Du nombre maximum d'ensembles fermés en 3 dimensions

Alexandre Bazin*, Laurent Beaudou**, Giacomo Kahn*** and Kaveh Khoshkhah****

*Laboratoire Electronique, Informatique et Image (Le2i),
Université de Bourgogne Franche-Comté
contact@alexandrebazin.com
**Laboratoire d'Informatique Fondamentale d'Orléans (LIFO),
Université d'Orléans
giacomo.kahn@univ-orleans.fr
*** Laboratoire d'Informatique, de Modélisation
et d'Optimisation des Systèmes (LIMOS),
Université Clermont Auvergne
laurent.beaudou@uca.fr
**** Institute of Computer Science, University of Tartu, TARTU, ESTONIA
khoshkhah@theory.cs.ut.ee

Résumé. Dans ce papier, nous étudions le nombre maximum d'ensembles fermés dans un cube de données de taille $n \times n \times n$. Nous montrons qu'il se situe entre 3.36^n et 3.38^n.

1 Introduction

L'étude des ensembles fermés, soient-ils fréquents ou non, est un des sujets centraux de la fouille de données. L'analyse formelle de concepts (FCA) (Ganter et Wille (1999)) est un des formalismes qui permettent d'étudier ces ensembles fermés, grâce à la structure qu'ils ont lorsqu'ils sont ordonnés par inclusion : le treillis des concepts. De nombreuses études utilisent ce formalisme, que ce soit dans des applications (Poelmans et al. (2013)) ou pour des résultats plus théoriques en complexité d'énumération (Gély et al. (2009)) ou de comptage (Kuznetsov et Obiedkov (2008)).

Il est bien connu que le nombre maximum d'ensembles fermés dans une table de données $n \times m$, avec n plus petit que m est 2^n. Dans ce papier, nous cherchons à généraliser au cas 3-dimensionnel la construction qui atteint 2^n, puis nous cherchons une borne supérieure au nombre maximum d'ensembles fermés en trois dimensions – nombre que nous appellerons par la suite $f_3(n)$. Pour ce faire, nous commençons par rappeler les définitions basiques en deux dimensions, puis en trois dimensions. La Section 3 donne une construction qui permet d'atteindre 3.36^n ensembles fermés. Dans la Section 4, nous donnons une esquisse de preuve pour une borne supérieure de 3.38^n. Les résultats présentés dans cet article sont disponibles en version longue sur ArXiv (Bazin et al. (2018)).

2 Définitions

Nous présentons notre travail dans le formalisme de l'analyse formelle de concepts (Ganter et Wille (1999)). En deux dimensions, un *contexte* est un triplet $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ dans lequel $\mathcal{O}$ et $\mathcal{A}$ sont des ensembles (appelés respectivement ensemble d'objets et ensemble d'attributs) et $\mathcal{R}$ est une relation entre $\mathcal{O}$ et $\mathcal{A}$. Moins formellement, on peut représenter un contexte par une table de croix, où une croix dans la case (o, a) signifie que $(o, a) \in \mathcal{R}$ et est lue "l'objet o possède l'attribut a".

Dans un contexte, il existe des motifs, appelés *concepts*, qui sont des paires (O, A) où O est un ensemble d'objets et A un ensemble d'attributs tels que $O \times A \in \mathcal{R}$ (toutes les croix entre les objets de O et les attributs de A sont présentes) et il n'est pas possible d'augmenter un de ces ensembles en conservant cette propriété. Les ensembles A et O correspondent aux ensembles fermés étudiés, notamment, en fouille de données. Un concept est un rectangle maximal de croix dans le contexte. L'ensemble des concepts d'un contexte, ordonnés par inclusion sur une de leur composante (traditionnellement l'ensemble d'objets), forme une structure ordonnée particulière : un treillis, que l'on appelle le treillis des concepts du contexte.

