FIASCO : un nouvel algorithme d'extraction d'itemsets fréquents dans les *flots de données*

Lionel VINCESLAS*, Jean-Émile SYMPHOR*, Alban MANCHERON* et Pascal PONCELET**

*GRIMAAG, Université des Antilles et de la Guyane, Martinique, France.
{lionel.vinceslas,je.symphor,alban.mancheron}@martinique.univ-ag.fr,

**EMA-LG2IP/site EERIE, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France.
pascal.poncelet@ema.fr

Résumé. Nous présentons dans cet article un nouvel algorithme permettant la construction et la mise à jour incrémentale du FIA$_\theta$[1] : **FIASCO**. Notre algorithme effectue un seul passage sur les données et permet de prendre en compte les nouveaux batches, itemset par itemset et pour chaque itemset, item par item.

1 Introduction

Le FIA$_\theta$ est un nouvel automate qui permet de traiter de façon efficace la problématique de l'extraction des itemsets fréquents dans les flots de données. **FIASCO** est l'algorithme qui permet de construire et de mettre à jour le FIA$_\theta$ en effectuant un seul passage sur les données. Notre objectif dans cet article est de présenter et d'illustrer par l'expérimentation l'applicabilité et le passage à l'échelle de **FIASCO** dans le cas des flots de données.

2 FIASCO (Frequent Itemset Automaton Stepwise Construction Operator)

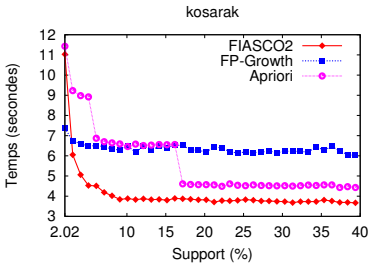
Le FIA$_\theta$ est un automate déterministe et acyclique, ce qui nous permet d'établir une relation d'ordre sur ses états (notée $\preccurlyeq$). De par cette relation d'ordre, nous introduisons un algorithme en deux passes pour la construction de cet automate, en utilisant des *bits positions* : **FIASCO**2. Cet algorithme utilise les propriétés d'Apriori afin d'optimiser sa construction, ce qui le rend efficace dans le cas d'une base de données (*cf.* section 3). Nous proposons aussi un algorithme en une passe (**FIASCO**1), pour les flots de données, permettant de mettre à jour incrémentalement le FIA$_\theta$, item par item, avec une phase d'élagage en utilisant un support statistique.

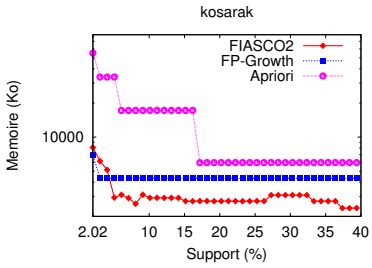
3 Expérimentations

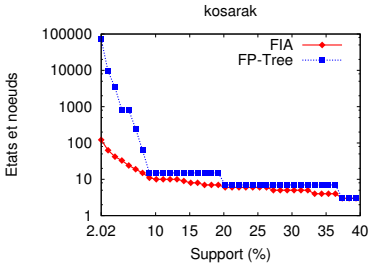
Les expérimentations ont été réalisées sur les jeux de données[2] kosarak et T10I4D100K, sur une machine munie d'un bi-processeur AMD ATHLON 3600+ 64 bits, avec 1Go de RAM.

[1] Le FIA$_\theta$ est présenté comme article long à EGC'08

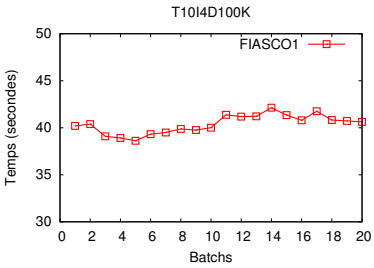
[2] disponibles à l'URL `http://fimi.cs.helsinki.fi/data`

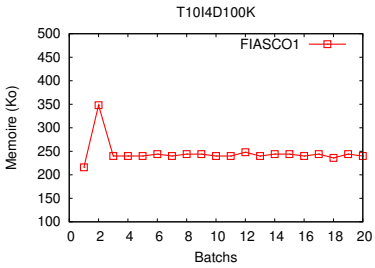






« Les résultats confirment bien l'applicabilité du FIA$_\theta$ *dans les flots de données (cf. courbes du jeu de données T10I4D100K). Les résultats obtenus sont comparables voire meilleurs pour certaines valeurs de support que Apriori et FP-Growth, sachant que le* FIA$_\theta$ *est une structure qui indexe les itemsets fréquents du flot de données. »*





4 Conclusion

Nous présentons dans cet article un nouvel algorithme, **FIASCO**, qui permet de construire et de mettre à jour incrémentalement le FIA$_\theta$ appliqué aux flots de données. Cet algorithme est en une passe, avec une granularité par item. Les expérimentations, avec une analyse en temps et en espace, montrent l'applicabilité et le passage à l'échelle de l'algorithme.

Summary

We present in this paper a new algorithm for constructing and incrementally updating the FIA$_\theta$: **FIASCO**. Our algorithm only needs one scan over the data and takes into account the new batches, itemset per itemset and for each itemset, item per item.