TMD-MINER : Une nouvelle approche pour la détection des diffuseurs dans un système communautaire

Mohamed Nidhal Jelassi*,**, Christine Largeron*, Sadok Ben Yahia**

* Laboratoire Hubert CURIEN, Université Jean Monnet, Saint-Etienne, France.
Christine.Largeron@univ-st-etienne.fr,
**Faculté des Sciences de Tunis, Université de Tunis-El Manar, Tunis, Tunisie.
Sadok.Benyahia@fst.rnu.tn

Résumé. Plusieurs méthodes ont été développées ces dernières années pour détecter, dans un réseau social, les membres qualifiés, selon les auteurs, d'influenceurs, de médiateurs, d'ambassadeurs ou encore d'experts. Dans cet article, nous proposons un nouveau cadre méthodologique permettant d'identifier des diffuseurs dans le contexte où seule l'information sur l'appartenance des membres du réseau à des communautés est disponible. Ce cadre, basé sur une représentation du réseau sous forme d'hypergraphe, nous a permis de formaliser la notion de diffuseur et d'introduire l'algorithme TMD-MINER, dédié à la détection des diffuseurs et basé sur les itemsets essentiels.

1 Introduction

C'est en s'appuyant sur des représentations et des concepts issus de la théorie des graphes que les réseaux sociaux ont été étudiés en sciences sociales dès les années soixante (Moreno (1934), Cartwright et Harary (1977)). Parmi les questions essentielles que l'analyse de réseau s'efforce de traiter figure l'identification d'individus occupant un rôle déterminant dans le réseau. Avec l'explosion des réseaux sociaux sur Internet, des travaux plus récents se sont attachés à repérer des acteurs qualifiés selon les auteurs d'influenceurs, de médiateurs, d'ambassadeurs ou encore d'experts (Opsahl et Hogan (2010); Scripps et al. (2007a,b); Agarwal et al. (2008)).

En particulier, plusieurs algorithmes ont été présentés récemment pour résoudre le problème de recherche d'influenceurs, redéfini comme un problème de maximisation (Richardson et Domingos (2002); Kempe et al. (2003)). Cependant, ces algorithmes exploitent la matrice d'adjacence associée au graphe décrivant le réseau ; ce qui suppose connues les relations entre les acteurs pris deux à deux. Or, dans de nombreuses applications, cette information n'est pas forcément accessible. Par contre, on sait à quelle(s) communauté(s) appartient un acteur. On cherche alors à identifier des acteurs, appelés diffuseurs, en nombre le plus limité possible, qui appartiennent à plusieurs groupes et qui du fait de cette position sont susceptibles d'assurer des échanges d'un groupe à un autre. Il peut s'agir par exemple des chercheurs qui ont participé à la conférence EGC et ceux qui ont assisté à Coria ou encore des clients ayant acheté différentes

versions d'un article et pour lesquels on ne ne connait pas les liens directs existants entre ces personnes. Dans le premier cas, les diffuseurs seront susceptibles de diffuser des idées, dans le second de faire connaître aux utilisateurs de l'ancien modèle les avantages du nouveau et plus généralement d'émettre des recommandations d'un groupe à un autre. L'objectif de cet article est de proposer un cadre méthodologique permettant de détecter de tels diffuseurs et d'analyser le réseau dans ce contexte d'information incomplète où on ne dispose pas de la matrice d'adjacence associée au graphe représentant le réseau mais où, en revanche, les communautés sont données.

Pour ce faire, nous représentons le système communautaire sous la forme d'un hypergraphe dans lequel les sommets représentent les acteurs et les hyperarêtes représentent les communautés. Dans cet hypergraphe, les diffuseurs pourront être déterminés à partir des traverses minimales de l'hypergraphe, elles-mêmes définies comme un ensemble de sommets, minimal au sens de l'inclusion, qui intersecte toutes les hyperarêtes (Berge,1989). Dans la section 2, nous rappelons des notions-clés, issues de la théorie des graphes et de la fouille de données, que nous utilisons pour définir formellement la notion de diffuseur dans un hypergraphe puis, dans la section 3, nous introduisons notre approche pour la détection des diffuseurs et présentons un algorithme original, appelé TMD-MINER, pour le calcul des traverses minimales diffuseurs.

