Génération des chaînes maximales d'un treillis

Anne Berry[1], Jean-Paul Bordat[2], Eric SanJuan[3], Alain Sigayret[1]

[1] LIMOS, UMR CNRS 6158
bat. ISIMA, 63173 Aubière cedex, France.
berry@isima.fr, sigayret@isima.fr et www.isima.fr/limos/berry, www.isima.fr/limos/sigayret

[2] LIRMM
161 Rue Ada, 34392 Montpellier, France
bordat@lirmm.fr

[3] LITA (EA 3097), IUT de Metz
Ile du Saulcy, 57045 Metz cedex 1, France.
eric.sanjuan@iut.univ-metz.fr et www.iut.univ-metz.fr/ esanjuan

Abstract : Nous utilisons un nouveau codage d'une relation binaire par un graphe pour proposer un nouvel algorithme de génération de toutes les chaînes maximales du treillis associé.

Mots-clés : Data Mining, Treillis des Concepts, Treillis de Galois.

1 Introduction

Etant donné un contexte qui représente les réponses d'enquêtés à un questionnaire binaire, Guttman a proposé de tester si le contexte correspond à une relation de Ferrer, auquel cas il est possible de dégager un ordre de "difficulté" sur les questions ((Guttman, 1950), (Barbut et Monjardet, 1970)). Plus on progresse dans cet ordre, plus l'enquêté hésite à répondre par l'affirmative. Guttman a utilisé ce type d'approche pour établir une échelle de préjugés racistes sur différentes populations tels les soldats de l'armée américaine lors de la deuxième guerre mondiale.

Pour établir la proximité entre le contexte réel et une relation de Ferrer, Guttman a introduit la notion de "scalogram" qui consiste à réordonner les lignes et les colonnes du contexte par nombre d'items croissant. Ensuite il a introduit deux indices qui évaluent la proportion des erreurs dans ce tableau vis à vis d'une matrice binaire triangulaire qui n'aurait que des 1 sur la partie supérieure et que des 0 sur la partie inférieure.

Sur un questionnaire cependant beaucoup de colonnes et de lignes du contexte peuvent avoir le même nombre d'items, or leur disposition influe directement sur les indices de Guttman. Guttman recommandait de prendre la disposition la plus favorable, mais n'a proposé aucun moyen algorithmique pour y parvenir. Les logiciels SPSS et SAS qui ont implanté ce type d'analyse n'effectuent pas cette recherche d'un meilleur

ordonnancement des lignes et des colonnes, ce qui a conduit à des implantations très faiblement utilisées et finalement retirées des distributions standard après SPSS X.

Le progrès pour ce type d'analyse est venu de l'approche par les modèles d'erreur de Rasch qui a introduit la notion d'échelle de Guttman stochastique ((Linacre, 1992)). Ce modèle qui applique les outils statistiques de correction d'erreurs dans une transmission est parfaitement adapté à l'analyse de vastes échantillons, mais bien moins à celle de résultats de questionnaires d'opinion. Ainsi malgré la popularisation de logiciels mettant en oeuvre les modèles de Rasch tel que Winsteps, la question de la recherche combinatoire de la relation de Ferrer la plus proche d'un contexte de données reste d'actualité.

Or il a été démontré dans (Habib et al., 1991) que les sous-relations de Ferrer maximales d'un contexte correspondent aux chaînes maximales du treillis de Galois associé.

Nous nous posons donc ici le problème d'engendrer efficacement les chaînes maximales du treillis de Galois afin d'extraire les sous-relations de Ferrer maximales du contexte, ce qui permet de proposer à l'analyste autant d'échelles de Guttman possibles.

Dans le cadre de l'Analyse Formelle de Concepts, et plus particulièrement de la génération de Treillis de Galois/Treillis des Concepts, nous avons introduit récemment ((Berry at Sigayret, 1970), (Sigayret, 2002)) un codage par un graphe non-orienté de la relation binaire de départ, codage qui permet de mettre à profit les nombreux travaux existants en algorithmique de graphes. Plus précisement, nous montrons l'équivalence entre l'ensemble des séparateurs minimaux du graphe codant et l'ensemble des Rectangles Maximaux/Concepts du treillis associé à la relation.

Ce codage nous a permis de proposer des approches nouvelles à la généra-tion des éléments du treillis (voir (Berry, Bordat et Sigayret, 2003), (Sigayret, 2002)), ainsi que des outils généraux de représentation des fermetures (voir (Berry, Sanjuan et Sigayret, 2003)).

