Extraction d'itemsets compacts

Bashar Saleh* Florent Masseglia*

*Inria Sophia-Antipolis Méditérranée
Equipe-Projet AxIS
2004 route des lucioles - BP 93
FR-06902 Sophia Antipolis
{Prénom.Nom}@sophia.inria.fr,
http://www-sop.inria.fr/axis

Résumé. L'extraction d'itemsets fréquents est un sujet majeur de l'ECD et son but est de découvrir des corrélations entre les enregistrements d'un ensemble de données. Cependant, le support est calculé en fonction de la taille de la base dans son intégralité. Dans cet article, nous montrons qu'il est possible de prendre en compte des périodes difficiles à déceler dans l'organisation des données et qui contiennent des itemsets fréquents sur ces périodes. Nous proposons ainsi la définition des itemsets compacts, qui représentent un comportement cohérent sur une période spécifique et nous présentons l'algorithme DEICO qui permet leur découverte.

1 Introduction

Le problème de la recherche de règles d'association, introduit dans Agrawal et al. (1993), est basé sur l'extraction de corrélations fréquentes entre les enregistrements et connaît de nombreuses applications dans le marketing, la gestion financière ou l'analyse décisionnelle (par exemple). Au cœur de ce problème, la découverte d'itemsets fréquents représente un domaine de recherche très étudié. Dans l'analyse du panier de la ménagère, par exemple, les itemsets fréquents ont pour but de découvrir des ensembles d'items qui correspondent à un nombre significatif de clients. Si ce nombre est supérieur à un support défini (par l'utilisateur) alors cet itemset est considéré comme fréquent. Cependant, dans la définition initiale des itemsets fréquents, l'extraction est effectuée sur la base de données toute entière (*i.e.* soit min_{supp}, le support minimum donné par l'utilisateur, les itemsets extraits doivent apparaître dans au moins $|D| \times min_{supp}$ enregistrements de D). Toutefois, il est possible que des itemsets intéressants reste ignorés malgré des caractéristiques particulières (y compris de support). Effectivement, les itemsets intéressants sont souvent liés au moment qui correspond à leur observation. On pourrait prendre pour exemple le comportement des utilisateurs d'un site de commerce en ligne pendant une offre spéciale sur les DVD et les CD vierges pour laquelle une publicité est faite par mailing. De la même manière, le site Web d'une conférence peut voir le nombre de connexions augmenter dans une fenêtre de quelques heures avant la date limite de soumission. Une condition nécessaire à la découverte de ce type de données est liée à l'aspect temporel des données. Cet aspect a déjà été abordé pour les règles d'association dans Ale et Rossi (2000);

Lee et al. (2001). Dans Ale et Rossi (2000), les auteurs proposent la notion de règles d'association temporelles. Leur idée consiste à extraire les itemsets qui sont fréquents sur des périodes définies par la durée de vie de chaque item (les périodes ne sont donc pas découvertes mais utilisées comme contrainte).

Dans cet article, nous proposons de découvrir les itemsets qui sont fréquents sur un sous-ensemble contigü de la base de données. Par exemple, les navigations sur les pages Web des DVD et CD vierges apparaissent de façon distribuée sur toute l'année. Cependant, la fréquence de ce comportement augmente très certainement pendant les quelques heures (ou jours) qui suivent le mailing. Ainsi, le défi consiste à trouver la fenêtre temporelle qui va optimiser le support de ce comportement. Considérons que le mailing soit envoyé le 3 mars et qu'il a influencé les clients pendant deux jours. Notre but est de découvrir que : "25% des clients, entre le 3 et le 5 mars, ont demandé la page des CD vierges, la page des DVD vierges et finalement la page des offres spéciales". Le support de ce comportement est sûrement trop faible pour permettre son extraction sur l'année toute entière mais cette connaissance (*i.e.* le comportement et la période sur laquelle il est fréquent) peut être très utile pour les décideurs qui veulent certainement découvrir ce comportement et sa période de fréquence pour finalement faire le lien avec le mailing.

2 Definitions

La définition 1 reprend le concept d'itemset fréquent de Agrawal et al. (1993). Nous y avons ajouté la notion d'estampille (donc une transaction peut couvrir plusieurs dates).

