

Analyse de Données Sociales et Suivi de Clusters dans les Réseaux Sociaux

Erick Stattner

Laboratoire LAMIA - EA4540
Université des Antilles
France
erick.stattner@univ-antilles.fr

Pointe-à-Pitre, Décembre 2016



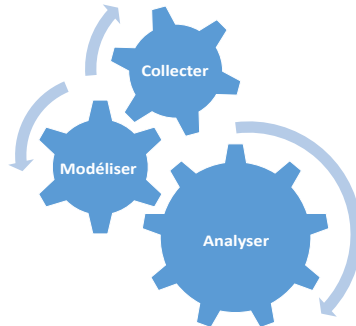
Introduction

Contexte :

- Explosion des études sur les réseaux
 - ▶ **Études sur** : réseaux d'amitiés, de collaboration, d'achats, de communications, d'échanges, ...
- Naît de l'observation que :
liens sociaux = facteurs déterminants dans l'évolution de nombreux phénomènes
 - ▶ Problème de diffusion (*rumeur, maladie, etc.*)
 - ▶ Phénomène d'achat (*lien social > attributs démographiques*)
 - ▶ La prise de décision (*lien social peut déterminer un comportement*)
- **La nouvelle science des réseaux** [Barabasi,2002]
Ensemble des méthodes qui s'intéressent aux interactions

Introduction

Principaux axes de recherche



Escalade de la collecte de données sociales

- Outils communautaire : Twitter, Facebook, Instagram, etc.
- Site de e-commerce : Amazon, Google, etc.
- Périphériques divers : déplacements, activités, etc.

Sommaire

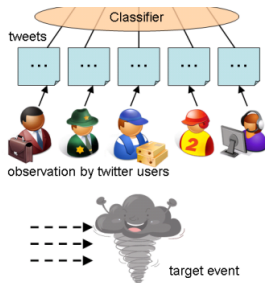
- 1 Analyser les données sociales
 - Identifier des événements
 - Prédire un événement
 - Etudier les comportements
 - Aller plus loin
- 2 Clustering de liens et suivi des clusters
- 3 Résultats expérimentaux
- 4 Conclusion et perspectives

Analyser les données sociales

Identifier des événements

Détecter les tremblements de terre au Japon [Sakaki et al., 2010]

- Détection plus rapide que l'agence nationale
- Implémenté dans un système qui fournit des notifications

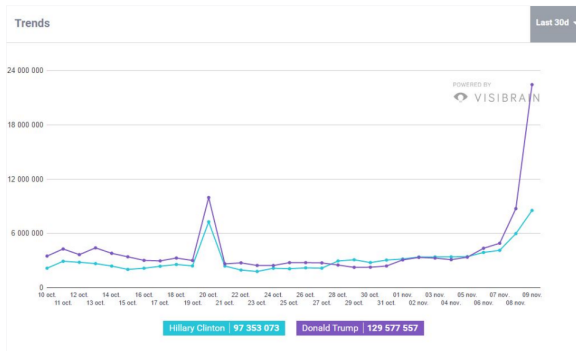


Anlyser les données sociales

Prédire un événement

Résultat d'élections [Tumasjan et al., 2010]

- Collecte messages sur les politiciens et les partis en Allemagne
- Corrélation entre le volume et le résultat

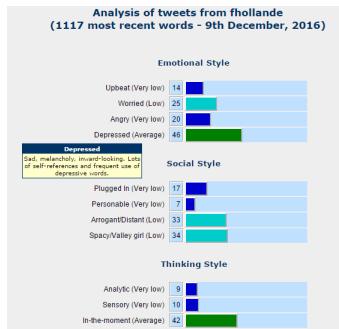


Analyser les données sociales

Etudier les comportements

Dresser un profil psycho-sociologique [Qiu et al., 2012]

- Extraire des Tweets des indicateurs de personnalités
- 3 styles : émotionnel, social, réflexion



Anlyser les données sociales

Etudier les comportements

Identifier quelqu'un grâce à ses déplacements [Blondel et al., 2014]

