Découverte de motifs compressés dans des flux de données

Contexte - La fouille d'itemsets fermes dans les flux de données [2, 3, 6, 7] (stream) vise à maintenir en temps-réel une représentation compressée sans pertes d'un ensemble de fragments représentatifs, utiles et potentiellement actionnables à partir d'un flux de transactions.

Problématique - Les flux de données [9] sont mis à jour en continu et imprévisibles. Il n'est pas envisageable de repasser le flux avec différents paramètres (ex: support). Contrairement aux approches de type batch, une méthode de fouille de flux doit traiter les transactions selon leur ordre d'arrivée. De plus, dans un mode fenêtre glissante, elle doit supprimer les transactions trop anciennes ou obsolètes. Dès lors, maintenir les motifs dans le flux pose deux problèmes: le temps de mise à jour de chaque modification et les coûts en mémoire nécessaires au maintien des motifs. Une approche efficace doit impérativement éviter la sur-utilisation des ressources (CPU, mémoire) menant à une dégradation des performances.

Deux nouvelles méthodes [6, 7] ont été développées récemment en exploitant le cadre mathématique de l'analyse des concepts [6, 7]. L'optique ici est double: D'abord, d'enrichir l'approche en appliquant les récents développements à des situations résolues en mode *par lot*, mais toujours pas abordées en flux. De commun accord, une direction sera choisie parmi : les règles rares [4], règles minimales [5], top-k règles [8], motifs hautement utiles [9], séquences [1], etc.

Objectifs du stage - Concevoir, implémenter et étudier expérimentalement des méthodes innovant permettant de maintenir depuis un flux les familles d'itemsets qui sous-tendent les problèmes visées (top-k motifs d'un certain type, les bases de règles d'association valides, les motifs les plus utiles, etc.)

Approche générale - Il s'agit de porter de problématiques déjà connues en batch dans un mode flux. Ainsi, suite à des études du comportement des structures mathématiques impliquées, des nouvelles méthodes seront conçues, implémentées et validées de manière expérimentale. Le but sera d'identifier leurs forces et leurs limites. Un soin particulier sera apporte au coût algorithmique (complexité au pire cas) ainsi qu'aux performances pratiques des implémentations.

Stratégie de réalisation - Ce travail reprend des travaux effectues au CRIA par plusieurs étudiants gradués (dont un stagiaire de l'École Polytechnique de Paris). Des formalisations partielles et des implémentations pour de nombreux primitives de fouille sont déjà disponibles. Les travaux seront accompagnés par un doctorant du CRIA, T. Martin, qui offrira un support de pointe en optimisation logicielle.

Conditions de réalisation - Le stage demande une double compétence en mathématiques discrètes et en programmation. Celle-ci sera faite en C/C++ et/ou Java. Des connaissances de base en fouille de motifs (combinatoire) ainsi qu'en optimisation des performances de code C/C++ sont souhaitables. La durée du stage est de 5 à 6 mois.

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the 1995 International Conference on Data Engineering (ICDE '95)*, pages 3–14, Mar 1995.

[2] Y. Chi et al. Moment: Maintaining closed frequent itemsets over a stream sliding window. In *ICDM'04*, pages 59–66, 2004.

[3] C. Gao and J. Wang. Efficient itemset generator discovery over a stream sliding window. In *18th CIKM*, pages 355–364, 2009.

[4] Y. Koh and S. Ravana. Unsupervised rare pattern mining: a survey. *ACM TKDD*, 10(4):45, 2016.

[5] M. Kryszkiewicz. Concise Representations of Association Rules. In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, pages 92–109, 2002.

[6] T. Martin, G. Francoeur, and P. Valtchev. Ciclad: A fast and memory-efficient closed itemset miner for streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1810–1818, 2020.

[7] T. Martin, L.-R. Roux, and P. Valtchev. Fgc-stream: A novel joint miner for frequent generators and closed itemsets in data streams. In *Proceedings of the 21st IEEE International Conference on Data Mining*, 2021.

[8] L. T. Nguyen, B. Vo, L. T. Nguyen, P. Fournier-Viger, and A. Selamat. Etarm: an efficient top-k association rule mining algorithm. *Applied Intelligence*, 48(5):1148–1160, 2018.

[9] H. Ryang and U. Yun. High utility pattern mining over data streams with sliding window technique. *Expert Systems with Applications*, 57:214–231, 2016.