Notre propos ici est d'utiliser ces outils d'algorithmique de graphes pour proposer un algorithme de génération des chaînes maximales du treillis (c'est à dire des chemins de bottom à top), en un temps intéresssant par chaîne maximale, tout en se restreignant à un espace mémoire de taille polynomiale, évitant ainsi à l'utilisateur l'obligation de stocker toutes les chaînes maximales précédemment calculées pour pouvoir continuer à en engendrer d'autres.

2 Notations et rappels

Nous commençons par quelques définitions et exposons les résultats anté-rieurs sur lesquels notre discussion est basée.

Nous considérerons un ensemble fini $\mathcal{P}$ de "propriétés" et un ensemble fini $\mathcal{O}$ "d'objets", ainsi qu'une relation $R \subseteq \mathcal{P} \times \mathcal{O}$. Cette relation sera classiquement représentée par une table dont les croix dénotent l'appartenance d'un couple à R.

On appelle *rectangle maximal* (ou *concept*) de R tout élément $(A \times B) \subseteq R$ tel que : $\forall x \in X-A, \exists y \in B \,|\, (x,y) \notin R$ et $\forall y \in Y-B, \exists x \in A \,|\, (x,y) \notin R$. A est appelé *intension* du rectangle $A \times B$ et B *extension* de ce rectangle. Dans la suite, on

representera un élément du treillis par son intension, qui est un sous-ensemble de $\mathcal{P}$, même si dans les algorithmes on aura besoin des deux parties du rectangle.

Les rectangles maximaux de R sont structurés en treillis par l'inclusion sur les intensions, treillis que l'on notera $L(R)$ et que l'on représentera par son diagramme de Hasse, c'est à dire en omettant les arcs de réflexivité et de transitivité. Ce treillis associé à R s'appelle le *treillis de Galois* ou *treillis des concepts* de R, suivant l'école dont on se réclame.

Pour l'Analyse Formelle de Concepts, le triplet $(\mathcal{P}, \mathcal{O}, R)$ est appelé un *contexte formel* ou plus simplement *contexte*, et les rectangles maximaux de R, éléments du treillis, sont appelés des *concepts formels* ou plus simplement *concepts*.

En l'absence de notation standard, nous avons choisi d'utiliser dans nos exemples des lettres pour représenter les propriétés (éléments de $\mathcal{P}$) et des nombres pour représenter les objets (éléments de $\mathcal{O}$).

On appellera *bottom* l'élément minimum du treillis et *top* l'élément maximum. On appelle *chaîne maximale* un chemin de bottom à top dans le graphe de Hasse. On dira que l'élément $A' \times B'$ est un *successeur* de l'élément $A \times B$ si $A \subset A'$ et qu'il n'y a aucun élément intermédiaire $A'' \times B''$ tel que $A \subset A'' \subset A'$. L'ensemble des successeurs d'un élément forme sa *couverture*. Nous noterons $|R|$ le nombre d'éléments de R, et $|\overline{R}|$ le nombre $|\mathcal{P}|.|\mathcal{O}| - |R|$. On notera $R(X, Y)$, $X \subseteq \mathcal{P}, Y \subseteq \mathcal{O}$, la sous-relation dont on n'a gardé que les lignes de Y et les colonnes de X.

Exemple 2.1
$\mathcal{P} = \{a, b, c, d, e, f, g, h\}, \mathcal{O} = \{1, 2, 3, 4, 5, 6\}$.
Relation :

R	a	b	c	d	e	f	g	h
1		×	×	×	×			
2	×	×	×				×	×
3	×	×				×	×	×
4				×	×			
5			×	×				
6	×							×

Le treillis associé $\mathcal{L}(R)$ est donné par la figure 1.

Notre codage par un graphe est basé sur le complémentaire de la relation. Nous utiliserons les notations suivantes, pour une relation$R$:
Pour $x \in \mathcal{P}, N_R^+(x) = \{y | (x, y) \notin R\}$, $\overline{N_R}^+(x) = \{y | (x, y) \in R\}$.
Pour $x \in \mathcal{O}, N_R^+(x) = \{y | (y, x) \notin R\}$, $\overline{N_R}^+(x) = \{y | (y, x) \in R\}$.

Nous utilisons la notion de *domination*, qui provient directement de la théorie des graphes :

Définition 2.2
Soient $x, y \in \mathcal{P}$; on dit que x domine *y, noté $x \geq y$, ssi $N^+(y) \subseteq N^+(x)$.*

On montre (voir (Berry at Sigayret, 1970), (Sigayret, 2002)) que cette relation de domination définit un préordre sur les propriétés; nous appelons les classes d'équivalence définies par ce préordre des *maxmods* (ce terme provient de l'expression "module complet maximal", issu lui aussi de la théorie des graphes) :

Définition 2.3
On notera $x \sim y$ *ssi* $N^+(x) = N^+(y)$ *ssi* $\overline{N}^+(x) = \overline{N}^+(y)$. *On appelle* maxmod *un ensemble maximal de propriétés équivalentes.*

La relation de domination définit un ordre partiel sur ces maxmods, qui est très classiquement l'ordre quotient du préordre évoque ci-dessus.