Un exemple de ces deux définitions est donné dans la Figure 1. L'ensemble d'objets $\mathcal{O}$ contient sept objets, l'ensemble d'attributs $\mathcal{A}$ contient cinq attributs. Dans ce contexte, la paire $(\{o_2, o_7\}, \{a_2, a_4, a_5\})$ est un concept. Afin d'alléger les notations, et quand cela n'induit aucune confusion, nous écrirons les ensembles sans leurs accolades. Ainsi, notre concept devient $(o_2o_7, a_2a_4a_5)$.

	a_1	a_2	a_3	a_4	a_5
o_1	×		×		×
o_2		×		×	×
o_3	×	×	×		
o_4			×	×	
o_5	×	×			×
o_6	×		×	×	
o_7	×	×		×	×

FIG. 1 – *Un exemple de contexte où* $\mathcal{O} = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$ *and* $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$. *Une croix dans la cellule* (o, a) *est lue "l'objet* o *a l'attribut* a*".*

Le passage en trois dimensions se fait naturellement (il a été fait pour la première fois dans (Lehmann et Wille (1995)). Un *3-contexte* est un quadruplet $(\mathcal{O}, \mathcal{A}, \mathcal{C}, \mathcal{R})$ où $\mathcal{O}$, $\mathcal{A}$ et $\mathcal{C}$ sont des ensembles, appelés respectivement ensemble d'objet, d'attributs et de conditions, et $\mathcal{R}$ est une relations ternaire entre ces ensembles.

Dans ce modèle, un *3-concept* est une boite maximale de croix dans le 3-contexte. Plus précisément, c'est un triplet (O, A, C) pour lequel on a que $O \times A \times C \in \mathcal{R}$ et on ne peut augmenter aucun de ces ensembles sans perdre la propriété. L'ensemble des 3-concepts suit une orientation différente de celle en deux dimensions, mais les concepts peuvent tout de même être ordonnés en un 3-treillis.

Les figures 2 et 3 donnent deux manière de visualiser un 3-contexte.

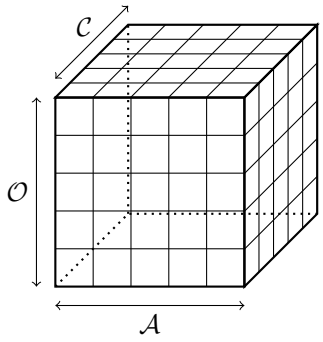

FIG. 2 – *Représentation visuelle d'un 3-contexte (sans les croix).*

	a	b	c	a	b	c	a	b	c
α	×	×		×			×		
β	×			×			×	×	
γ	×			×		×			×
		1			2			3	

FIG. 3 – *Un 3-contexte* $(Nombres, Grec, Latin, \mathcal{R})$ *avec* $Nombres = \{1, 2, 3\}$*,* $Grec = \{\alpha, \beta, \gamma\}$ *et* $Latin = \{a, b, c\}$*.*

3 2^n, 3^n et plus si affinité

En deux dimensions, dans un contexte de taille $n \times n$, il est possible d'atteindre 2^n concepts. Cela signifie que toutes les parties de l'ensemble $[n]$ sont fermées. Le contexte permettant d'atteindre 2^n concepts est appelé *contranominal scale* en anglais, nous l'appellerons contexte anti-diagonal ici, pour la raison suivante. Basé sur un ensemble S de taille n, il correspond au contexte $(S, S, \neq)$. Soit X un sous ensemble de S. Les concepts du contexte anti-diagonal sont de la forme $(X, \overline{X})$, avec $\overline{X} = S \setminus X$. Toutes les parties de S sont fermées, le treillis de concepts correspondant est le treillis booléen de dimension n. [1]

Un équivalent du contexte anti-diagonal en 3-dimension est le contexte $(S, S, S, S^3 \setminus \{(a, a, a) \mid a \in S\})$, montré en Figure 4.

	1	2	3	1	2	3	1	2	3
1		×	×	×	×	×	×	×	×
2	×	×	×	×		×	×	×	×
3	×	×	×	×	×	×	×	×	
		1			2			3	

FIG. 4 – *Le contexte anti-diagonal sur un ensemble de taille 3. Ce 3-contexte a* $3^3 = 27$ *3-concepts.*

1. Prenez garde ! Le terme de dimension est très trompeur ici, un treillis booléen de dimension n correspond à un contexte en dimension 2 !

Ce 3-contexte, basé sur un ensemble de taille n, donne 3^n 3-concepts. Introduit par Lehmann et Wille (1995), il a été également étudié par Biedermann (1998, 1999). La comparaison avec le cas 2-dimensionnel s'arrête ici, car on peut construire des contextes donnant plus de 3^n concepts.

