé de politique, il n'est plus aussi simple d'obtenir des données.

Travail à réaliser L'objectif est de créer un outil de collecte des données de Twitter.

De façon plus précise, il s'agira de développer :

- Un portail web ;
- Un entrepôt de données ;
- Une application Twitter.

L'application Twitter aura la charge de la récupération des données. Celle-ci a fait l'objet d'un stage de L3 l'an dernier et ne demande plus qu'à être complétée et mise à jour.

Le portail web permettra aux visiteurs de voir en temps réel les informations concernant l'avancement de la récupération des données, ainsi que des statistiques concernant la portion de graphe collecté. Une interface *dynamique* type HTML5 est à privilégier. Les visiteurs pourront également se connecter à l'application Twitter via leur propre compte Twitter. Ceci permettra à l'application de disposer de comptes supplémentaires pour la collecte de données.

Enfin, l'entrepôt de données devra fournir un moyen sûr, durable et efficace de stocker les données collectées. Des moyens d'exporter et de sauvegarder ces données seront ainsi mis en place. Une partie de ces travaux ont également été étudiés l'an dernier.

Encadrement Vous serez encadrés durant ce TER par Nicolas Dugué (équipe Contraintes et Apprentissage) et Anthony Perez (équipe Graphes, Algorithmes et Modèles de Calcul).

Moyens à disposition des étudiants

- Une application Twitter déjà développée en marche
- Une base de données MYSQL en marche

Contact nicolas.dugue@univ-orleans.fr

Références

[1] The Telegraph : Twitter in numbers, 2013. `http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html`.

[2] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer : quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, 2011.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter : The Million Follower Fallacy. In *ICWSM '10 : Proc. of int. AAAI Conference on Weblogs and Social*, 2010.

[5] Nicolas Dugué and Anthony Perez. Detecting social capitalists on twitter using similarity measures. In *Complex Networks IV*, volume 476 of *Studies in Computational Intelligence*, pages 1–12. 2013.

[6] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter : understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, 2007.

[7] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media ? In *Proc. of the 19th int. conference on World wide web*, WWW '10, pages 591–600, 2010.

[8] B. Suh, Lichan Hong, P. Pirolli, and Ed H. Chi. Want to be retweeted ? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184, 2010.

[9] Alex Hai Wang. Don't follow me : Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10, 2010.

[10] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank : finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, 2010.