- 4 points suffisent pour identifier 95% des individus
- *"Nos données de déplacements sont encore plus personnelles que nos empreintes digitales."*



Anlyser les données sociales

Aller plus loin

Accueil / News / Monde / Twitter utilisé pour prédire crimes et délits

Twitter utilisé pour prédire crimes et délits

Par Direct Matin | Mis à jour le 29 Septembre 2016 à 08:59 | Publié le 29 Septembre 2016 à 08:16



D'ici trois ans, la police de Los Angeles pourrait être la première à tester cette nouvelle méthode. [Andrew Burton / GETTY IMAGES NORTH AMERICA / AFP]



Prédire à l'avance les crimes et délits en scannant des données récoltées sur Twitter. Aux Etats-Unis, les autorités viennent de se lancer dans un vaste chantier de surveillance numérique pour prévenir des faits de délinquance.

DERNIÈRE MINUTE

- 22:58 Le footballeur Antoine Conté mis en examen après une violente agression
- 22:43 Koh-Lanta : tout savoir sur l'épreuve des poteaux
- 22:39 Syrie : Daesh est de retour aux portes de Palmyre
- 22:22 Australie : sa maison est entièrement détruite par erreur
- 22:07 Attentat déjoué en France : un troisième suspect présenté à la justice antiterroriste
- 21:42 «Football Leaks» : Pogba recourt au paradis fiscal des îles anglo-normandes
- 21:14 Hollande vante des résultats «impressionnants» contre Daesh

Direct Matin

Le club

DES PLACES VIP,
DES RENCONTRES AVEC LES ARTISTES,
DES AVANT-PRÉMIÈRES
ET D'AUTRES CADEAUX INÉDITS



Gagnez votre Wonderbox !

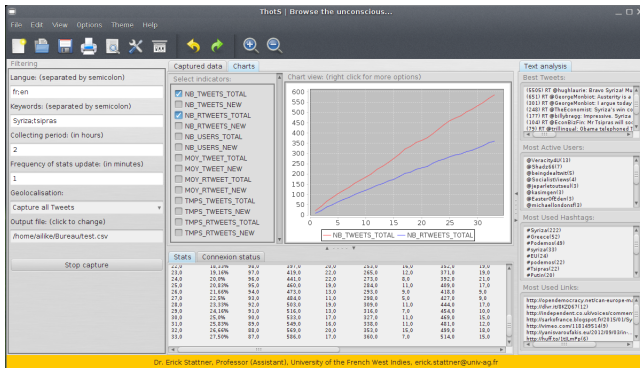
Rejoignez-nous

Analysar les données sociales

Etudier les comportements

Etudier diffusion [Stattner et al., 2015]

- Collecte des messages sur des sujets ciblés
- Extraire de la connaissance
- <http://erickstattner.com/thots-analytics/>



Dr. Erick Stattner, Professor (Assistant), University of the French West Indies, erick.stattner@univ-agg.fr

Sommaire

- 1 Analyser les données sociales
- 2 Clustering de liens et suivi des clusters
 - Clustering traditionnel dans les réseaux
 - Liens conceptuels
 - Suivi des liens conceptuels
- 3 Résultats expérimentaux
- 4 Conclusion et perspectives

Clustering de liens et suivi des clusters

Clustering traditionnel dans les réseaux

Extraction de clusters dans les réseaux

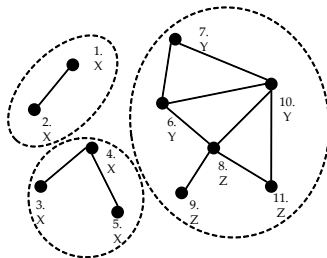
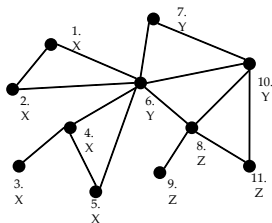
- Une des tâches les plus courantes
- Identifier des groupes de noeuds qui partagent des propriétés communes
- 2 grandes approches
 - ▶ Approche traditionnelle
 - ▶ Approche hybride

