4-IF-FD – Projet de fouille de données

Fouille de données du Web : Découverte de points d'intérêts à partir de medias sociaux géo-localisés

Mehdi Kaytoue – Jean François Boulicaut – 2013/2014

Contexte

Depuis quelques années, les applications Web ou smart-phones fleurissent pour fournir des services divers et variés. En exemple récent, le service Mapado permet à un utilisateur de trouver des activités dans une ville donnée. Encore en version béta, le service Tapastreet permet à tout utilisateur géo-localisé de trouver des photos de points d'intérêt à visiter à sa proximité. Dans un tel cas, on imagine un système capable de récupérer des informations à partir du Web (crawling, scraping), comme des photos géo-taguées. 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.

Concepts principaux (mais non limité à !)

- **Clustering**
  - Partitionnements avec K-means, clustering hiérarchique
  - Approches « densité » : DBSCAN et Mean Shift
- **Evaluation de clusterings**
- **Motifs fréquents et règles d'associations**
- **Visualisation**
- **Knime, Sci-Kit Learn (python), Web Api (Google, Bing, Yahoo, …)**

Objectifs et résultats attendus

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. Pour cela, vous avez déjà réalisé une collecte de médias géo-localisés (photos) à travers l'API du service Flickr de Yahoo. Vous disposez donc d'une base NoSQL (Cassandra) de plus de 80 000 photos prises au cours des 3 dernières années. Chaque photo est décrite comme un tuple :

<id_photo,id_photographe,latitude,longitude,tags, description, dates>

A partir de l'exportation de cette base, votre mission est de

1. **Préparer**, nettoyer, décrire les données (doublons, incohérences, distributions…)
2. **Analyser** : trouver de bons clusters caractérisant des points d'intérêts.
3. **Evaluer, comparer** les résultats de clustering
4. **Décrire les clusters obtenus** : non plus par extension, mais par intension. Motifs fréquents et règles à partir des tags et descriptions seront utiles. Il est même possible d'utiliser d'autres services Web (e.g. Google Places)
5. **Visualisation des résultats** : Projection des clusters sur une carte (Google maps, Bing, ou manuelle)
6. **Interprétation des résultats** : Comment votre analyse peut-elle aider le Grand Lyon ? Quelles connaissances lui apporte-t-elle ?
7. **Aller plus loin :** Passage à l'échelle ? Tâches prédictives ? Analyse dynamique et non statique ? Autres sources ? (Tweets, Instagram,…)

Conseils : Utiliser Knime pour **1.** Utiliser Knime(+Weka) et Sci-Kit pour **2.**
Utiliser Knime/Weka/SciKit pour **3.** mais une discussion est primordiale
Utiliser Knime ou LCMv5.3 (http://research.nii.ac.jp/~uno/) pour **4.**
Utiliser Javascript/Google maps pour **5.** (http://gis.yohman.com/up206b/tutorials/api-access-flickr/)