Définition 1 *Soit $\mathcal{I} = \{i_1, i_2, ..., i_n\}$ un ensemble d'items. Soit $X = \{i_1, i_2, ..., i_k\}/k \leq n$ et $\forall j \in [1..k]\ i_j \in \mathcal{I}$. X est un* **itemset** *(ou un $k-$***itemset***). Soit $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ un ensemble d'estampilles, sur lesquelles un ordre linéaire $<_T$ est défini, et où $t_i <_T t_j$ signifie que t_i précède t_j. Une* **transaction** *T est un couple $T = (tid, X)$ ou tid est l'identifiant de la transaction et X est l'itemset associé. À chaque item i de X est associé l'estampille t_i qui représente la date d'apparition de i dans T.*
Une transaction $T = (tid, I)$ supporte un itemset $X \in \mathcal{I}$ si $X \subseteq I$. Une **base de transactions** *D est un ensemble de transactions. La* **couverture** *d'un itemset X sur D est l'ensemble des identifiants de transactions dans D qui supportent X : couverture$(X, D) = \{tid/(tid, I) \in D, X \in I\}$. Le* **support** *d'un itemset X dans D est le nombre de transactions dans la couverture de X sur D : support$(X, D) = |couverture(X, D)|$. La* **fréquence** *d'un itemset X sur D est le rapport entre la taille de la couverture de X sur D et la taille de D : fréquence$(X, D) = \frac{support(X,D)}{|D|}$. Soit $\gamma \in]0..1]$ le support minimum donné par l'utilisateur, un itemset X est dit* **fréquent** *si fréquence$(X, D) \geq \gamma$.*

Définition 2 *L'ensemble F des itemsets fréquents de D avec un support minimum γ est noté $F(D, \gamma) = \{X \in \mathcal{I}/frequence(X, D) \geq \gamma\}$.*

Étant donnés un ensemble d'items $\mathcal{I}$, une base de transactions D et un support minimum γ, le problème de **l'extraction d'itemsets fréquents** vise à trouver $F(D, \gamma)$ ainsi que le support des itemsets de F. L'exemple 1 donne une illustration des concepts définis dans cette section.

Exemple 1 *La figure 1(a) montre un exemple de base de données D. Pour simplifier la lecture, nous supposons que les transactions de D sont affichées par ordre de date (i.e. T_1 est enregistrée avant T_2, etc.) et qu'une estampille unique est associée à tous les items d'une transaction (alors que dans la définition 1 chaque item est estampillé). Avec $\gamma = \frac{1}{2}$, les items fréquents (en gras dans les transactions de la figure 1(a)) sont a, b et c. Les itemsets fréquents de D sont (a), (b), (c), avec un support de $\frac{6}{10}$, et (a, c), avec un support de $\frac{1}{2}$.*

Tid	items	F(D,1/2)
1	**a b c**	
2	**a c** d	
3	**b** e f	
4	**c** j h	(a)
5	**a** i j	(b)
6	**b** k l	(c)
7	**a b c**	(a c)
8	m n o	
9	**a b c**	
10	**a b c**	

Fig. 1(a) Itemsets fréquents.

date	items	SI_1	SI_2	SI_3
1	**a b c**			
2	**a c** d			
3	**b** e f	(a)		
4	**c** j h			
5	**a** i j	(b)	(a c)	
6	**b** k l			
7	**a b c**	(c)		
8	m n o		(a b)	
9	**a b c**		(b c)	(a b c)
10	**a b c**			

Fig. 1(b) Itemsets compacts.

FIG. 1 – *itemsets fréquents et itemsets compacts sur D avec $\gamma = \frac{1}{2}$*

Notre problème est basé sur les estampilles et vise à extraire des itemsets qui sont fréquents sur des périodes particulières de D. Nous présentons maintenant les notions d'itemset temporel et d'itemset compact, qui sont au cœur de cet article.

Définition 3 *Une* **période** $P = (P_s, P_e)$ *est définie par une date de départ P_s et une date de fin P_e. L'ensemble des transactions qui appartiennent à une période P est défini par $Tr(P) = \{T/T \subseteq D, \forall i \in T, P_s \leq P_i \leq P_e\}$ avec P_i l'estampille de l'item i dans la transaction T. Enfin, PR est l'ensemble des périodes possibles sur D.*

En d'autres termes, l'ensemble des transactions qui appartiennent à P est l'ensemble des transactions dont tous les items sont estampillés dans les limites de P.

Définition 4 *Un* **itemset temporel** *x est un tuple (x_i, x_p, x_σ) où x_i est un itemset, x_p est une période associée à x_i et x_σ est le support de x_i sur x_p. Soit k la taille de x_i, alors x est un k-itemset temporel.*

Soit γ, le support minimum, nous présentons les caractéristiques d'un itemset compact dans la définition 5.