Clustering de liens et suivi des clusters

Clustering traditionnel dans les réseaux

Approche traditionnelle

- Extraction de communautés : groupes de noeuds fortement connectés
 - ▶ Algorithmes agrégatifs [Newman2003]
 - ▶ Algorithmes séparatifs [Fortunato2009]
 - ▶ Algorithmes basés sur des fonctions d'optimisation [Blondel2008]

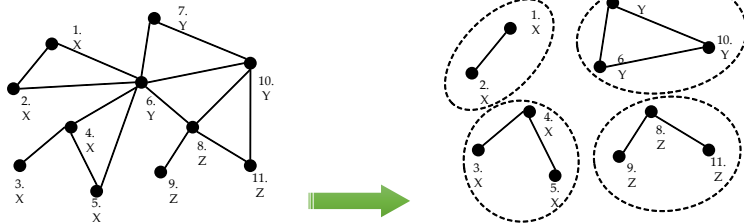


Clustering de liens et suivi des clusters

Clustering traditionnel dans les réseaux

Approche hybride

- Exploite structure et attributs
- Extraction de communautés dans lesquelles les noeuds partagent des propriétés communes
 - ▶ Idem + prend en compte une similarité interne [Zhou2009]



Clustering de liens et suivi des clusters

Clustering traditionnel dans les réseaux

Limites :

- Les motifs extraits ne permettent pas de répondre à des questions telles que :
 - ▶ Quels sont les groupes de noeuds les plus connectés ?
 - ▶ Quelles sont les caractéristiques les plus fréquemment retrouvées en connexion ?

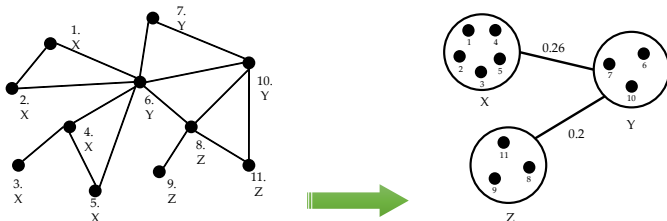
Clustering de liens et suivi des clusters

Liens conceptuels

Approche "liens conceptuels"

- Exploite structure et attributs
- Recherche des régularités dans les liens parmi des groupes de noeuds
- Extraire des **clusters de liens**

Groupe de noeuds (vérifiant certaines propriétés) fréquemment connecté à un autre groupe de noeuds



Clustering de liens et suivi des clusters

Liens conceptuels

Définition :

- $G = (V, E)$: Un réseau social
- V défini comme une relation $R(A_1, \dots, A_p)$ où A_i est un attribut
- Chaque noeud $v \in V$ est défini par un **itemset**
 $(A_1 = a_1 \text{ et } \dots \text{ et } A_p = a_p)$ ou (a_1, \dots, a_p)
- Soit m **itemset**
On note V_m l'ensemble des noeuds vérifiant la propriété m

Clustering de liens et suivi des clusters

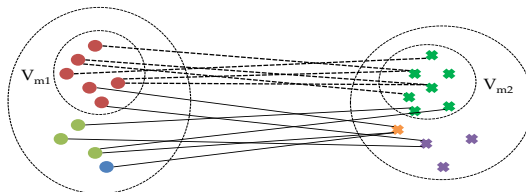
Liens conceptuels

Définition :

- Soient m_1 et m_2 **deux itemsets**

(m_1, m_2) : **Lien conceptuel** (*cluster de liens*)

$$(m_1, m_2) = \{e \in E; \quad e = (a, b) \quad a \in V_{m_1} \text{ et } b \in V_{m_2}\}$$



Lien entre deux concepts

Soit (m_1, m_2) un lien conceptuel

V_{m_1} : extension, i.e. l'ensemble des objets impliqués

m_1 : intension, i.e. l'ensemble des attributs partagés

Clustering de liens et suivi des clusters

Liens conceptuels

Définition :

- (m_1, m_2) : **lien conceptuel**

Support de (m_1, m_2) : Pourcentage de liens de type (m_1, m_2)

$$\text{support}[(m_1, m_2)] = \frac{|\{e \in E; \quad e = (a, b) \quad a \in V_{m_1} \text{ et } b \in V_{m_2}\}|}{|E|}$$