Observation 1 *Il existe un 3-contexte* $5 \times 5 \times 5$ *avec quatre cent vingt-huit concepts. Ce contexte est donné en Figure 5.*

	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e
1		×	×	×	×	×	×	×	×		×	×	×		×	×	×		×	×	×		×	×	×
2	×		×	×	×		×	×	×	×	×	×	×	×		×	×	×		×	×	×		×	×
3	×	×		×	×	×		×	×	×		×	×	×	×	×	×	×	×		×	×	×		×
4	×	×	×		×	×	×		×	×	×		×	×	×		×	×	×	×	×	×	×	×	
5	×	×	×	×		×	×	×		×	×	×		×	×	×		×	×	×		×	×	×	×
	α					β					γ					δ					ϵ				

FIG. 5 – *Ce contexte a 428 concepts.*

On remarque que 428 est strictement supérieur à $3^5 = 243$. Notons également que $428 > 3.36^5$. Cela nous permet de conclure à l'existence de petits contextes ayant beaucoup de concepts, mais pas sur le cas général. Afin d'étendre cette observation, nous présentons maintenant une construction permettant de coller deux contextes et de multiplier leur nombre de concepts.

Soient $\mathbb{K}_1 = (\mathcal{O}_1, \mathcal{A}_1, \mathcal{C}_1, \mathcal{R}_1)$ et $\mathbb{K}_2 = (\mathcal{O}_2, \mathcal{A}_2, \mathcal{C}_2, \mathcal{R}_2)$ deux 3-contextes, tels que $\mathcal{O}_1 \cap \mathcal{O}_2 = \emptyset$, $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$ et $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. On construit le 3-contexte $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{C}, \mathcal{R})$ à partir de C_1 et C_2 en

1. fusionnant leurs dimensions : $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ et $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$;
2. gardant les croix existantes, et en en ajoutant de nouvelles dans les cellules empruntant des coordonnées aux deux contextes.

Deux exemples sont montrés dans la Figure 6 : en deux dimensions puis en trois dimensions.

Proposition 2 *Soient* $\mathbb{K}_1$ *et* $\mathbb{K}_2$ *deux 3-contextes avec respectivement* N_1 *et* N_2 *concepts. Alors le contexte* $\mathbb{K}$ *résultant de la fusion de* $\mathbb{K}_1$ *et* $\mathbb{K}_2$ *par la procédure décrite ci-dessus a* $N_1 \times N_2$ *concepts.*

Cela nous permet de répéter notre contexte $5 \times 5 \times 5$ pour créer de grands contextes avec beaucoup de concepts. Nous avons donc le théorème suivant :

Théorème 3 *Il existe une constante* c *telle que, pour tout entier* n, $f_3(n) \geq c3.36^n$.

4 Approche pour une borne supérieure

Notre approche pour obtenir une borne supérieure ne s'appuie pas directement sur les concepts d'un contexte, mais sur l'équivalence qu'ils ont avec les traverses minimales d'une certaine classe d'hypergraphes. Dans le cadre des 3-concepts, ils sont équivalents aux traverses minimale des hypergraphes 3-uniformes (chacune des arêtes a arité 3), 3-partis. Chaque arête correspond alors à un "trou" du contexte.

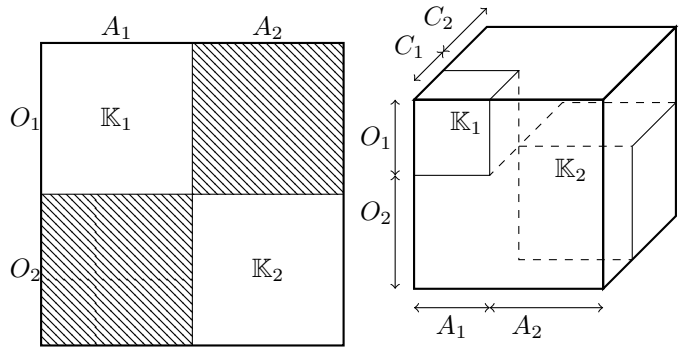

FIG. 6 – *La procédure revient à coller les deux contextes sur une diagonale, tout en remplissant le reste de croix. Les parties grisées en deux dimensions représentent ces croix. Elles sont absentes en trois dimensions, pour des raisons évidentes de lisibilité.*

