Université d'Angers

M2 Intelligence Décisionnelle

Recherche de motifs fréquents par algorithmes évolutionnaires

Auteur :
Ugo Rayer

Encadrants :
Benoit Da Mota
Béatrice Duval
David Lesaint

8 mars 2017

Remerciements

Avant toute chose, je tiens à remercier l'ensemble des personnes qui ont participé, de près ou de loin, à la réalisation de ce stage et à l'écriture de ce rapport.

Des remerciements spéciaux sont adressés d'une part au Laboratoire d'Etude et de Recherche en Informatique d'Angers pour son accueil et d'autre part au projet GRIOTE de la région Pays de la Loire qui a financé cette étude. Ensuite, je tiens à remercier chaleureusement Madame Duval et Messieurs Da Mota et Lesaint pour la qualité de leur encadrement et les différentes remarques qu'ils m'ont prodiguées pendant ces quelques mois.

Enfin, un grand merci à Josépha pour ses précieux conseils d'écriture malgré l'incompréhension générale des sujets abordés.

Table des matières

Chapitre 1

Introduction

Le calcul des motifs fréquents est une notion essentielle dans de nombreux domaines liés à la découverte et l'extraction de connaissances. Initiallement introduit par Agrawal et al. dans [1], ces motifs étaient alors utilisés pour l'établissement de règles d'associations visant à caractériser les habitudes d'achats de clients d'un supermarché. Par exemple, une règle de la forme "Pain & Beurre $\Rightarrow$ Jambon (75%)" signifie que 3 clients sur 4 achetant du pain et du beurre achètent également du jambon. Le calcul de telles règles se fait en deux étapes, dont la principale (i.e. disposant de la plus grosse complexité calculatoire) correspond au calcul des motifs fréquents. Depuis son introduction, le problème du calcul des motifs fréquents a été très largement étudié et appliqué à de nombreux domaines comme en bio-informatique, en cybersécurité et bien évidemment en marketing.

L'avènement de l'ère du Big Data a fait rentrer le problème de calcul des motifs fréquents dans une nouvelle dimension. En effet, le volume de données produites dans tous les domaines a cru de manière vertigineuse ces dernières années, rendant l'extraction de connaissances d'autant plus nécessaire et délicate. De fait, la problématique du passage à l'échelle des méthodes exactes est devenue primordiale. Bien que divers efforts en ce sens aient été faits au travers de nombreuses publications, ils se concentrent généralement sur l'optimisation et la parallélisation des méthodes existantes.

Les algorithmes évolutionnaires font partie des techniques de résolution de problèmes combinatoires appelées méta-heuristiques. Les méta-heuristiques regroupent un ensemble de méthodes approchées à la résolution de problème combinatoire. Elles sont naturellement envisagée lorsque la complexité du problème étudié ne permet pas l'usage de méthodes exactes (aussi bien en temps qu'en espace). Le principe des algorithmes évolutionnaires est de manipuler un ensemble d'individus représentant chacun une solution au problème étudié. A chaque itération, certains individus sont modifiés via des opérateurs de croisement et de mutation. Chaque individu est évalué au regard d'une fonction à optimiser dépendant du problème étudié.

Ainsi, nous formaliserons le problème de calcul de motifs fréquents dans la section suivante avant d'effectuer un état de l'art des méthodes existantes en

section 3. Le chapitre 4 sera dédié à la présentation de notre méthode dont nous présenterons les résultats en section 5. Le dernier chapitre sera consacré à la conclusion de cette étude et à une ouverture vers de futurs travaux.

Chapitre 2

Le problème du calcul motifs fréquents

2.1 Définition du problème

La définition la plus courante du problème de calcul des motifs fréquents se fait par la théorie des ensembles. Nous verrons cependant dans le chapitre suivant qu'il peut également être décrit par la théorie des graphes. Nous présenterons dans un premier temps un cadre formel nécessaire à la définition du problème que nous illustrerons ensuite. Enfin, nous introduirons quelques propriétés dérivés de la définition du problème.

