

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

J.-F BOULICAUT ET M. KAYTOUE

Travaux pratiques 4IF-FD – Année scolaire 2015–2016

1 Contexte



FIGURE 1 – 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 événements et utilisateurs d'influence sur les réseaux sociaux, etc. Les entreprises cherchent aussi à évaluer automatiquement la présence de leur marque dans les différents réseaux sociaux, identifier les acteurs influents, *hashtags* spontanés inconnus, etc. En fait, les possibilités d'application ne sont limitées que par notre imagination : des services d'emplois mettent relation employeurs/employés [2], des événements sont détectés, des galeries géo-localisées sont créées, 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 coût. 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 (votre capteur *social*, 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 : $\langle id_photo, id_photographe, latitude, longitude, tags, description, dates \rangle$.

Table "flickr.zip" -- Rows: 83851 Spec -- Columns: 16 Properties Flow Variables															
Row ID	D id	S user	D lat	D long	S tags	S title	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...
Row0	22,653,655,0...	77161041@N...	45.768	4.802	square,sierra,squareformat,i...	Enfin. #instabeer #beer #chimay #ap...	46	18	24	11	201				
Row1	22,884,818,2...	113280318@...	45.76	4.842	square,squareformat,iphone...	https://www.facebook.com/PascalFro...	3	17	24	11	201				
Row2	23,277,598,0...	132999708@...	46.028	4.7	compagnons_dev_arnas20 (1)		0	15	7	11	201				
Row3	22,883,485,2...	132999708@...	46.028	4.7	compagnons_dev_arnas20 (3)		1	15	7	11	201				
Row4	23,249,102,1...	133835212@...	45.699	4.475	sunset,sky,cloud,sun,soleil,c...	Un soir dans les Monts du Lyonnais	20	20	31	8	201				
Row5	23,243,740,7...	129394312@...	45.763	4.85	france,architecture,lyon,offic...	InCity, Lyon, France, 2015	11	16	7	9	201				
Row6	22,642,697,4...	19710808@N...	45.739	4.814	orange,building,architecture,...		29	12	25	6	201				
Row7	22,972,701,4...	35210768@N...	45.763	4.827	square,squareformat,iphone...	@Bidule_officiel C'est à la Renaissance...	2	23	23	11	201				
Row8	22,971,623,1...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	55	13	3	10	201				
Row9	22,971,621,9...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	54	13	3	10	201				
Row10	22,873,337,7...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Martin-pêcheur d'Europe (Alcedo atthis)	39	13	3	10	201				
Row11	22,873,336,0...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Martin-pêcheur d'Europe (Alcedo atthis)	39	13	3	10	201				
Row12	23,267,456,3...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Martin-pêcheur d'Europe (Alcedo atthis)	38	13	3	10	201				
Row13	22,873,332,5...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row14	22,639,030,9...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row15	23,241,316,7...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row16	23,241,315,0...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row17	22,971,608,6...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row18	22,640,326,5...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row19	23,241,309,2...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				
Row20	23,267,441,0...	124810342@...	45.587	4.774	france,animaux,fr,oiseau,rho...	Grand Cormoran (Phalacrocorax carbo)	33	13	3	10	201				

FIGURE 2 – Échantillon brut du jeu de données à votre disposition

3 Découverte de points d'intérêt grâce au clustering

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. Il faudra par exemple : vérifier la cohérence des données (dates, positions GPS) ; supprimer les doublons, afficher les points sur une carte monde, ... On utilisera entre autres les nœuds *File Reader*, *GroupBy*, *Row Filter*, *Geo-Coordinate Row Filter*, *OSM Map View*, *Missing Value*.
- Sélection des attributs intéressants pour l'analyse courante (*Column Filter*).
- Fouille de données avec du *clustering* : comparer, discuter *k-means*, *clustering hiérarchique*, et *DBSCAN*. On utilisera les nœuds *k-Means*, *Color Manager*, *Color Appender*, *OSM Map View*, *Hierarchical Clustering*, *DBScan 3.x*, *Weka Cluster Assigner*, *Missing Value*.
- 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.

4 Description des points d'intérêt grâce à la fouille de motifs

Si l'étape précédente nous a permis d'extraire des points d'intérêt candidats intéressants, une étape de validation/compréhension est manquante. On va alors chercher à décrire les clusters obtenus non plus en extension, mais en intension. Pour cela, on utilisera le tutoriel proposé par Knime sur la fouille de texte. On construit alors une matrice document/terme que l'on peut rendre binaire. On peut chercher des *motifs fréquents* de termes pour chaque cluster, ou encore pour aller plus loin *des motifs discriminants*.

5 Ressources utiles

- Description de l'API de Flickr [1]
- Récupération de données à partir du Web [4]
- Exemple de résultats sur le jeu de données [3] [5]
- Lecture scientifique pour aller plus loin *Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, Thomas S. Huang : Diversified Trajectory Pattern Ranking in Geo-tagged Social Media. SDM 2011 : 980-991* [6]

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] Z. Yin, L. Cao, J. Han, J. Luo, and T. 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 / Ompress, 2011.