- β : **seuil de support des liens**

(m_1, m_2) est un **lien conceptuel fréquent (FCL)** ssi

$$\text{support}[(m_1, m_2)] > \beta$$

Clustering de liens et suivi des clusters

Liens conceptuels

Définition :

- (m'_1, m'_2) est un **sur-lien conceptuel** de (m_1, m_2) ssi

$$m_1 \subseteq m'_1 \quad \text{et} \quad m_2 \subseteq m'_2$$

Ex. (ab, b) *sur-lien conceptuel* de (a, b)

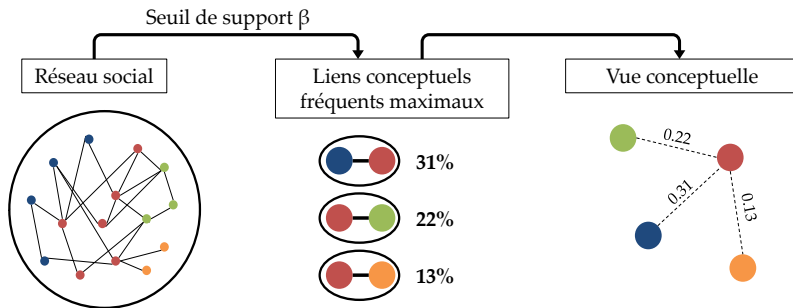
- (m_1, m_2) est un **sous-lien conceptuel** de (m'_1, m'_2)
- (m_1, m_2) **Lien conceptuel fréquent maximal (MFCL)** ssi
 \nexists pas de sur-lien conceptuel (m'_1, m'_2) de (m_1, m_2) qui soit fréquent

Clustering de liens et suivi des clusters

Liens conceptuels

Vue conceptuelle :

- Connaissance sur les groupes de noeuds les plus connectés
- Fournissent une "**vue conceptuelle**"

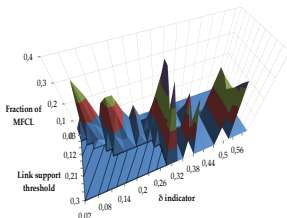


Clustering de liens et suivi des clusters

Liens conceptuels

Nos travaux récents

- Optimisation de l'algorithme [*IJISMD'*2013, *RCIS'*2013]



- Intersection avec clusters traditionnels [*ASONAM'*2013, *SNAM'*2014]

Questions ouvertes

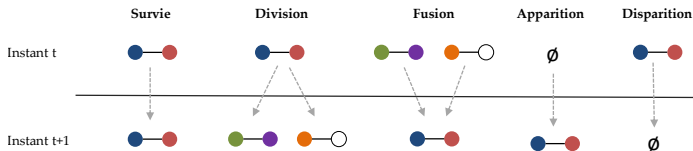
- Clusters extraits sur des réseaux statiques
- La plupart des réseaux évoluent
- Aucune information sur le devenir des clusters

Clustering de liens et suivi des clusters

Suivi des liens conceptuels

Évolution des liens conceptuels

- Comment les liens conceptuels évoluent sur les réseaux dynamiques ?
- **Objectif** : caractériser l'évolution des clusters entre l'état G_t et G_{t+1} du réseau
- 5 comportements identifiés



Clustering de liens et suivi des clusters

Suivi des liens conceptuels

Évolution des liens conceptuels

- On introduit la similarité entre deux liens conceptuels L et L'

$$\text{sim}(L, L') = \min\left(\frac{|L \cap L'|}{|L|}, \frac{|L \cap L'|}{|L'|}\right) \quad (1)$$