2.1.1 Formalisation

Soit $\mathcal{I}$ un ensemble de *symboles* appelées **items**. Quelque soit $I \subseteq \mathcal{I}$, I est un *motif* appelé **itemset**.

Soit $\mathcal{T} = \{ t_1, \ldots, t_n \}$ un ensemble de **transactions**. Chaque élément t_i est un couple $< tid, I >$ où tid est l'identifiant de la transaction et $I \subseteq \mathcal{I}$. $\mathcal{T}$ est communément appelé **base de transactions**.

Pour tout itemset $I \subseteq \mathcal{I}$ la **couverture** de I par $\mathcal{T}$ est définie par :

$$\mathbf{cover}_{\mathcal{T}}(I) = \{ t \in \mathcal{T} |\ I \subseteq t \}$$

La cardinalité de la couverture d'un itemset I par $\mathcal{T}$

$$\mathbf{sup}_{\mathcal{T}}(I) = |\ \mathbf{cover}_{\mathcal{T}}(I)\ |$$

est appelée **support** de I. Etant donné un support minimal $minsup$ l'ensemble des **itemsets** (i.e. motifs) **fréquents** est défini par :

$$\mathbf{F}_{\mathcal{T},minsup} = \{ I \subseteq \mathcal{I} |\ sup_{\mathcal{T}}(I) \geq minsup \}$$

Le problème du **calcul des itemsets fréquents** (**FIM** - *Frequent Itemsets Mining*) est, étant donné une base de transactions $\mathcal{T}$ et un support minimal *minsup* de calculer l'ensemble F des itemsets fréquents. Comme F. Boden le remarque dans [10], bien qu'historiquement définie comme une valeur relative et donc asujettie à un support minimal défini dans l'intervalle [0,1], le support est de nos jours mesuré de manière absolue. Si nécessaire, nous y ferons référence sous la notion de fréquence :

$$\mathbf{Freq}_{\mathcal{T}}(I) = \frac{|\ \mathbf{cover}_{\mathcal{T}}(I)|}{|\mathcal{T}|}$$

avec *minfreq* simplement définie par $\frac{minsup}{|\mathcal{T}|}$.

2.1.2 Exemple

Situons nous dans le contexte de la définition historique de problème et considérons la base de transactions suivante (que nous conserverons tout au long de cet article) :

[2] [3] [4] [5] [6] [7] [8] [9] [10] [11]

Bibliographie

[1] R. Agrawal, T. Imielinski, and A.N. Swani. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–213. ACM Press, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB '94 : Proceedings of the 20th International Conference on Very Large Databases*, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.

[3] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26(2) of *SIGMOD Record*, pages 255–264. ACM Press, 1997.

[4] J. Pei, J. Han, and R. Mao. Closet : An efficient algorithm for mining frequent closed itemsets. In *Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD2000)*, pages 11–20, Dallas, TX, 2000.

[5] C. Borglet. Efficient implementations of a priori and eclat. In *Proceedings of the 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003)*, page 90, 2003.

[6] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang. H-mine : Hyper-structure mining of frequent patterns in large databases. In *Proc. 2001 Int. Conf. Data Mining (ICDM2001)*, pages 441–448, San Jose, CA, 2001.

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *roc. 7th Int. Conf. Database Theory (ICDT'99)*, pages 398–416, Jerusalem, Israel, January 1999.

[8] M.J. Zaki and C. Hsiao. Charm : An efficient algorithm for closed association rule mining. Technical report, Rensselaer Polytechnic Institute, 1999.

[9] J. Han, P. Jian, Y. Yiwen, and M. Runying. *Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach*, volume 8, pages 53–87. 2004.

[10] F. Bodon. A survey on frequent itemset mining. 2006.

[11] B. Goethals. Survey on frequent pattern mining.