SONDY : une plateforme open-source d'analyse et de fouille pour les réseaux sociaux en ligne

Adrien Guille, Cécile Favre, Djamel Abdelkader Zighed

SONDY : une plateforme open-source d'analyse et de fouille pour les réseaux sociaux en ligne

Adrien Guille*, Cécile Favre*
Djamel Abdelkader Zighed**

*Laboratoire ERIC - Université Lyon 2
{adrien.guille|cecile.favre}@univ-lyon2.fr,
**Institut des Sciences de l'Homme de Lyon, Laboratoire ERIC - Université Lyon 2
Abdelkader.Zighed@ish-lyon.cnrs.fr

Résumé. Ce papier décrit la plateforme SONDY qui permet l'analyse et la fouille de données issues de réseaux sociaux en ligne. La plateforme permet d'explorer et de visualiser l'évolution des thématiques populaires et la structure du réseau social de manière interactive. Elle permet également de comparer les méthodes utilisées et d'en intégrer de nouvelles. La démonstration consistera à expérimenter tous les services de SONDY sur des données issues de Twitter.

1 Introduction

Les réseaux sociaux en ligne permettent à des millions d'internautes à travers le monde de produire et consommer des contenus en temps réel. Ce flux continu de contenus est un mélange non structuré d'idées, d'informations, d'opinions, *etc*. Etant donné l'impact de ces réseaux sur la société, il est devenu important de pouvoir fouiller efficacement ces données. C'est pourquoi de nombreuses approches ont été proposées récemment afin de pouvoir identifier à partir de cette masse d'information, notamment, des thématiques émergentes, des évènements ou encore des personnes influentes ou des communautés.

1.1 Constats et besoins

Bien que de nombreuses méthodes aient été proposées, nous constatons la difficulté de pouvoir les réutiliser simplement. Pour cette raison, il apparaît pertinent de proposer une plateforme les intégrant et permettant de les manipuler. Deux types de public sont concernés : d'une part, les utilisateurs finaux, tels que des journalistes ou des analystes médias, qui souhaitent explorer l'activité sociale, et d'autres part les chercheurs qui désirent expérimenter et comparer des méthodes d'analyse et de fouille sur ces données. Un outil de veille basé sur l'analyse de la diffusion d'informations permettrait aux journalistes de mieux appréhender les informations circulant sur un média social comme Twitter par exemple. Une plateforme open-source permettrait aux chercheurs d'implémenter leurs algorithmes sans se soucier de la gestion des données en entrée, des moyens de visualisation en sortie, et leur permettrait également de les comparer aisément avec d'autres.

1.2 Proposition : SONDY

Nous proposons une plateforme open-source permettant à la fois de traiter les flux de messages et la structure associés aux réseaux sociaux (FIG. 1), nommée SONDY [1] (*i.e.* SOcial Networks Dynamics, "sondy" étant également le terme tchèque pour sonde), développée avec le langage JAVA (environ 10000 lignes de code) en raison de sa simplicité et de sa grande compatibilité. Le stockage des données se fait avec une base de données MySQL [2] et l'indexation avec la librairie LUCENE [3]. La visualisation interactive de graphe utilise l'API GraphStream [4].

2 Architecture de la plateforme

Les données d'entrée de l'application sont un ensemble de messages accompagnés de leur date de publication (donc un flux de messages) et leur auteur, ainsi que le graphe social connectant ces derniers. Pour traiter ces données, la plateforme se décompose en quatre services :

1. *Le service de manipulation des données*, pour importer et préparer les données afin d'en optimiser leur utilisation par les autres services.
2. *Le service de détection de thématiques*, pour identifier et localiser temporellement des thématiques populaires.
3. *Le service d'analyse du réseau*, pour observer la structure du réseau des auteurs et trouver, par exemple, des personnes influentes ou des communautés.
4. *Le service de gestion des extensions*, pour importer de nouveaux algorithmes utilisés par le service de détection de thématiques ou le service d'analyse du réseau.

2.1 Service de manipulation des données

Ce service gère une collection de jeux de données et permet non seulement d'en importer de nouveaux, mais également de les pré-traiter en vue de leur future exploitation par les autres services. Lorsqu'un nouveau jeu de données est importé dans la collection, l'application le stocke dans une base de données indexée *ad-hoc*. Les filtres proposés sont les suivants :

1. *Discretisation temporelle du flux de messages*, afin de pouvoir appliquer les méthodes se basant sur le calcul de la fréquence des termes.
2. *Redimensionnement du flux de messages*, afin d'étudier un extrait d'un jeu de données.
3. *Suppression des mots outils*, pour nettoyer les données, en enlevant par exemple des termes trop courants ou des termes spécifiques à la source de données, selon une liste intégrée ou à la discrétion de l'utilisateur.
4. *Stemming*, pour désuffixer les termes afin d'améliorer l'efficacité de certains algorithmes de détection de thématiques.