2 Définition d'un diffuseur dans un hypergraphe

Un réseau social peut être défini comme un ensemble d'entités interconnectées les unes avec les autres (Wasserman et Faust (1994)). Les relations décrivent des interactions entre les entités. Dans la pratique, on ne connaît pas toujours de façon précise les relations existants entre les entités prises deux à deux. En revanche, il peut exister des sous-groupes formant des communautés au sein du réseau et on peut savoir à quelle(s) communauté(s) appartient chaque entité. C'est dans ce cadre, que nous proposons de définir la notion de diffuseur au sein du réseau, en faisant appel aux concepts d'hypergraphe et de traverse minimale.

Définition 1 HYPERGRAPHE *(Berge (1989)) Soit le couple* $H = (\mathcal{X}, \xi)$ *avec* $X = \{x_1, x_2, \ldots, x_n\}$ *un ensemble fini et* $\xi = \{E_1, E_2, \ldots, E_m\}$ *une famille de parties de* $\mathcal{X}$. *H constitue un hypergraphe sur* $\mathcal{X}$ *si* $E_i \neq \emptyset$, $i \in \{1, \ldots, m\}$ *et* $\bigcup_{i=1,\ldots,m} E_i = \mathcal{X}$.

Les éléments x_i de $\mathcal{X}$, appelés sommets de l'hypergraphe, correspondent aux entités du réseau et les éléments de ξ, appelées hyperarêtes de l'hypergraphe, correspondent aux communautés.

Exemple 1 *La figure 1 illustre un hypergraphe* $H = (\mathcal{X}, \xi)$ *tel que* $\mathcal{X} = (1, 2, 3, 4, 5, 6, 7, 8)$ *et* $\xi = (\{1, 2\}, \{2, 3, 7\}, \{3, 4, 5\}, \{4, 6\}, \{6, 7, 8\}, \{7\})$.

Définition 2 ENSEMBLE TRANSVERSAL *(Berge (1989)) Un ensemble transversal (ou traverse) des arêtes d'un hypergraphe* $H = (\mathcal{X}, \xi)$ *où* $\xi = \{E_1, E_2, \ldots, E_m\}$ *est un ensemble* $T \subset \mathcal{X}$ *avec* $T \bigcap E_i \neq \emptyset \; \forall i = 1, \ldots, m$.

On note γ_H, *l'ensemble des traverses définies sur* H, *une traverse* T *de* γ_H *est dit* minimale *si il n'existe pas une autre transverse* S *telle que* S *est un sous-ensemble de* T. *Le nombre minimum de sommets d'un ensemble transversal est appelé le* nombre de transversalité *de l'hypergraphe* H *et on le désigne par :* $\tau(H) = min\,|T|$.

On notera par $\mathcal{M}_H$, *les traverses minimales définies sur* H.

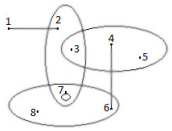

FIG. 1 – *Un exemple d'hypergraphe.*

Dans l'exemple illustré par la figure 1, l'ensemble $\mathcal{M}_H$ de l'hypergraphe est : { {1, 4, 7}, {2, 4, 7}, {1, 3, 6, 7}, {1, 5, 6 7}, {2, 3, 6, 7} et {2, 5, 6, 7} }. Il importe de mettre en exergue que dans le contexte d'un hypergraphe, où nous ne disposons que des communautés présentes dans le réseau, nous considérons qu'il doit y avoir au moins un diffuseur par communauté et que le nombre de ces diffuseurs doit être minimum. Ainsi, il apparaît qu'une traverse minimale constitue un cadre idéal pour localiser ces diffuseurs. Ceci nous conduit à formaliser la notion de traverse minimale diffuseur de la façon suivante.