Définition 5 *Soit x un itemset temporel. x est un* **itemset compact (IC)** *ssi les conditions suivantes sont respectées :*
1) $x_\sigma \geq \gamma$
2) $\forall p_2 \in PR / x_p \subseteq p_2$ alors on observe a) ou b) ou les deux :
a) support$(x_i, p_2) < \gamma$
b) couverture(x_i, p_2) =couverture(x_i, x_p)
3) $\forall p_2 \in PR / p_2 \subseteq x_p$, couverture$(x_i, p_2)$ <couverture(x_i, x_p)
Soit k la taille de x_i, alors x est un **k-itemset compact**. *Enfin, $\mathbf{SI_k}$ est l'ensemble de tous les k-itemsets compacts.*

La première condition de la définition 5 assure que x représente un itemset qui est fréquent sur sa période. La seconde condition assure que la taille de x_p est maximale. En fait, si une période plus grande existe, alors, sur cette période, x_i n'est pas fréquent ou la couverture de x_i reste identique (*i.e.* étendre la période de x_p à p_2 n'apporte rien au support). Enfin, la troisième condition assure que la taille de x_p est également minimale. En effet, si x_i est supporté par la première et la dernière transaction de x_p, alors si il existe un période plus petite sur laquelle x_i est fréquent, la couverture sera plus faible (*i.e.* passer de x_p à p_2 implique d'ignorer des transactions qui supportent x_i et doivent donc être gardées). Une illustration de cette définition est donnée dans l'exemple 2.

Exemple 2 *La figure 1(b) montre les k-itemsets compacts qui sont extraits avec $\gamma = 0,5$. On peut y constater que les itemsets compacts de taille 1 sont (a), (b) et (c), que leur support est de $\frac{6}{10}$, et que leur période correspond à la base D entière. Ensuite, on peut observer trois itemsets compacts de taille 2 :*

- *$(a\ c)$, avec un support de $\frac{5}{10}$ et une période qui couvre toute la base D.*
- *$(a\ b)$ et $(b\ c)$, sur la période $[7..10]$ avec un support de $\frac{3}{4}$.*

Finalement, il y a un itemset compact de taille 3 : $(a\ b\ c)$ qui apparaît dans la période $[7..10]$ avec un support de $\frac{3}{4}$.

Définition 6 *L'ensemble des* **Itemsets Compacts Maximaux (ICM)** *est défini comme suit : soit x un IC, x est un ICM si les conditions suivantes sont respectées :*
$\forall y \in SI / x \neq y$ si $x_i \subseteq y_i$ alors $x_p \neq y_p$.

Dans la suite de ce papier, nous proposons un algorithme optimisé pour la découverte de l'ensemble des ICM, comme décrits par la définition 6.

3 DeICo : principe général

DEICO introduit un nouveau principe de comptage pour les itemsets candidats. Considérons un itemset temporel t qui n'est pas compact (*i.e.* $t_\sigma < \gamma$). Tout surensemble $u = (u_x, u_p, u_\sigma) / t_x \subseteq u_x \wedge u_p \subseteq t_p$ de t ne peut pas être un itemset compact (*i.e.* $u_\sigma < \gamma$). DEICO étend le principe d'apriori afin de générer des itemsets compacts candidats et compter leur support. Le principe de génération est modifié par l'ajout d'un filtre sur les intersections possibles entre candidats (*i.e.* si deux itemsets compacts de taille k ont un préfixe commun mais ne partagent pas la même période, alors leur croisement ne peut pas générer un itemset compact). Cependant, l'étape de comptage d'apriori ne peut pas s'appliquer directement dans notre cas. Considérons c, un itemset temporel candidat. Une solution consisterait à compter le nombre d'apparitions de c dans c_p. Ce n'est pas une solution correcte. Considérons en effet le candidat $c = ((a\ b), [1..10], c_\sigma)$ (généré à partir de $x = ((a), [1..10], \frac{6}{10})$ et $y = ((b), [1..10], \frac{6}{10})$). c n'est pas compact car $c_\sigma = \frac{4}{10}$. Toutefois, sur c_p, il existe un itemset compact $c' = ((a\ b), [7..10], \frac{3}{4})$. Notre but, pendant le comptage, est de construire des kernels qui correspondent aux périodes de fréquence des itemsets temporels candidats. Ensuite, ces kernels seront fusionnés dans le but d'obtenir les itemsets compacts. La définition 7 précise ces concepts. Cette définition récursive s'adapte bien au fait que l'on effectue des passes successives sur les données afin de trouver les périodes qui correspondent aux itemsets compacts.

En effet, la façon dont une passe est effectuée (soit dans l'ordre séquentiel des transactions) implique de découvrir les kernels "à la volée".

Définition 7 *Un* **kernel** *est une période. L'ensemble $K(x, P, \gamma)$ des kernels de l'item x sur la période P pour un support γ est défini comme suit :*
Soit $k \subseteq P$ une période telle que $x \subseteq Tr(k_s) \wedge Tr(k_s)$ est la première apparition de x sur P. Si k n'existe pas, alors $K = \emptyset$. Sinon, soit N l'ensemble des estampilles telles que $\forall n \in N, n \in P \wedge n > k_s \wedge frequence(x, [k_s..n]) < \gamma$ (en d'autres termes, N est l'ensemble des estampilles de P telles que toute extension de k à une estampille de N implique la perte de fréquence pour x). Si N est vide, alors k_e est défini comme la dernière apparition de x dans P et $K(x, P, \gamma) = \{k\}$. Sinon (i.e. $N \neq \emptyset$), soit $m \in N / \forall n \in N, n > m$ (m est la première estampille telle que la fréquence de x est perdue sur $[k_s..m]$). Alors k_e est défini comme la dernière apparition de x sur $[k_s..m]$ et $K(x, P, \gamma) = \{k\} \cup K(x, P - [k_s..k_e], \gamma)$.