1. http://eric.univ-lyon2.fr/~aguille/Adrien_Guille_-_Software.html
2. http://www.mysql.com
3. http://lucene.apache.org
4. http://graphstream-project.org

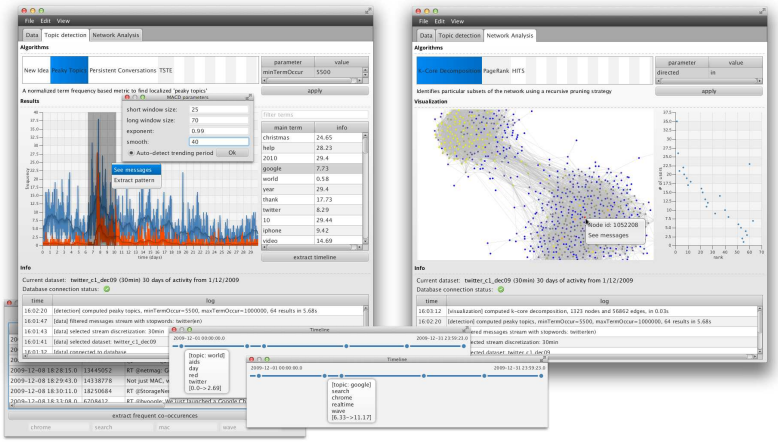

FIG. 1 – *SONDY : détection de thématiques (à gauche), analyse du réseau (à droite).*

2.2 Service de détection de thématiques

Ce service permet d'appliquer différents algorithmes sur un jeu de donnée choisi afin d'extraire des thématiques d'intérêt. Les résultats des algorithmes peuvent être exportés pour être comparés ensuite, leur temps de calcul est affiché, et il est possible de les explorer selon plusieurs moyens : en parcourant la liste classée des thématiques détectées, en générant des "timelines" pour résumer l'information, en sélectionnant une thématique en particulier puis en visualisant l'évolution de sa popularité dans le temps et en la comparant avec d'autres. Les algorithmes actuellement implémentés sont les suivants :

1. *Peaky Topics* (Shamma et al., 2011), pour détecter des thématiques très populaires sur une période très localisée.
2. *Persistent Conversations* (Shamma et al., 2011), pour détecter des thématiques moins saillantes mais qui continuent de générer de l'activité plus longtemps.
3. *TSTE (Temporal and Social Term Evaluation)* (Cataldi et al., 2010), pour détecter des thématiques émergentes en prenant en compte l'autorité des auteurs (à l'aide de l'algorithme PageRank).

Afin d'aider à la localisation temporelle des périodes de popularité des thématiques, l'application implémente également l'indicateur MACD (Lu et al., 2012) (Moving Average Convergence Divergence).

2.3 Service d'analyse du réseau

Ce service permet de visualiser le réseau des auteurs en rapport avec la thématique et la période sélectionnées dans le service de détection. Les graphes présentés sont colorés, il est possible de se déplacer ou zoomer, ainsi que de sélectionner les nœuds pour les identifier ou explorer leurs messages. Pour colorer les graphes et ainsi en faire ressortir la structure, les algorithmes suivants sont implémentés :

1. *K-Cores Decomposition* (Batagelj et Zaversnik, 2003), un algorithme pour identifier des sous-ensembles particuliers du graphe appelés *k-cores*. Les plus grandes valeurs de k correspondent aux nœuds les plus centraux du réseau.
2. *PageRank* (Page et al., 1998), un algorithme classique pour quantifier l'autorité des nœuds au sein du réseau.

2.4 Service de gestion des extensions

SONDY propose une interface de programmation permettant d'implémenter de nouveaux algorithmes, qui fournit les définitions à respecter ainsi que les moyens permettant de manipuler les données. L'import de nouveaux algorithmes se fait à l'aide d'une classe compilée en format JAR. Une fois l'algorithme importé via le service de gestion des extensions, il apparaît dans l'interface utilisateur, ce qui permet de l'appliquer de manière interactive et de faire varier ses paramètres.

3 Perspectives

L'idée étant de disposer d'une plateforme comparative des méthodes, l'objectif est bien sûr d'implémenter au fur et à mesure les méthodes les plus pertinentes d'analyse de ces données pour enrichir la plateforme, ainsi que de tester nos propres contributions dans ce domaine. Il sera intéressant par la suite d'intégrer au service de manipulation de données la possibilité de récupérer des données directement depuis des réseaux sociaux en ligne. On pourra aussi intégrer de nouveaux services, comme des services d'inférence ou encore de prédiction du graphe de diffusion par thématique.

Références

Batagelj, V. et M. Zaversnik (2003). An o(m) algorithm for cores decomposition of networks.

Cataldi, M., L. Di Caro, et C. Schifanella (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. MDMKDD '10, pp. 4–13.

Lu, R., Z. Xu, Y. Zhang, et Q. Yang (2012). Trends predicting of topics on twitter based on macd. IACSIT '12, pp. 44–49.

Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. WWW '98, pp. 161–172.

Shamma, D. A., L. Kennedy, et E. F. Churchill (2011). Peaks and persistence : modeling the shape of microblog conversations. CSCW '11, pp. 355–358.

Summary

This paper presents the SONDY platform that allows to analyse and mine both messages stream produced by a social network and its structure.