Définition 3 TRAVERSE MINIMALE DIFFUSEUR *Soit $H = (\mathcal{X}, \xi)$, un hypergraphe et $I \subset \mathcal{X}$, on dit que I est une traverse minimale diffuseur, notée* TMD*, si I vérifie les trois conditions suivantes :*

1. *(Condition Nécessaire) I est une traverse minimale : $I \in \mathcal{M}_H$.*
2. *(Condition de composition) Le nombre de transversalité de I est minimal dans $\mathcal{M}_H$ au sens de la cardinalité : $\tau(H) = Min\ \{|T|, \forall T \in \mathcal{M}_H\}$.*
3. *(Condition de recouvrement maximum) : $\sum_{E_i / E_i \bigcap I \neq \emptyset} |E_i| = Max\ \{ \sum_{E_i / E_i \bigcap T \neq \emptyset} |E_i|, \forall\ T \in \mathcal{M}_H$ tel que $|T| = \tau(H)$ }.*

Ainsi, un ensemble de sommets constitue une TMD s'il forme une traverse minimale, si sa taille est la plus petite possible et s'il maximise la condition de recouvrement. Plus précisément, la première condition indique qu'il y aura au moins un diffuseur présent dans chaque communauté. La seconde condition vise à prendre un ensemble de diffuseurs le plus petit possible. Ainsi l'objectif est de représenter toutes les communautés à l'aide d'un minimum de sommets du graphe. Enfin, la troisième condition, dite de recouvrement maximum, permet, s'il existe plusieurs traverses minimales vérifiant la condition de composition, de privilégier celles qui sont composées de noeuds appartenant aux communautés de plus grandes tailles.

Avant de présenter TMD-MINER, nous allons définir la notion de contexte d'extraction, montrer son équivalence avec la notion d'hypergraphe et fournir une autre définition d'une traverse minimale en termes d'itemset essentiel, utilisée dans l'algorithme.

Définition 4 CONTEXTE D'EXTRACTION *Un contexte d'extraction est un triplet $\mathcal{K} = (\xi, \mathcal{X}, \mathcal{R})$ où ξ et $\mathcal{X}$ sont, respectivement, des ensembles finis d'hyperarêtes (ou transactions) et de sommets (ou items), et $\mathcal{R} \subseteq \xi \times \mathcal{X}$ est une relation binaire entre les hyperarêtes et les sommets.*

Ainsi, il existe une relation entre les notions d'hypergraphe et de contexte d'extraction. En effet, le contexte d'extraction $\mathcal{K} = (\xi, \mathcal{X}, \mathcal{R})$ associé à l'hypergraphe $H = (\mathcal{X}, \xi)$ permet de

représenter la matrice d'incidence de H :
$\forall x_j \in \mathcal{X}$ and $\forall E_i \in \xi$ on a : $\mathcal{R}(x_j, E_i) = 1$ si $x_j \in E_i$ et $\mathcal{R}(x_j, E_i) = 0$ si $x_j \notin E_i$.

Le tableau 1 représente le contexte d'extraction relatif à l'hypergraphe de la figure 1.

	1	2	3	4	5	6	7	8
$\{1,2\}$	1	1	0	0	0	0	0	0
$\{2,3,7\}$	0	1	1	0	0	0	1	0
$\{3,4,5\}$	0	0	1	1	1	0	0	0
$\{4,6\}$	0	0	0	1	0	1	0	0
$\{6,7,8\}$	0	0	0	0	0	1	1	1
$\{7\}$	0	0	0	0	0	0	1	0

