Détection efficace des traverses minimales d'un hypergraphe par élimination de la redondance

M. Nidhal Jelassi*, Christine Largeron**, Sadok Ben Yahia*

*Faculté des Sciences de Tunis, Tunis, Tunisia
{nidhal.jelassi, sadok.benyahia}@fst.rnu.tn
**Université de Lyon, F-42023, Saint-Étienne, France
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
christine.largeron@univ-st-etienne.fr

Résumé. L'extraction des traverses minimales d'un hypergraphe est une problématique réputée comme particulièrement difficile et qui a fait l'objet de plusieurs travaux dans la littérature. Dans cet article, nous établissons un lien entre les concepts de la fouille de données et ceux de la théorie des hypergraphes, proposant ainsi un cadre méthodologique pour le calcul des traverses minimales. Le nombre de ces traverses minimales étant, souvent, exponentiel même pour des hypergraphes simples, nous proposons d'en représenter l'ensemble de manière concise et exacte. Pour ce faire, nous introduisons la notion de traverses minimales irrédondantes, à partir desquelles nous pouvons retrouver l'ensemble global de toutes les traverses minimales, à l'aide de l'algorithme IMT-EXTRACTOR. Une étude expérimentale de ce nouvel algorithme a confirmé l'intérêt de l'approche introduite.

1 Introduction

La théorie des hypergraphes se propose de généraliser la théorie des graphes en introduisant le concept d'hyperarête où les arêtes ne relient plus un ou deux sommets, mais un nombre quelconque de sommets. De ce fait, un hypergraphe est défini comme une extension du traditionnel graphe, dans lequel les liens entre des ensembles de sommets sont appelés hyperarêtes. Une traverse minimale correspond à un ensemble de sommets qui intersecte toutes les hyperarêtes d'un hypergraphe, en étant minimal au sens de l'inclusion. L'extraction de ces traverses minimales a fait l'objet de plusieurs travaux dans littérature du fait de la diversité de ses applications dans des domaines variés tels que l'intelligence artificielle, la fouille de données, la cryptographie, le web sémantique, etc. Hagen (2008).

Pour résoudre ce problème, plusieurs approches ont été envisagées. Mannila et Toivonen (1997) ont prouvé que les traverses minimales d'un hypergraphe H peuvent être calculées à partir de la bordure négative d'un ensemble d'itemsets vérifiant la contrainte d'anti-monotonie Mannila et Toivonen (1997), Hébert et al. (2007). C'est en se basant sur ces liens entre la fouille de données et la théorie des hypergraphes que nous proposons une nouvelle définition des traverses minimales, ainsi que leur représentation de manière succincte.

Notre idée, inspirée des travaux de Hamrouni *et al.* est donc de chercher un sous-ensemble représentant de manière concise et exacte l'ensemble des traverses minimales Hamrouni et al. (2008). Ce sous-ensemble, qu'on appellera ensemble irrédondant de traverses minimales, sera construit en considérant l'ensemble des hyperarêtes auxquelles appartient chaque sommet, appelé extension et en construisant un hypergraphe irrédondant limité à un sous-ensemble des sommets initiaux ayant des extensions différentes. L'espace de recherche s'en trouve alors réduit puisque plusieurs candidats ne seront pas générés par l'algorithme. Cette intuition se base sur le fait que si deux sommets X et Y appartiennent aux mêmes hyperarêtes, (*i.e*, ils ont la même extension) et si X appartient à une traverse minimale T alors en substituant X par Y dans T on obtient une nouvelle traverse minimale. De plus, le risque de redondance est éliminé puisque Y n'est pas pris en compte dans l'exploration de l'espace de recherche.

Cet article est organisé comme suit : dans la section 2, nous proposons une nouvelle définition de la notion de traverse minimale et nous introduisons la notion de traverse minimale irrédondante qui nous permettra de proposer, dans la section 3, un algorithme original, appelé IMT-EXTRACTOR. Enfin, une étude expérimentale sur des hypergraphes aléatoires générés à partir des données issues du site de marque-page social DEL.ICIO.US sera décrite dans la section 4.