Exemple 3 *Considérons l'itemset temporel candidat de taille 1, $c = ((b), [1..10], c_\sigma)$. La figure 2 donne la table booléenne des apparitions de b. Il y a deux kernels de (b) sur c_p ($[1..3]$ et $[6..10]$). Ces kernels peuvent être fusionnés pour obtenir un itemset compact $((b), [1..10], \frac{6}{10})$.*

date	b	kernels	merge
1	1	Kernel 1: [1..3] threshold=2/3	Itemset: (b) period: [1..10] threshold: 6/10
2	0		
3	1		
4	0		
5	0		
6	1	Kernel 2: [6..10] threshold=4/5	
7	1		
8	0		
9	1		
10	1		

FIG. 2 – *Kernels et période de l'itemset (b)*

Algorithm MERGEKERNELS
While $(\exists q, r \in K / \frac{couverture(x,q)|+|couverture(x,r)|}{|q \cup r|} \geq \gamma)$
$K \leftarrow K + q \cup r - q - r$;
$couverture(x, q \cup r) = couverture(x, q) \cup couverture(x, r)$
End while
End Algorithm MERGEKERNELS

Notre algorithme d'extraction se base sur le principe de la génération de candidats. Pendant la passe sur les données, le but de l'algorithme d'extraction est de mettre à jour les informations sur les kernels des itemsets temporels candidats dont la période englobe l'estampille de la transaction courante. A la fin de chaque passe de l'algorithme nous obtenons tous les kernels de chaque candidat pour cette passe. A la fin de chaque passe, les kernels obtenus pour chaque itemset temporel candidat sont fusionnés pour obtenir des itemsets compacts.

4 Expérimentations

Les expérimentations sont réalisées sur les fichiers log Web de l'Inria Sophia de Mars 2004 à Juin 2007 qui représentent 253 Go de données brutes et 36 710 616 navigations après le prétraitement. Voici un exemple de comportement trouvé grâce aux itemsets compacts.

1) Joan Miro : *start* : Thu Apr 20 07 :05 :39 2006 ; *end* : Thu Apr 20 17 :21 :06 2006 ; *frequence* : 0.024565 ; *cover* : 120 ; *itemset* : (préfixe "omega/personnel/Christophe.Berthelot/") - {css/style.css, Omega/JoanMiro/joanmiro.html}

Pour interpréter ce comportement, il faut savoir que : 1) cette page est dédiée à Joan Miro (un artiste célèbre) 2) Joan Miro est né le 20 avril 1893 et 3) la page de Christophe est classé cinquième dans les résultats Google avec les mots clés "Joan Miro" (au moment de ces expérimentations). Notre conclusion est donc que pour l'anniversaire de Miro (20 avril), les internautes ont massivement fait des recherches sur l'artiste et sont en partie passés par la page de Christophe. Ce comportement se trouve également en 2004, 2005 et 2007.

5 Conclusion

Nous avons proposé une nouvelle définition pour la découverte d'itemsets qui correspondent à des fréquences élevées sur des périodes précises sans connaissance préalable sur ces périodes. Cette découverte posait la difficulté de découvrir en même temps les itemsets et leurs périodes optimales de fréquence. De plus, le nombre de combinaisons possible devait être réduit et nous avons apporté les bases théoriques nécessaires à la résolution du problème. Notre algorithme, basé sur la découverte de 'kernels' et leurs fusions s'est révélé efficace et capable d'extraire ce nouveau type de connaissance de manière précise et exhaustive. D'après nos expérimentations, les itemsets compacts constituent un résultat lisible et instructif pour mieux comprendre les données étudiées.

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *SIGMOD*, Washington, D.C., USA, pp. 207–216.

Ale, J. M. et G. H. Rossi (2000). An approach to discovering temporal association rules. pp. 294–300.

Lee, C.-H., C.-R. Lin, et M.-S. Chen (2001). On mining general temporal association rules in a publication database. pp. 337–344.

Summary

Frequent pattern mining is a very important topic of knowledge discovery, intended to extract correlations between items recorded in databases. However, those databases are usually considered as a whole and hence, itemsets are extracted over the entire set of records. In this paper, we introduce the definition of solid itemsets, which represent a coherent and compact behavior over a specific period.