TAB. 1 – *Le contexte d'extraction relatif à l'hypergraphe de la figure 1*

Définition 5 SUPPORT DISJONCTIF D'UN MOTIF ET MOTIF ESSENTIEL *Soient $\mathcal{K} = (\xi, \mathcal{X}, \mathcal{R})$ un contexte d'extraction et un motif formé d'un sous-ensemble non vide de sommets I de $\mathcal{X}$. Le support disjonctif du motif I, noté $Supp(\vee I)$, et désignant le nombre d'hyperarêtes (ou transactions) qui contiennent au moins un sommet (ou item) de I est défini par : Supp$(\vee I) = |\{E \in \mathcal{E} \mid (\exists\, x \in I, \mathcal{R}(x, E) = 1)\}|$. $I \in \mathcal{X}$ est un motif essentiel si et seulement si : $Supp(\vee I) > \max\{Supp(\vee I \backslash i) \mid i \in I\}$*

Corollaire 1 TRAVERSE ET TRAVERSE MINIMALE
Un motif $I \subseteq \mathcal{I}$ du contexte d'extraction $\mathcal{K} = (\xi, \mathcal{X}, \mathcal{R})$ est une traverse de l'hypergraphe $H = (\mathcal{X}, \xi)$ si son support disjonctif est égal au nombre d'hyperarêtes (ou transactions) de H : Supp$(\vee I) = |\xi|$. Un motif $I \subseteq \mathcal{I}$ est dit Traverse minimale si et seulement si I est un itemset essentiel dont le support disjonctif est égal au nombre d'hyperarêtes de H.

Dans la section suivante, nous proposons un algorithme de recherche des diffuseurs dans un hypergraphe basé sur une détection efficace des traverses minimales qui vérifient la condition de composition puis celle de recouvrement.

3 Méthodologie et algorithme d'extraction des diffuseurs

La recherche des diffuseurs dans un hypergraphe nécessite l'extraction des traverses minimales dans un hypergraphe. Berge a été le premier à proposer un algorithme (Berge (1989)), dont plusieurs extensions ont été présentées dans la littérature comme celle de (Kavvadias et Stavropoulos (2005)) ou celle de (Hébert et al. (2007)).

Dans ce qui suit, nous introduisons, un nouvel algorithme, appelé TMD-MINER pour l'extraction des traverses minimales diffuseurs basé sur la fonction $Apriori\_Gen$ (Agrawal et Ramakrishnan (1994)). L'algorithme TMD-MINER, dont le pseudo-code est décrit par l'Algorithme 1 prend en entrée un contexte d'extraction (ou hypergraphe) et donne en sortie l'ensemble des TMD. L'algorithme effectue un parcours en largeur, *i.e* il opère par niveau pour calculer les itemsets essentiels, et exploite la propriété d'idéal d'ordre vérifié par l'ensemble des itemsets essentiels.

A chaque niveau *k*, un appel à $Apriori\_Gen$ (Agrawal et Ramakrishnan (1994)), avec un support minimal égal à zéro, permet de calculer les candidats de taille *k*, à partir des itemsets

Algorithme 1: TMD-MINER

Données : $\mathcal{H}$: Contexte d'extraction $\mathcal{K} = (\xi, \mathcal{X}, \mathcal{R})$
Résultat : $\mathcal{TMD}$

```
1  début
2      L_1 := {1-itemsets};
3      i := 1;
4      si ∃x ∈ L_i/supp(x) = |ξ| alors
5          TMD = {x};
6          retourner TMD;
7      trouve = false ;
8      tant que L_i ≠ ∅ or trouve = false faire
9          C_{i+1} := Apriori_Gen(L_i);
10         L_{i+1} := {X ∈ C_{i+1} | ∄ x ∈ X : Supp(∨X) = Supp(∨X\x)};
11         pour chaque X ∈ L_{i+1} faire
12             si supp(∨X) = |ξ| alors
13                 TMD = TMD ∪ {x};
14                 trouve = true ;
15         i := i + 1;
16     TMD = Recouvrement(TMD) ; retourner TMD
```

