Analyse de la structure latente des réseaux sociaux par graphlets

Hiba Abou Jamra

Analyse de la structure latente des réseaux sociaux par graphlets

Hiba Abou Jamra

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté
9, Avenue Alain Savary
F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr

RÉSUMÉ.

L'exploitation des données des réseaux sociaux peut révéler des structures latentes qui peuvent être des précurseurs de changements de la structure du réseau ou de phénomènes de diffusion importants. Contrairement aux phénomènes de viralité qui atteignent très rapidement des niveaux de diffusion très importants, les signaux faibles ne sont pas détectables facilement par des outils statistiques simples. Dans cet article, nous présentons une approche pour détecter des signaux faibles en utilisant les graphlets. Nous utilisons un algorithme d'énumération en graphlets afin de montrer leur capacité à détecter des signaux faibles. À partir du cas concret de l'incendie de l'usine Lubrizol et des tweets qui en ont découlés, nous construisons une série temporelle extraite à partir des données Twitter, nous analysons les intervalles précédentes, pendant un évènement significatif en terme de vitesse de croissance/décroissance du nombre de graphlets.

ABSTRACT.

The exploitation of data from social networks reveals latent structures which can be precursors to changes in the structure of the network or to important diffusion phenomena. Unlike virality phenomena which reach quickly very important diffusion levels, weak signals are not easily detectable by simple statistical tools. In this article, we present an approach to detect weak signals using graphlets. We use the graphlet enumeration and decomposition algorithms to demonstrate their ability to detect weak signals. From the concrete case of the Lubrizol factory fire and the tweets that ensued, we build a time series extracted from Twitter data, we analyze the previous intervals, during a significant event in terms of rate growth / decrease in the number of graphlets.

MOTS-CLÉS : Signaux faibles, Graphlets, Structure des réseaux, Twitter, Détection d'évènements

KEYWORDS: Weak signals, Graphlets, Network structure, Twitter, Event detection

.

1. Introduction et Problématique

La détection de signaux faibles à partir d'informations cachées dans la masse de données produites quotidiennement sur les réseaux sociaux est un enjeu important puisqu'elle permet d'anticiper des prises de décision en matière de politique industrielle et commerciale et de stratégie de communication tout en projetant des scénarios d'avenir.

La première théorisation des signaux faibles a été proposée par Ansoff (Ansoff, 1975) qui place son étude dans le contexte de la planification et de la gestion des enjeux stratégiques des entreprises. Il définit les signaux faibles comme les premiers symptômes de discontinuités stratégiques qui agissent comme une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un évènement important. Lesca et Blanco (2002) présentent des caractéristiques permettant d'identifier un signal faible : **fragmentaire** car un seul fragment d'information sur un évènement susceptible d'être anticipé est disponible ; **visibilité faible** car le signal faible n'est pas perçu à côté des données qui font du bruit ; **peu ou pas familier** car le type d'information est non attendu, ni utilisé de manière répétitive ; **utilité faible** car l'information est apparemment dépourvue de signification opératoire, et son utilité ne saute pas aux yeux ; **fiabilité faible** car l'information n'est pas factuelle, mais perçue comme subjective et alors potentiellement erronée.

La quantité des données produite par les réseaux sociaux est si importante que les méthodes classiques s'appuyant sur des statistiques simples ne permettent pas d'extraire les signaux faibles. La construction d'outils algorithmiques travaillant plus localement est nécessaire. Nous avons envisagé trois approches possibles : les décompositions en ondelettes, les graphons (Glasscock, 2016) et les graphlets. Cependant, une décomposition en ondelettes sur une matrice d'adjacence ou de distance entre nœuds est difficilement interprétable. Les graphons s'appliquent plutôt sur des graphes denses, ce qui n'est pas le cas des graphes générés par les données des réseaux sociaux. Nous avons donc choisi de privilégier la piste des graphlets, notre hypothèse de travail est qu'ils permettent de déterminer des structures récurrentes ou patterns (petits graphes) qui se révèlent être des précurseurs d'évènements.

Afin de tester notre hypothèse, nous avons réalisé des expérimentations sur les données du projet ISITE Cocktail [1] dont le but est de créer un observatoire en temps réel des tendances, des innovations et des signaux faibles circulant dans les discours des contextes métiers alimentaire et santé sur Twitter.

2. Approche et Expérimentation

De nombreux travaux ont recherché des signaux faibles dans des documents en utilisant des méthodes de *text mining* telles que LSA et word2vec (Tonta, Darvish,

1. Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003), https://projet-cocktail.fr/

2010). Ils ont pour hypothèse que des mots-clés jouant le rôle de patterns, permettent de relier entre eux des documents et donc de trouver de l'information cachée. Nous faisons cette même hypothèse qu'il existe des patterns caractéristiques des signaux faibles, les graphlets, que l'on peut trouver à partir de la topologie du réseau.

Les graphlets ont été introduits pour la première fois par Pržulj *et al*. (2004). Un graphlet est un sous-graphe non isomorphe induit connecté (2 à 5 nœuds) choisi parmi les nœuds d'un large graphe. La figure 1 montre les 9 graphlets différents de 2 à 4 nœuds.

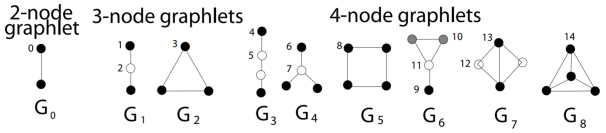

Figure 1. Représentation des 9 graphlets de 2 jusqu'à 4 nœuds

Il est ainsi possible de calculer pour un graphe une signature en terme de graphlets, c'est-à-dire en comptant le nombre de graphlets de chaque type qui apparaissent dans le graphe. Cette approche est appelée décomposition en graphlet.