Définition 2.4
On dit qu'un maxmod est non-dominant *s'il est un élément minimal de l'ordre de domination.*

Nos travaux nous permettent, pour un concept donné, de définir localement sa couverture, sans nécessiter d'information sur les autres éléments du treillis. Cette technique est la base de l'algorithme de Bordat ((Bordat, 1986)).

Définition 2.5
On appellera sous-relation de Bordat associée à un élément $A \times B$ *du treillis la relation* $R(\mathcal{P} - A, B)$.

Enfin, notre approche s'appuie sur le théorème fondamental suivant, ébauché dans (Bordat, 1986), et étendu à un opérateur de fermeture quelconque dans (Berry, Sanjuan et Sigayret, 2003) :

Théorème 2.6
((Berry at Sigayret, 1970), (Sigayret, 2002)) Soit $R \subseteq \mathcal{P} \times \mathcal{O}$, *soit* $A \times B$ *un élément du treillis associé, soit* X *un maxmod de la sous-relation de Bordat associée à* $A \times B$; *alors* $A \cup X$ *est l'intension d'un élément qui couvre* $A \times B$ *ssi* X *est un maxmod non dominant de cette sous-relation.*

3 Procédé algorithmique

Notre approche se base sur le fait que nous savons, grâce au graphe codant, calculer chaque élément de la couverture d'un concept donné en temps $O(|\overline{R}|)$.

Pour le problème de la génération des rectangles maximaux, nous avons introduit des structures de données assez sophistiquées, qui sont rendues necessaires par le fait qu'il est impératif de ne pas engendrer plus d'une fois un même concept.

Dans le contexte de la génération des chaînes maximales, il n'en va pas de même, puisqu'on est obligé d'engendrer chaque arc du graphe de Hasse du treillis.

Pour engendrer la couverture d'un élément, le principe que nous utilisons ici met en jeu un procédé d'affinement de partitions inspiré du célèbre algorithme LexBFS de Rose, Tarjan et Lueker ((Rose, Tarjan et Lueker, 1976)), ainsi que du travail subséquent de Hsu et Ma ((Hsu et Ma, 1999)), qui nous permet, en temps $O(|\overline{R'}|)$:

- de calculer la partition en maxmods d'une sous-relation R'.
- de fournir cette partition sous la forme d'une suite de maxmods qui est une extension linéaire de l'ordre de domination entre les maxmods, ce qui se concrétise par : si un maxmod Y domine un maxmod X, alors Y est après X dans la suite que l'on appellera une *partition ordonnée* des maxmods.

En conséquence, le premier élément de la suite est un maxmod non dominant. Par ailleurs, on sait, en temps $O(|\overline{R'}|)$, pour un maxmod X donné, calculer l'ensemble des maxmods qui dominent X. Ceci s'effectue par une simple recherche de sommets universels dans le sous-graphe codant défini par $N^+(X)$, ce qui nécessite au plus un parcours de ce graphe, qui a $O(|\overline{R'}|)$ arêtes significatives. Ainsi, on obtiendra l'ensemble des minimaux recherchés en répétant :

- choisir le maxmod X le plus à gauche et l'extraire de la suite;
- calculer les maxmods qui dominent X et les extraire également de la suite.

En ce qui concerne l'obtention des chaînes maximales, le principe sera le suivant : on part du bottom; on choisit un élément de la couverture de bottom; on recommence récursivement sur cet élément jusqu'à atteindre le top.