- Soit L un cluster extrait dans G_t , on note $\text{match}(L)$ l'ensemble des clusters de liens L' dans G_{t+1} dont la similarité avec L dépasse un seuil
 - Fusion** : L dans G_t fusionne avec d'autres clusters pour devenir L' dans G_{t+1} si $L' \in \text{match}(L)$ et $\exists Z \neq L$ dans G_t tel que $L' \in \text{match}(Z)$.
 - Division** : L dans G_t se divise en plusieurs liens conceptuels L'_1, L'_2, \dots, L'_k dans G_{t+1} si $\forall i, L'_i \in \text{match}(L)$.
 - Survie** : L dans G_t devient L' dans G_{t+1} si $L' \in \text{match}(L)$ et $\forall Z \neq L$ dans $G_t, L' \notin \text{match}(Z)$.
 - Disparition** : L dans G_t disparaît si aucun des cas précédents ne survient.
 - Apparition** : L' dans G_{t+1} apparaît si $\forall L$ dans $G_t, L' \notin \text{match}(L)$.

Sommaire

- 1 Analyser les données sociales
- 2 Clustering de liens et suivi des clusters
- 3 Résultats expérimentaux**
 - Environnement de tests
 - Exemple clusters extraits
 - Résultats
 - Outils d'extraction
- 4 Conclusion et perspectives

Résultats expérimentaux

Environnement de tests

Jeux de données utilisé

- Réseau de communications téléphoniques (Orange Caraïbe)
 - ▶ Noeuds : Abonnés
 - ▶ Liens : Appels téléphoniques
- Étude sur 10h : Journée du 1^e Juin de 5h du matin à 15h
- Chaque noeud est caractérisé par 10 attributs
 - 1 numéro
 - 2 localisation (Martinique, Guadeloupe ou Guyane)
 - 3 tranche horaire sur laquelle il est le plus actif
 - 4 type de forfait
 - 5 nombre moyen d'appels passés
 - 6 durée moyenne des appels passés
 - 7 nombre moyen d'appels reçus
 - 8 durée moyenne des appels reçus
 - 9 nombre de sms envoyés
 - 10 nombre de sms recus

Résultats expérimentaux

Environnement de tests

Jeux de données utilisé

- de 6 786 noeuds à 246 253 noeuds
- de 3 799 liens à 255 947 liens

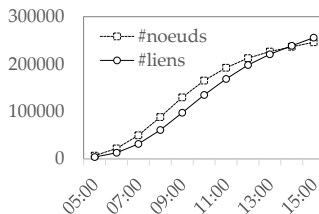


FIGURE – Évolution du nombre de liens et de noeuds sur la période

Résultats expérimentaux

Exemple clusters extraits

Exemple de clusters extraits

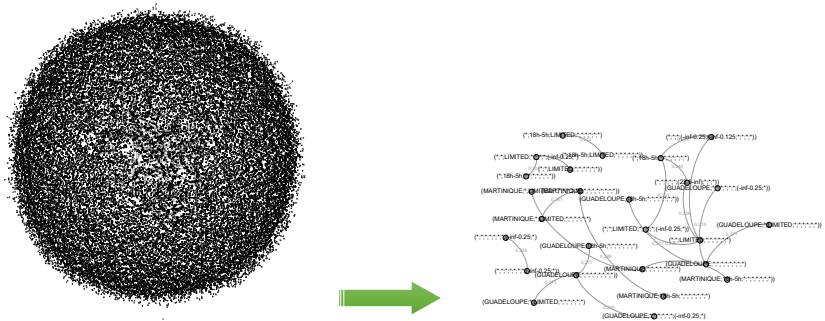


FIGURE – Extraction des liens conceptuels à 9h, avec $\beta = 0.2$

Résultats expérimentaux

Exemple clusters extraits

Exemple de clusters extraits

Un lien conceptuel obtenu

Support = 0,209

$(*,* ; 18h-5h ; *,* ; *,* ; *,* ; *,* ; *) (*,*,* ; LIMITED ; *,* ; *,* ; *,* ; (-inf-0.25 ; *))$

20% des appels sont passés entre des individus actifs sur la tranche
 $18h - 5h$
et des individus ayant un forfait limité et envoyant peu de SMS.

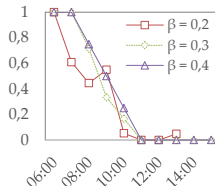
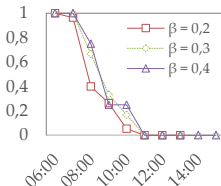
Comment évoluent ces clusters ?