2 Traverses minimales irrédondantes

Dans cette section, nous proposons de présenter des définitions et notations que nous utiliserons tout au long des sections suivantes. Pour aboutir à notre approche d'extraction des traverses minimales, nous avons défini la notion de traverse minimale irrédondante à partir des notions de la théorie des hypergraphes suivantes.

Définition 1 HYPERGRAPHE *Berge (1989)*
Soit $H = (\mathcal{X}, \xi)$ avec $\mathcal{X}$ = $\{x_1, x_2, \ldots, x_n\}$ un ensemble fini d'éléments et ξ = $\{e_1, e_2, \ldots, e_m\}$ une famille de parties de $\mathcal{X}$. H constitue un hypergraphe sur $\mathcal{X}$ si :

1. $e_i \neq \emptyset, i \in \{1,\ldots, m\}$;
2. $\bigcup_{i=1,\ldots,m} e_i = \mathcal{X}$.

Définition 2 TRAVERSE MINIMALE ET NOMBRE DE TRANSVERSALITÉ *Berge (1989)*
Soit un hypergraphe $H = (\mathcal{X}, \xi)$. $T \subset \mathcal{X}$ est une traverse de H si $T \bigcap e_i \neq \emptyset \; \forall i = 1, \ldots, m$. γ_H désigne l'ensemble des traverses définies sur H. Une traverse T de γ_H est dite minimale *si $\nexists T_1 \subset T$ t.q. $T_1 \in \gamma_H$. On notera $\mathcal{M}_H$, l'ensemble des traverses minimales définies sur H.*

Le nombre minimum de sommets formant une traverse est appelé le nombre de transversalité *de l'hypergraphe H et on le désigne par : $\tau(H) = min\ |T|$.*

Définition 3 EXTENSION D'UN SOMMET
Soit un hypergraphe $H = (\mathcal{X}, \xi)$ et $x \in \mathcal{X}$. $E = (e_1, e_2, \ldots, e_l) \subseteq \xi$ est une extension de x si $x \in e_i$, $\forall\ e_i \in E$. Nous noterons EXTENT*(x), l'extension de x. Le lien entre l'extension d'un sommet x et son poids est donné par la formule : $Poids_{(\mathcal{H})}(x) = |Extent(x)|$. De plus, pour $X \subseteq \mathcal{X}$, on définit le poids de X de la façon suivante : $Poids_{(\mathcal{H})}(X) = |\ \{e \in \xi \mid (\exists\ x \in X,$ tel que $x \in e)\}\ |$*

Définition 4 CLASSE DE SUBSTITUTION
Une classe de substitution est un ensemble formé de tous les sommets de $\mathcal{X}$ qui ont la même extension. Ainsi, si deux sommets x_i et x_j appartiennent à des classes de substitution distinctes, alors EXTENT*(x_i) $\neq$* EXTENT*(x_j) et réciproquement, si deux sommets ont des extensions différentes, ils relèveront de classes de substitution différentes.*

Définition 5 REPRÉSENTANT D'UNE CLASSE DE SUBSTITUTION
Etant donnée une classe de substitution S, on dit que $x \in \mathcal{X}$ est le représentant *de cette classe S si x est le premier élément de la liste triée par ordre lexicographique des sommets qui forment la classe S.*

La méthode originale et efficace pour calculer les traverses minimales d'un hypergraphe proposée dans cet article, consiste à réduire l'hypergraphe en le représentant de manière plus concise mais sans perte d'information. Elle exploite le fait que parmi l'ensemble des traverses minimales, il y a une redondance d'information. Cette redondance est liée au fait que deux ou plusieurs sommets qui ont la même extension, (*i.e*, appartiennent exactement aux mêmes hyperarêtes) tiennent, à tour de rôle, la même position dans une traverse minimale mais ne peuvent y appartenir en même temps.