Exemple 3.1
Dans notre exemple, dans R, la partition de $\mathcal{P}$ en maxmods est : $\{a,h\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}$. *On peut l'obtenir de la manière suivante :*
$N^+(1) = \{a,f,g,h\}$, $\mathcal{P}$ *est partitionné en :* $\{b,c,d,e\}$ *suivi de* $\{a,f,g,h,\}$,
$N^+(2) = \{d,e,f\}$ *donne la suite :* $\{b,c\}, \{d,e\}, \{a,g,h\}, \{f\}$,
$N^+(3) = \{c,d,e\}$ *donne :* $\{b\}, \{c\}, \{d,e\}, \{a,g,h\}, \{f\}$,
$N^+(4) = \{a,b,c,f,g,h\}$, *pas de changement,*
$N^+(5) = \{a,b,e,f,g,h\}$, *donne :* $\{b\}, \{c\}, \{d\}, \{e\}, \{a,g,h\}, \{f\}$,
$N^+(6) = \{b,c,d,e,f,g\}$, *donne enfin :* $\{b\}, \{c\}, \{d\}, \{e\}, \{a,h\}, \{g\}, \{f\}$.
$\{b\}$ *est non dominant, il est dominé par* $\{f\}$ *et* $\{g\}$ *;* $\{c\}$ *est non dominant et n'est pas dominé ;* $\{d\}$ *est non dominant, il est dominé par* $\{e\}$ *; il reste* $\{a,h\}$ *dans la suite. Par conséquent, les maxmods non dominants sont :* $\{b\}, \{c\}, \{d\}$ *et* $\{a,h\}$, *ils constituent la couverture de bottom dans* $\mathcal{L}(R)$.

La section 4 donne les algorithmes détaillés.

4 Algorithmes

L'algorithme **POMM** construit une partition ordonnée des maxmods associés à un contexte. L'algorithme **MMND** extrait de cette partition les maxmods non dominants. Enfin, l'algorithme **CM** réalise le parcours en profondeur du treillis en utilisant en plus de la pile implicite des appels récursifs une pile permettant de stocker la chaîne maximale en cours de construction. L'algorithme **CM** est appelé initialement par **CM**($\emptyset$), la pile explicite étant initialisée avec l'intension de bottom.

Algorithme POMM
Donnée : un contexte $(\mathcal{P}, \mathcal{O}, R)$.
Résultat : une partition ordonnée $(K_1, ..., K_p)$ de $\mathcal{P}$ en maxmods de R.
PART $\leftarrow$ $(\mathcal{P})$;
pour $y \in \mathcal{O}$ FAIRE :
  pour chaque classe K de PART telle que $|K| > 1$ **faire :**
    $K_1 \leftarrow K - N_R^+(y)$;
    $K_2 \leftarrow K \cap N_R^+(y)$;
    si $K_1 \neq \emptyset$ et $K_2 \neq \emptyset$ **alors**
      Dans PART, remplacer K par K_1 suivi de K_2;
retourner PART.

Algorithme MMND
Donnée : un contexte $(\mathcal{P}, \mathcal{O}, R)$ et une partition ordonnée PART=$(K_1, ..., K_p)$ de $\mathcal{P}$ en maxmods de R.
Résultat : une partition PART réduite aux maxmods non dominants de $\mathcal{P}$.
MARQUÉ $\leftarrow \emptyset$;
pour i **de** 1 **à** p **faire :**
  Choisir x dans K_i;
  si $x \in$ MARQUÉ **alors** Supprimer K_i de PART
  sinon
    $D \leftarrow$ Ensemble des propriétés des maxmods qui dominent K_i dans R;
    pour $x \in D$ **faire :** MARQUÉ $\leftarrow$ MARQUÉ+$\{x\}$;
retourner PART.

Algorithme CM
Donnée : un concept $A \times B$ du contexte $(\mathcal{P}, \mathcal{O}, R)$.
Résultat : prolongation par les successeurs de $A \times B$ de la construction des chaînes maximales passant par cet élément du treillis et dont le début se trouve dans la pile.
Si $A = \mathcal{P}$
  alors AFFICHER les éléments de la pile depuis le fond jusqu'au sommet
  sinon
    PART $\leftarrow$ POMM($(\mathcal{P}-A, B, R(\mathcal{P}-A, B))$);
    PART $\leftarrow$ MMND($(\mathcal{P}-A, B, R(\mathcal{P}-A, B))$, PART);
    pour chaque X_i **dans** PART **faire :**
      EMPILER(X_i); *//pré-traitement.*
      CM($(A+X_i) \times (B-N^+(X_i))$); *//appel récursif.*
      DEPILER ; *//post-traitement.*

Complexité en temps

Pour chaque appel à **CM**, une exécution de **POMM** suivi de **MMND** fournit la liste des maxmods non dominants de la sous-relation considérée. Conformément à ce qui a été dit dans la section précédente, **POMM** peut être réalisé en $O(|\bar{R}|)$ et **MMND** en $O(|\bar{R}|)$ par maxmod non dominant obtenu. Chaque appel à **CM** se fait donc en $O(|\bar{R}|)$, coût de construction du concept passé en paramètre dans l'appel.

La construction d'une chaîne maximale à k éléments nécessite récursive-ment k appels à **CM** – de bottom à top – et donc un temps global en $O(k|\bar{R}|)$.