En utilisant ce formalisme, nous avons utilisé une approche de *measure and conquer* (Kullmann (1999); Fomin et al. (2009)) pour borner le nombre de traverses minimales dans un hypergraphe de cette classe. Cette approche nous permet d'obtenir le théorème suivant :

Théorème 4 *Pour tout entier n, $f_3(n) \leq 3.38^n$.*

Preuve (version courte) Une version longue de cette preuve peut-être consultée dans (Bazin et al. (2018)).

Notre preuve utilise le theorème de Kullmann (1999) qui borne le nombre de feuilles d'un arbre muni de probabilités de transition sur ses arêtes. Nous commençons par montrer, pour un hypergraphe $\mathcal{H}$, qu'il est possible de construire un tel arbre de façon à ce que $\mathcal{H}$ soit la racine et les feuilles les traverses de $\mathcal{H}$. L'essentiel de la preuve se résume alors à identifier les probabilités de transition correspondantes aux différentes configurations possibles dans l'hypergraphe. Cela nécessite une étude de cas fastidieuse. Une fois les probabilités trouvées, le théorème nous assure que le nombre maximum de traverses minimales dans un hypergraphe tri-parti 3-uniforme à $3n$ sommets est inférieur à 1.5012^{3n} et donc que le contexte correspondant à une tripartition des sommets en trois dimensions égales possède, au plus, 3.38^n concepts.

5 Questions sans réponses

Bien que nous donnions un encadrement assez petit de la valeur de $f_3(n)$, nous n'avons pas de certitude quant à sa véritable valeur. Une recherche plus approfondie pourrait éventuellement permettre de trouver des 3-contextes ayant plus de concepts, qui pourraient alors être répliqués en utilisant la construction multiplicative.

De même, la borne supérieure que nous donnons peut possiblement être améliorée, soit en utilisant une évaluation plus fine en terme de *measure and conquer* ou par une tout autre approche.

La question reste ouverte par rapport au nombre maximum de concepts dans un contexte de dimension d (Voutsadakis (2002)).

Références

Bazin, A., L. Beaudou, G. Kahn, et K. Khoshkhah (2018). Bounding the number of minimal transversals in tripartite 3-uniform hypergraphs. *CoRR abs/1807.09030*.

Biedermann, K. (1998). Powerset trilattices. In *6th International Conference on Conceptual Structures, ICCS '98, Montpellier, France, 1998, Proceedings*, pp. 209–224.

Biedermann, K. (1999). An equational theory for trilattices. *Algebra Universalis 42*(4), 253–268.

Fomin, F. V., F. Grandoni, et D. Kratsch (2009). A measure & conquer approach for the analysis of exact algorithms. *J. ACM 56*(5), 25 :1–25 :32.

Ganter, B. et R. Wille (1999). *Formal concept analysis - mathematical foundations*. Springer.

Gély, A., L. Nourine, et B. Sadi (2009). Enumeration aspects of maximal cliques and bicliques. *Discrete Applied Mathematics 157*(7), 1447–1459.

Kullmann, O. (1999). New methods for 3-sat decision and worst-case analysis. *Theor. Comput. Sci. 223*(1-2), 1–72.

Kuznetsov, S. O. et S. A. Obiedkov (2008). Some decision and counting problems of the duquenne-guigues basis of implications. *Discrete Applied Mathematics 156*(11), 1994–2003.

Lehmann, F. et R. Wille (1995). A triadic approach to formal concept analysis. In *Third International Conference on Conceptual Structures, ICCS '95, Santa Cruz, USA, 1995, Proceedings*, pp. 32–43.

Poelmans, J., D. I. Ignatov, S. O. Kuznetsov, et G. Dedene (2013). Formal concept analysis in knowledge processing : A survey on applications. *Expert Syst. Appl. 40*(16), 6538–6560.

Voutsadakis, G. (2002). Polyadic concept analysis. *Order 19*(3), 295–304.

Summary

We study the maximum number of closed sets in a 3-dimensional dataset of size $n \times n \times n$. We show that it is between 3.36^n and 3.38^n.