Découverte de points d'intérêts à partir de médias sociaux géo-localisés

Mehdi Kaytoue, Jean-François Boulicaut

Projet 4IF-Fouille de données – Année scolaire 2014-2015

1 Contexte

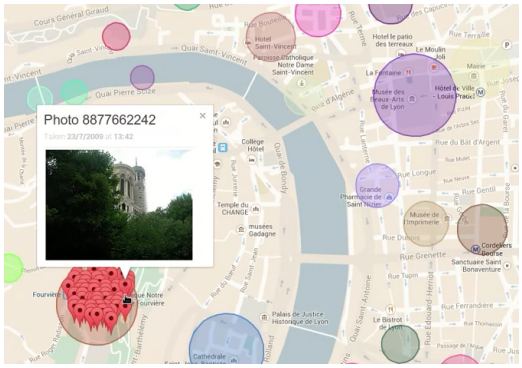

FIGURE 1 – Résultat final d'un projet 4IF (2013)

Les applications Web, smart phones et tablettes fleurissent pour fournir des services divers et variés. Certaines utilisent la masse d'information des réseaux sociaux (Facebook, twitter, instagram, ...) pour proposer des services où la géolocalisation des médias en question joue un rôle crucial. Ces *distilleries du Web social* filtrent la masse de messages pour n'en garder que l'essence, ou valeur ajoutée (e.g. 500 millions de tweets par jour en 2015). Les collectivités territoriales et gouvernements sont aussi intéressés par la valorisation de ces masses : on peut monitorer les mouvements de foules dans une ville, suivre une épidémie de *dengue* au Brésil, découvrir des évènements et utilisateurs d'influence sur les réseaux sociaux, etc. En fait, les possibilités d'application ne sont limitées que par notre imagination : les marques suivent leur popularité, des services d'emplois mettent relation employeurs/employés [2], des évènements sont détectés, des galeries géo-localisées personnalisées sont créées (e.g. `www.tapastreet.com`), etc.

2 Données

Dans un souci d'améliorer ses transports en communs et la vie des touristes visitant Lyon, le Grand Lyon vous demande de trouver de manière non-intrusive les zones à fortes densités de touristes à moindre cout. On imagine ici une architecture capable de récupérer des informations à partir du Web (crawling, scraping), comme des photos géo-localisées. Il faut alors trouver de manière automatique les points d'intérêt principaux à partir d'une large collection de photographies géo-localisées. En effet, 3000 photos prises autour de la tour Eiffel correspondent à un unique point d'intérêt. Pour cela, vous avez déjà réalisé une collecte de médias géo-localisés (quelle efficacité !) à travers l'API du service Flickr de Yahoo. Vous disposez de plus de 80,000 photos prises au cours de plusieurs années. Chaque photo est décrite comme un tuple :

$$\langle id\_photo, id\_photographe, latitude, longitude, tags, description, dates \rangle$$

3 Travail demandé

Votre mission est de trouver de manière automatique des points d'intérêts intéressants dans la ville de Lyon, définis par une activité forte de prise de photos. Pour cela, on veillera à détailler chaque étape du processus de KDD (à l'aide du logiciel Knime) :

- Compréhension, nettoyage des données, visualisation et statistiques (doublons, incohérences, ...)
- Sélection des attributs et objets intéressants pour l'analyse courante
- Fouille de données avec du *clustering* : comparer, discuter *k-means*, *clustering hiérarchique*, et *DBSCAN*. Les étudiants les plus motivés pourront utiliser l'algorithme *Mean-Shift* de la librairie python *scikit-learn*.
- Décrire les clusters obtenus non plus en extension, mais en intension (*itemsets* ou outils de traitement du langage – *natural langage processing* –).
- Évaluation, interprétation, visualisation (sur une carte), discussion des résultats. Comment votre analyse peut-elle aider le Grand Lyon ? Quelles connaissances lui apporte-t-elle ?

La dernière étape est souvent négligée, mais elle est capitale. Un résultat de fouille de données ne sert à rien s'il n'est pas *actionnable* : il doit servir à quelque chose, et le mode d'emploi doit être donné. On pourra seulement alors parler éventuellement de connaissances découvertes.

Références

[1] Advanced GIS : Web GIS. API Access : Flickr, Tutorial. `http://gis.yohman.com/up206b/tutorials/api-access-flickr/`.

[2] Article de le monde. `http://www.lemonde.fr/economie/article/2015/02/25/votrejob-quand-twitter-s-aventure-sur-le-terrain-de-pole-emploi_4582863_3234.html`.

[3] Autre démo étudiante, ucbl, lyon. `http://liris.cnrs.fr/mehdi.kaytoue/sujets/ter-meanshift/demo1.html`.

[4] Data publica : Crawling et au scraping (livre blanc). `http://www.data-publica.com/content/2013/09/le-livre-blanc-de-data-publica-consacre-au-crawling-et-au-scraping/`.

[5] Démo d'un excellent projet 4IF, INSA de Lyon. `https://www.youtube.com/watch?v=aM-zhxyVE54`.

[6] Saket Anand, Sushil Mittal, Oncel Tuzel, and Peter Meer. Semi-supervised kernel mean shift clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6) :1201–1215, 2014.

[7] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8) :790–799, 1995.

[8] Dorin Comaniciu and Peter Meer. Mean shift : A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5) :603–619, 2002.

[9] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339. ACM, 2007.

[10] Arijit Khan, Xifeng Yan, and Kun-Lung Wu. Towards proximity pattern mining in large graphs. In *SIGMOD*, pages 867–878. ACM, 2010.

[11] Zhenhui Li, Ming Ji, Jae-Gil Lee, Lu-An Tang, Yintao Yu, Jiawei Han, and Roland Kays. Movemine : Mining moving object databases. In *SIGMOD*, pages 1203–1206. ACM, 2010.

[12] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext : A location predictor on trajectory pattern mining. In *KDD*, pages 637–646. ACM, 2009.

[13] Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas S. Huang. Diversified trajectory pattern ranking in geo-tagged social media. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 980–991. SIAM / Omnipress, 2011.

[14] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800. ACM, 2009.