essentiels de taille *k-1*. TMD-MINER calcule, ensuite, le support disjonctif des *k*-candidats, générés dans la *ligne 9*, et vérifie si ce dernier est strictement supérieur à ceux de ses sous-ensembles directs (*ligne 10*). Si parmi les itemsets essentiels calculés dans un niveau *k*, il en existe, au moins un, dont le support disjonctif est égal au nombre d'hyperarêtes de l'hypergraphe d'entrée *(ligne 12)*, alors la boucle de la *ligne 8* s'arrête, ne passe pas au niveau *k+1* et $\mathcal{TMD}$ contient l'ensemble des plus petites traverses minimales qui sont aussi minimales au sens des cardinalités. Ces traverses candidates vérifient donc la condition nécessaire et la condition de composition de la définition 3. Au final, les TMD sont déterminés, à partir des traverses minimales qui vérifient la condition de composition par la fonction *Recouvrement (ligne 16)*.

4 Conclusion

Dans cet article, nous avons introduit une nouvelle approche pour la détection des diffuseurs au sein d'un système communautaire représenté sous la forme d'un hypergraphe. Les perspectives de prolongement du présent travail sont nombreuses, à commencer par l'optimisation de TMD-MINER, en utilisant un encodage des données sous forme de nombres premiers et le PGCD et le PPCM comme opérateurs de décodage.

Remerciements Ce travail est partiellement soutenu par St-Etienne Metropole (`http://www.agglo-st-etienne.fr/`) et le projet Utique CMCU 11G1417

Références

Agarwal, N., H. Liu, L. Tang, et P. S. Yu (2008). Identifying the influential bloggers in a community. In *Proceedings of the International Conference on Web Search and web Data Mining (WSDM '08)*, Stanford, USA.

Agrawal, R. et S. Ramakrishnan (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, Santiago, Chili, pp. 487–499.

Berge, C. (1989). Hypergraphs : Combinatorics of finite sets. pp. 256.

Cartwright, D. et F. Harary (1977). A graph theoretic approach to the investigation of system-environment relationships. *Journal of Mathematical Sociology 5*, 87–111.

Hébert, C., A. Bretto, et B. Crémilleux (2007). A data mining formalization to improve hypergraph minimal transversal computation. *Fundamenta Informaticae. 80*(4), 415–433.

Kavvadias, D. J. et E. C. Stavropoulos (2005). An efficient algorithm for the transversal hypergraph generation. *Journal of Graph Algorithms and Applications 9*(2), 239–264.

Kempe, D., J. Kleinberg, et E. Tardos (2003). Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'03)*, Washington, USA, pp. 137–146.

Moreno, J. (1934). *Who shall survive ? : a new approach to the problem of Human Interrelations*, Volume 58 of *Nervous and mental disease monograph series*.

Opsahl, T. et B. Hogan (2010). Growth mechanisms in continuously-observed networks : Communication in a facebook-like community. *CoRR*.

Richardson, M. et P. Domingos (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the of the 8th International Conference on Knowledge Discovery and Data mining (KDD '02)*, Edmontom, Canada, pp. 61–70.

Scripps, J., P.-N. Tan, et A.-H. Esfahanian (2007a). Exploration of link structure and community-based node roles in network analysis. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM'07)*, Omaha, USA, pp. 649–654.

Scripps, J., P.-N. Tan, et A.-H. Esfahanian (2007b). Node roles and community structure in networks. In *Proceedings of the 1st Workshop on Web Mining and Social Network Analysis (SNA-KDD'07)*, San José, California, pp. 26–35.

Wasserman, S. et K. Faust (1994). *Social Network Analysis, methods and application*.

Summary

With the spread of collaborative systems of Web 2.0. looking for influencers is grasping a renewed interest. Indeed, the dedicated literature witnesses a wealthy number of approaches looking for influencers (*aka* mediators, ambassadors or experts). In this paper, we introduce a new approach to "unveil" influencers even in the absence of an adjacency matrix associated to graph-based network. The main thrust of the approach stands in its consideration of as hypergraph to reformalize the concept of influencer using the snugness connection with the "essentiel" pattern".