Notre approche pour une représentation concise des traverses minimales repose sur deux notions importantes : les classes de substitution et leurs *Représentants*. Les classes de substitution, calculées sur l'hypergraphe H=($\mathcal{X}$, ξ), permettent tout d'abord de construire un hypergraphe irrédondant H' associé à H.

Définition 6 HYPERGRAPHE IRRÉDONDANT ET TRAVERSE MINIMALE IRRÉDONDANTE
Soit l'hypergraphe H=($\mathcal{X}$, ξ), $\mathcal{X}' \subseteq \mathcal{X}$ l'ensemble des représentants des différentes classes de substitution associées aux sommets de H et ξ' l'ensemble des hyperarêtes de H privées de éléments de $\mathcal{X}$-$\mathcal{X}'$ et défini par $\xi' = \{e_i \cap \mathcal{X}', e_i \cap \mathcal{X}' \neq \emptyset, \forall e_i \in \xi\}$, alors l'hypergraphe H' = ($\mathcal{X}'$, ξ') est appelé hypergraphe irrédondant associé à H.

En remarquant que toute traverse minimale de H' constitue une traverse minimale irrédondante de H, on en déduit que $\mathcal{M}'_H$, l'ensemble des traverses minimales de H' (*i.e* des traverses irrédondantes de H) permet de construire l'ensemble des traverses minimales de H. En effet, toute traverse minimale irrédondante $T = \{x_1, ...x_l\}$ de H, composée de l représentants $x_i, i = 1, .., l$ des classes de substitution $S_i, i = 1, .., l$ permet de générer $\prod_{i=1,l} |S_i|$ traverses minimales en remplaçant les représentants de chaque classe de substitution par les autres éléments de la classe.
Dans la section suivante, nous introduisons, à partir du cadre méthodologique présenté, l'algorithme IMT-EXTRACTOR pour l'extraction des traverses minimales irrédondantes.

3 L'algorithme IMT-EXTRACTOR

L'algorithme IMT-EXTRACTOR, dont le pseudo-code, est décrit par l'Algorithme 1 prend en entrée un hypergraphe et fournit en sortie l'ensemble des traverses minimales. On suppose que les sommets de l'hypergraphe sont triés par ordre lexicographique. L'algorithme effectue un parcours en largeur, *i.e*, il opère par niveau pour générer les ensembles essentiels de candidats en exploitant la propriété d'idéal ordre vérifiée par les ensembles essentiels de sommets.

L'algorithme commence par calculer les extensions de chaque sommet de $\mathcal{X}$ à partir desquelles seront construites les différentes classes de substitution (ligne 3). Cette tâche est effectuée par la procédure SEARCH-SUBSTITUTION. Cette procédure fournit en sortie les différentes classes de substitution, conformément à la proposition 4 avec pour chaque classe la liste des sommets qui la compose et son *représentant*. A partir de ces données, un nouvel hypergraph $\mathcal{H}'$ est généré à l'aide de la procédure CHANGE-HYP suivant la définition 6. Opérant

Algorithme 1: IMT-EXTRACTOR

Entrées : H=($\mathcal{X}$, ξ)
Sorties : $\mathcal{MT}$, ensemble des traverses minimales de H

```
1  début
2      MT_irr = φ ;
3      Classes := SEARCH-SUBSTITUTION(H) ;
4      H' = CHANGE-HYP(H, Classes) ;
5      Level := GETMINTRANSVERSALITY(H') ;
6      C_level := GENERATE-CANDIDATES(H', level) ;
7      pour i ← level à | Classes | faire
8          pour chaque X ∈ C_i faire
9              si ∄ x ∈ X tel que Poids_(H')(X) = Poids_(H')(X\x) alors
10                 si Poids_(H')(X) = |ξ| alors
11                     MT_irr = MT_irr ∪ {X} ; C_i = C_i \ X ;
12             sinon
13                 C_i = C_i \ X ;
14         si C_i ≠ ∅ alors
15             C_{i+1} := APRIORI-GEN(C_i) ;
16     MT = GET-ALL-MT(MT_irr, Classes) ;
17     retourner MT
```