Résultats expérimentaux

Résultats

Apparition et Disparition des clusters

- Tests avec 3 seuils : $\beta = 0.2$, $\beta = 0.3$ et $\beta = 0.4$
- Au début : clusters très instables
 - Premières heures : apparition et disparition à l'itération suivante

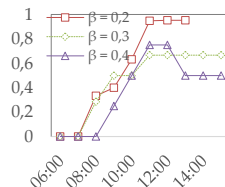
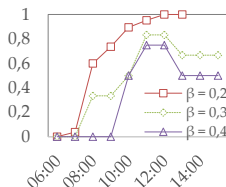
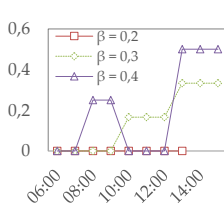


Résultats expérimentaux

Résultats

Survie, Fusion et Division des clusters

- Taux de survie relativement faible
- Bcp de fusion et de division
- Les clusters semblent se maintenir à travers la fusion et la division



Résultats expérimentaux

Résultats

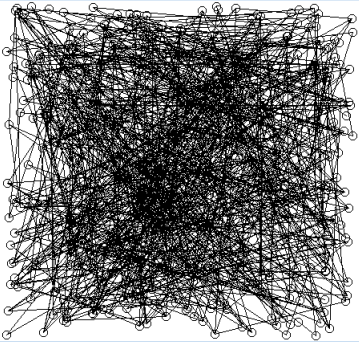
Outil GT-FCLMin

File Edit View Options Help

Calibrating

Network :
Geographical con...
Attributes :
6 attributes
Beta :
0.15
Measure :
Support
☐ By reducing space
Alpha :
0.1
☐ By querying
Start

Graphical mode Node Attributes



Frequent Links

1. Extraction of Frequent Links...
--
((*,*,2;*,*),(*,*,2;*,*)),[0.215]
((*,*,2;*,*),(*,*,1;*,*)),[0.295]
((*,*,2;*,*),(*,*,*,2;*,*)),[0.287]
((*,*,2;*,*),(*,*,*,1;*,*)),[0.223]
((*,*,1;*,*),(*,*,1;*,*)),[0.193]
((*,*,1;*,*),(*,*,*,2;*,*)),[0.294]
((*,*,1;*,*),(*,*,*,1;*,*)),[0.194]
((*,*,*,2;*,*),(*,*,*,2;*,*)),[0.343]
((*,*,*,2;*,*),(*,*,*,1;*,*)),[0.237]
((*,*,*,1;*,*),(*,*,*,1;*,*)),[0.181]
((*,*,1;1;*,*),(*,*,2;*,*)),[0.157]
((*,*,2;2;*,*),(*,*,*,2;*,*)),[0.194]
((*,*,2;2;*,*),(*,*,*,1;*,*)),[0.182]
((*,*,*,2;*,*),(*,*,*,2;2;*,*)),[0.194]
((*,*,*,1;*,*),(*,*,*,2;2;*,*)),[0.182]
((*,*,2;*,*),(*,*,*,1;1;*,*)),[0.157]
2. Summarizing...
--
[FL] = 16
Time = 0.269 sec

GT-FCLMin: Tool for Extracting Frequent Links in Social Networks

Erick STATNER - PhD Candidate, University of the French West-Indies

Sommaire

- 1 Analyser les données sociales
- 2 Clustering de liens et suivi des clusters
- 3 Résultats expérimentaux
- 4 Conclusion et perspectives

Conclusion et perspectives

Conclusion

- Explosion des travaux sur les données sociales
- De nombreuses études sur le clustering de réseaux sociaux
 - ▶ Hypothèse de réseaux statiques
- Contributions
 - ▶ Lien conceptuels : nouvelles approches de clustering de liens
 - ▶ Suivi des clusters de liens [Stattner et Collard, 2017]

Perspectives

- Étudier l'évolution des clusters sur des intervalles plus long et non-consécutif
- Améliorer l'algorithme d'extraction

Conclusion et perspectives

Merci de votre attention !