Il est à noter que cette analyse de complexité temporelle est probablement surévaluée.

Complexité en espace

La partition ordonnée PART et la pile explicite contiennent une suite d'ensembles disjoints de propriétés, leur taille ne dépasse donc pas $|\mathcal{P}|$.D'autre part, la pile de récursivité contient au plus $|\mathcal{P}|$ concepts dont chacun a une taille en $O(|\mathcal{P}|+|\mathcal{O}|)$.

La complexité spatiale est donc en $O(|\mathcal{P}| \times (|\mathcal{P}|+|\mathcal{O}|))$.

5 Conclusion

L'algorithme que nous proposons ici est générique, mais on peut imaginer facilement la mise en place des critères de sélection permettant de le guider, par exemple dans le cas, exposé en introduction, des échelles de Guttman.

Cette nouvelle utilisation du codage d'une relation et de son treillis par un graphe, codage que nous avons déjà appliqué dans différents contextes, nous confirme qu'il est important de continuer à proposer des passerelles inter-disciplinaires qui ne peuvent qu'enrichir le domaine en plein essor de l'Analyse Formelle de Concepts.

References

M. BARBUT AND B. MONJARDET. Ordre et classification. *Classiques Hachette*, 1970.

A. BERRY AND A. SIGAYRET. Representing a concept lattice by a graph. *Proceedings of Discrete Maths and Data Mining Workshop, 2nd SIAM Conference on Data Mining (SDM'02), Arlington (VA)*, April 2002, submitted to Discrete Applied Mathematics.

A. BERRY, J.-P. BORDAT AND A. SIGAYRET. Efficient Concept Generation. *To appear in proceedings of JIM 2003*.

A. BERRY, E. SANJUAN AND A. SIGAYRET. Generalized Domination in Closure Systems. *To appear in Proceedings of Discrete Maths and Data Mining Workshop, 3rd SIAM Conference on Data Mining (SDM'03), San Francisco*, May 2003.

J.-P. BORDAT. Calcul pratique du treillis de Galois d'une correspondance. *Mathématiques, Informatique et Sciences Humaines*, 96:31–47, 1986.

L. GUTTMAN. The basis for scalogram analysis. *In Stouffer et al. (1950). Measurement and prediction. The American Soldier Vol. IV. New York, Wiley.*

M.HABIB, M. MORVAN, M. POUZET, J-X. RAMPON. Extensions intervallaires minimales. *C.R.A.S. Paris, t 313, ser. I, (1991) p.893-898.*

W.-L. HSU AND T.-H. MA. Substitution decomposition on chordal graphs and its applications. *SIAM Journal on Computing*, 28:1004-1020, 1999.

J.-M. LINACRE. Stochastic Guttman order. *Rash Measurement Transactions, 1992, 5:4 p.189.*

D. J. ROSE, R. E. TARJAN, AND G. S. LUEKER. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 5:266–283, 1976.

A. SIGAYRET. *Data Mining : une approche par les graphes*. Thèse, LIMOS, Université Clermont-Ferrand II, 2002.

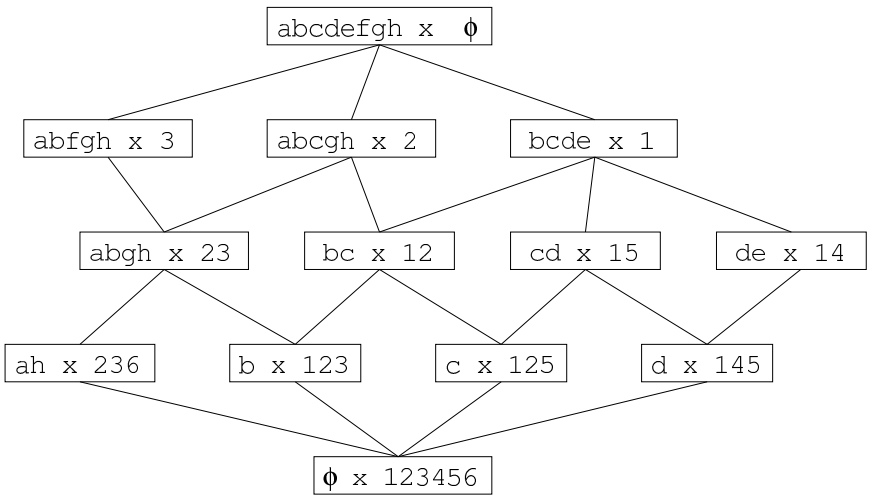

Figure 1: *Treillis $\mathcal{L}(R)$ associé à R.*