par niveau, l'algorithme IMT-EXTRACTOR invoque la procédure GETMINTRANSVERSALITY (ligne 5) pour calculer le nombre de transversalité, (cf. définition 2) noté $level$ dans l'algorithme, et qui correspond au niveau renfermant les plus petites traverses minimales. Le nombre de transversalité correspond ainsi à la taille des premières traverses minimales candidates. La procédure GENERATE-CANDIDATES détermine ensuite tous les sous-ensembles de sommets de $\mathcal{X}$ de cardinalité égale à $level$ (ligne 6). Ceci permet à l'algorithme d'éviter le balayage de l'espace de recherche compris entre 1 et $level - 1$.

Une fois les ensembles de sommets de taille $level$ générés, IMT-EXTRACTOR calcule le poids de chaque candidat avant de vérifier s'il est strictement supérieur au poids maximum de ses sous-ensembles directs ou pas. Si cette dernière propriété est vérifiée et que le poids du candidat est égal à la cardinalité de ξ (ligne 10), alors le candidat est stocké dans MT_{irr}, l'ensemble des traverses minimales irrédondantes et il est supprimé de l'ensemble des candidats (ligne 12). Si, de plus, le poids du candidat est strictement inférieur à la cardinalité de ξ alors il servira pour la génération des candidats de taille $level + 1$. Sinon, il est élagué de C_i (ligne 14). Le processus est itéré jusqu'à épuisement des candidats. A chaque itération, les candidats

de taille *i+1* (C_{i+1}) sont générés à partir de C_i par la procédure APRIORI-GEN(C_i) (Agrawal et Ramakrishnan (1994)). MT_{irr} représente les traverses minimales de H' calculée par IMT-EXTRACTOR autrement dit les traverses minimales irrédondantes de H. Il ne reste plus qu'à utiliser les classes de substitution pour générer l'ensemble complet des traverses minimales de H grâce à la procédure GET-ALL-MT(MT_{irr}, $Classes$) (ligne 17) décrite en fin de section précédente.

SEARCH-SUBSTITUTION($\mathcal{H}$) renvoie à l'algorithme IMT-EXTRACTOR le représentant et la liste des sommets composant chaque classe de substitution générée.

4 Etude expérimentale

Dans cette expérimentation, quatre hypergraphes ($H1$, $H2$, $H3$ et $H4$) ont été générés à partir des données DEL.ICIO.US en considérant que les sommets correspondent aux utilisateurs et que chaque hyperarête représente une communauté constituée par les utilisateurs ayant partagé, au moins une ou plusieurs pages web. Le tableau 1 détaille les caractéristiques de chacun de ces hypergraphes. Les deux premières colonnes fournissent les probabilités minimale (p_l) et maximale (p_u) qu'un sommet appartienne aux hyperarêtes dans l'hypergraphe. La troisième colonne et la quatrième colonne indiquent respectivement le nombre de sommes $|\mathcal{X}|$ et le nombre d'hyperarêtes $|\xi|$ de l'hypergraphe alors que la cinquième indique le nombre de *Représentants*, i.e., le nombre de classes de substitution. Ainsi, le nombre de sommets qui ne seront pas pris en compte lors de la génération des candidats est égal à $|\mathcal{X}|$ - *Repr*. Le nombre de transversalité τ*(H)* est donné dans la sixième colonne. Enfin, la septième colonne représente le nombre de traverses minimales alors que l'avant-dernière représente celui des traverses minimales irrédondantes. Pour ce qui est de la dernière colonne, nous y trouvons le taux de compacité, noté θ, de chacun des hypergraphes, et dont la valeur est donnée par la formule : $1- |\text{IMT}(H)| \ / \ |\mathcal{M}_H|$. θ représente donc le pourcentage de traverses minimales que notre approche élimine, sans perte d'informations.