Cas d'étude : incendie à l'usine Lubrizol et visite du président Macron à Rouen le 30 octobre 2019

Le 26 septembre 2019, une partie de l'usine Lubrizol à Rouen a été ravagée par un incendie, l'intérêt de l'évènement nous a amené à lancer une collecte qui nous a ramené environ 2 millions de tweets entre le 26 septembre et le 26 novembre 2019. Nous avons ensuite nettoyé ces données en filtrant avec les mot-clés `lubrizol` et `rouen` entre le 11 octobre et le 24 novembre. À l'issue de cette opération, le corpus est réduit à 137 561 tweets, parmi ces tweets, 16 100 sont des tweets originaux et 57 649 comportent des mentions. Nous avons choisi un période qui inclus un évènement important : la venue du président Macron à Rouen le 30 octobre au soir. Nous avons découpé notre corpus en trois périodes : P1 du 11 au 30 octobre midi (période avant l'évènement), P2 du 30 octobre midi au 3 novembre minuit (période autour de l'évènement), P3 du 4 au 24 novembre (période qui suit l'évènement). Les tailles des graphes de chaque période sont comparables. Afin de déterminer une signature topologique avant, pendant et après l'évènement, nous avons effectué les analyses suivantes : 1) une décomposition en graphlets de 2 à 4 nœuds (G_0 à G_8 figure 1) par pas d'un jour ; 2) pour observer l'évolution de la décomposition en graphlets, nous avons calculé les pentes normalisées comme suit : soit t un jour d'une période P, $Pente\_graphlets_{P(t)} = (G_{X_{(t)}} - G_{X_{(t-1)}})/G_{X_{(t-1)}}, X \in \{0, \dots 8\}$ et nous avons mesuré la pente pour observer la croissance des graphlets par rapport au nombre des mentions n dans le graphe initial : $Pente\_mentions_{P(t)} = (G_{X_{(t)}} - G_{X_{(t-1)}})/n, X \in \{0, \dots 8\}$.

Sur les 3 périodes nous avons constaté que les signatures issues de la décomposition en graphlet sont différentes. Nous nous sommes ensuite concentré sur la période P1 en incluant une partie de P2. Le tableau 1 est un extrait des pentes normalisées

par rapport au nombre de mentions entre le 23 et le 30 octobre (derniers jours de la période P1). Le calcul des pentes met en évidence deux singularités :

1. le 24 octobre la pente du graphlet de type G_5 devient positive avant celle des autres graphlets (valeur 0.543 mise en évidence dans le tableau), le jour suivant les pentes sont toutes positives. En regardant les évènements du monde réel, nous avons remarqué que le 25 octobre la société Lubrizol a beaucoup communiqué peut-être en réaction à un phénomène qui prenait de l'ampleur ;

2. Le 29 octobre (dernier jour de P1), la pente du graphlet de type G_6 devient positive et plus grande que celle des autres graphlets (valeur 2.13 mise en évidence dans le tableau), le lendemain (date de l'évènement qui nous intéresse) les pentes sont toutes positives.

Tableau 1. Pentes normalisées par mention (fin P1, début P2)

Jour	G0	G1	G2	G3	G4	G5	G6	G7
23/10	-0.550	-28.537	-0.041	-81.172	-1771.727	-0.562	-7.517	-0.631
24/10	-0.373	-23.305	0.020	-21.709	-1190.142	0.543	-9.415	-0.301
25/10	1.164	96.023	0.265	388.391	6704.945	48.029	52.198	6.501
26/10	-2.563	-207.338	-0.771	-875.825	-15158.076	-109.604	-122.724	-15.877
27/10	-1.976	-63.498	0.180	-124.576	-1915.135	-10.531	-2.220	1.151
28/10	0.301	1.837	-0.180	-15.277	15.017	3.419	-3.138	-1.183
29/10	0.195	-0.057	0.190	1.355	-13.057	-1.260	2.130	1.249
30/10	1.235	141.230	0.305	367.037	26899.298	6.197	152.724	5.765

3. Conclusion et Perspectives

Cette étude liminaire a permis de conforter notre hypothèse qui considère que les graphlets sont des précurseurs d'évènements et qu'ils peuvent être vus comme des signaux faibles. En effet, les graphlets sont de petits patterns caractéristiques qui présentent des anomalies dans leur nombre avant et autour de l'apparition d'un évènement. L'étude des trois périodes montre bien des signatures en terme de nombre de graphlets différents, les données et les programmes de l'expérimentation sont disponibles (https://github.com/hibaaboujamra/GraphletLubrizol). Cette première étude doit être poursuivie avec l'étude d'autres évènements comme par exemple la diffusion de `#sansmoile7mai` entre les deux tours de l'élection présidentielle de 2017. Les graphlets ont été catégorisés en chemin, triangulés, troués, etc., à partir de cette catégorisation, nous voulons pour chaque évènement et chaque type de graphlet précurseur d'évènement savoir si l'évènement aboutira à la construction d'une communauté (identification de graphlet de type troué et/ou triangulé) et/ou à une diffusion virale (graphlet de type chemin).

Bibliographie

Ansoff H. I. (1975). Managing strategic surprise by response to weak signals. *California management review*, vol. 18, n° 2, p. 21–33.

Glasscock D. (2016). *What is a graphon?* Consulté sur https://arxiv.org/abs/1611.00718

Lesca H., Blanco S. (2002). Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles. In *6è congrès international francophone sur la pme*, p. 10–1.

Pržulj N., Corneil D. G., Jurisica I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, vol. 20, n° 18, p. 3508–3515.

Tonta Y., Darvish H. R. (2010). Diffusion of latent semantic analysis as a research tool: A social network analysis approach. *Journal of Informetrics*, vol. 4, n° 2, p. 166–174.