Extraction de séquences fréquentes avec intervalles d'incertitude

Asma Ben Zakour*,**, Sofian Maabout*, Mohamed Mosbah*, Marc Sistiaga**

*LaBRI, Université Bordeaux 1, CNRS, FRANCE
maabout@labri.fr, mosbah@labri.fr
**2MoRO Solutions, Bidart, FRANCE
marc.sistiaga@2moro.fr, asma.ben-zakour@2moro.fr

Résumé. Lors de l'extraction des séquences, la granularité temporelle est plus ou moins importante selon les besoins des utilisateurs et les contraintes du domaine d'application. Nous proposons un algorithme d'extraction de séquences fréquentes par intervalles à partir de séquences à estampilles temporelles discrètes. Nous intégrons une relaxation des contraintes temporelles en introduisant la définition de "séquences temporelles par intervalles" (STI). Ces intervalles reflètent une incertitude sur les occurrences précises des évènements. Nous formalisons ce nouveau concept en exhibant certaines de ses propriétés et nous menons quelques expériences afin de comparer (qualitativement) nos résultats avec une autre proposition assez proche de la nôtre.

1 Introduction

L'extraction de séquences fréquentes (ESF) a été introduite par Agrawal et Srikant (1995) comme une extension de leur algorithme *A Priori* Agrawal et Srikant (1994) qui calcule les ensembles fréquents. Lors de l'ESF, la chronologie d'occurrence des événements est plus ou moins importante selon la nature des connaissances à extraire. En effet, il est parfois nécessaire de relaxer les contraintes de chronologie afin d'extraire des informations utiles. Dans cet article, nous proposons de *fusionner* des évènements consécutifs et proches temporellement en un ensemble d'évènements simultanés dont l'estampille temporelle n'est plus discrète mais exprimée par un intervalle temporel. Cet intervalle représente une incertitude sur les instants exacts des occurrences des évènements regroupés. Ce regroupement est contraint par la taille de la fenêtre glissante fixée par l'utilisateur. Ce travail est réalisé dans le cadre d'un projet industriel. Il s'agit d'anticiper des opérations de maintenance des équipements aéronautiques. Par exemple, considérons un historique d'utilisation d'avions. V_i désigne le vol i et M_j désigne l'opération de maintenance j. Soit $\mathcal{S} = \{S_1, S_2\}$ un ensemble de séquences avec : $S_1 = \langle(0, V_1)(1, V_2)(2, V_3)(5, M_1)\rangle$ et $S_2 = \langle(0, V_1)(1, V_3)(2, V_2)(6, M_1)\rangle$. En fixant un support minimal à 2 et une taille de fenêtre glissante égale à 1, notre approche retourne la séquence : $\langle([0, 0]V_1)([1, 2]V_2\ V_3)([5, 6]M_1)\rangle$ qui traduit les faits suivants : « Lorsque le vol V_1 est effectué, V_2 et V_3 ont lieu *dans n'importe quel ordre* entre une et deux unités de temps après V_1, *i.e.*, dans l'intervalle $[1, 2]$. Par la suite, la réparation M_1 est effectuée dans l'inter-