	p_l	p_u	$\lvert\mathcal{X}\rvert$	$\lvert\xi\rvert$	*Repr*	τ*(H)*	$\lvert\mathcal{M}_H\rvert$	$\lvert$IMT$(H)\rvert$	θ
H1	0.05	0.13	95	38	65	8	3860	512	87%
H2	0.04	0.1	119	91	89	12	70226	6390	90%
H3	0.05	0.1	156	117	143	15	6010	848	86%
H4	0.03	0.06	248	179	206	23	95268	17700	81%

TAB. 1 – *Caractéristiques de nos jeux de données.*

Le fait que O-M2D cible directement le niveau correspondant au nombre de transversalité, grâce à la procédure GETMINTRANSVERSALITY, permet à l'algorithme d'éviter la génération de candidats et les tests associés dans des niveaux où il ne peut pas y avoir de traverses minimales. Ceci est surtout avantageux quand le nombre de transversalité est élevé comme c'est le cas pour **H3** (15) et, encore plus pour **H4** (23). Venons-en maintenant à IMT-EXTRACTOR. Le nombre des *Représentants* étant toujours inférieur ou égal au nombre de sommets, les temps d'exécution de IMT-EXTRACTOR s'en trouvent améliorés par rapport à ses concurrents. En analysant les cardinalités de $|\mathcal{M}_H|$ et de IMT*(H)*, et le taux de compacité θ, nous constatons,

pour l'hypergraphe **H1** par exemple, que IMT-EXTRACTOR génére 848 traverses minimales irrédondantes, alors que dans le même temps ses concurrents en génèrent 6010. Ce qui représente un taux de compacité de **86%**. IMT-EXTRACTOR parvient donc à représenter l'ensemble des traverses minimales en ne calculant que **14%** du nombre de traverses minimales calculées par ses concurrents. Les traverses éliminés sont retrouvées par notre algorithme, en utilisant les classes de substitution. Par conséquent, aucune perte d'information n'est à déplorer, ce qui permet à IMT-EXTRACTOR de représenter l'ensemble des traverses minimales de manière concise et exacte. Les taux de compacité des trois autres hypergraphes, **H2**, **H3** et **H4**, oscillent entre **81%** et **90%** puisqu'ils dépendent de la structure de ces hypergraphes et des valeurs de p_l et p_u. Généralement, plus ces valeurs sont faibles, plus la taille des hyperarêtes est petite et, donc, moins l'hypergraphe est dense. Ceci implique, d'après nos premières constatations, un meilleur taux de compacité et donc un nombre, plus élevé, de traverses minimales rédondantes.

Remerciements

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

Références

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* (3rd ed.). North-Holland.

Hagen, M. (2008). *Algorithmic and Computational Complexity Issues of MONET*. Phd dissertation, Institut für Informatik, Friedrich-Schiller-Universität Jena.

Hamrouni, T., S. Ben Yahia, et E. M. Nguifo (2008). Succinct minimal generators : Theoretical foundations and applications. *Int. J. Found. Comput. Sci. 19*(2), 271–296.

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

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge discovery 1*(3), 241–258.

Summary

In this article, we introduced a new approach for the generation of minimal transversals, given a hypergraph. This approach extracts a lossless concise representation of the set of minimal transversals, using the notions of substitution class and representants. These notions, combined to data mining technics, allows us to provide a methodological framework to present a new algorithm the generation of minimal transversals IMT-EXTRACTOR. Carried out experiments on different datasets showed that the O-M2D algorithm provides very interesting performances compared to those obtained by classical algorithms.