PROPOSITION ACTION POUR LA FEDERATION ICVL

1- Titre

Hybridation des techniques de clustering et d'extraction de motifs

2- Description succincte

L'objectif de cette action est une fertilisation croisée entre les techniques de clustering et celles d'extraction de motifs. Ces deux thématiques fortes en fouille de données se sont largement influencées ces dernières années dans la littérature. Au sein de la fédération ICVL, nous souhaitons mettre en place une collaboration nouvelle autour de l'hybridation de ces méthodes aussi bien de l'extraction de motifs vers le clustering que du clustering vers l'extraction de motifs pour conduire à de nouveaux algorithmes pour la fouille de données.

3- Porteurs de l'action et participants

- Sylvie Billot (CA - LIFO)
- **Guillaume Cleuziou (CA – LIFO) - porteur**
- Thi-Bich-Hanh Dao (CA – LIFO)
- Matthieu Exbrayat (CA - LIFO)
- Arnaud Giacometti (BdTln – LI)
- **Nicolas Labroche (BdTln – LI) - porteur**
- **Arnaud Soulet (BdTln – LI) - porteur**
- Marcilio de Souto (CA – LIFO)
- Christel Vrain (CA - LIFO)

4- Historique des collaborations

Cette action s'inscrit dans une volonté commune d'initier une nouvelle thématique de recherche reposant sur les compétences du LI et du LIFO en clustering et recherche de motifs.

5- Proposition de travail

Les travaux portant sur l'interaction entre le clustering et l'extraction de motifs sont déjà présents au sein de la littérature [1]. Nous proposons de les étendre selon deux directions complémentaires en observant tout d'abord ce que l'extraction de motifs peut apporter au clustering et ensuite ce que le clustering peut apporter à l'extraction de motifs.

Extraction de motifs pour le clustering

Cette approche consiste soit à réutiliser les techniques algorithmiques pour extraire des clustering dans des sous-espaces distincts (subspace clustering [1]), soit à exploiter directement les motifs extraits pour construire un clustering (clustering conceptuel [2]).

Lorsque le nombre de dimensions décrivant les données devient très grand la qualité du clustering a souvent tendance à se dégrader. Dans ce contexte, l'idée clé du subspace clustering est de découvrir des clustering alternatifs en considérant des sous-espaces alternatifs. La recherche de ces sous-

espaces alternatifs repose sur les techniques d'énumération de motifs correspondants aux dimensions de l'espace des données. Un autre apport des méthodes d'extraction de motifs pour le clustering est la combinaison de motifs pour construire une partition (parfois, avec recouvrement) des données. Ces approches de clustering conceptuel (aussi appelées, bi-clustering ou co-clustering) ont l'avantage de fournir une description en intension des clusters. Nos travaux dans ce cadre porteront sur les aspects sémantiques, visant une description efficace des clusters à partir des sous-espaces, mais également sur des aspects liés à la pertinence des méthodes de subspace clustering : comment évaluer l'intérêt d'un sous-espace et comment déterminer le nombre idéal de dimensions pour chaque cluster.

Nous explorerons enfin dans cette action une troisième voie autour du clustering muti-vues et multi-objectifs hybride : dans cette problématique nous considérerons un même jeu de données selon les deux aspects (vues ou objectifs) numériques et conceptuels dans le but de faire émerger de nouvelles solutions (voire des ensembles de solutions) réalisant un consensus sur les objectifs à la fois numériques (inertie ou vraisemblance) et conceptuels (définition intensionnelle et motif clos).

Enfin, dans le contexte actuel des grandes masses de données, les méthodes proposées devront être efficaces et pour le moins s'affranchir des problèmes classiques d'explosion combinatoire en recherche de motifs liés à l'exploration exhaustive des sous-ensembles de dimensions.

Clustering pour l'extraction de motifs

A l'inverse, il est envisageable de tirer profit des approches de clustering pour améliorer les méthodes d'extraction de motifs. Cette voie a été moins explorée dans la littérature et souvent, le clustering est utilisé en simple post-traitement de l'extraction de motifs. Typiquement, les techniques de clustering sont utilisées pour regrouper les motifs similaires afin de réduire leurs redondances intrinsèques à l'approche. Cependant, nous pensons que le clustering pourrait être intégré plus en amont. Certaines propriétés du clustering devraient être intégrées dans le processus d'extraction de motifs pour en améliorer son fonctionnement. Par exemple, les méthodes d'extraction de motifs peu efficaces sur les données numériques pourraient exploiter une métrique à l'instar des méthodes de clustering.

Objectifs

Les thématiques de l'extraction de motifs et du clustering, présentes au sein de l'ICVL, sont propices à une collaboration nouvelle et immédiate. Plus particulièrement, nous souhaitons avancer sur les points suivants :

- Intégration d'une vue « concetpuelle » dans un processus de clustering multi-vues afin de faire émerger de nouvelles solutions consensuelles
- Intégration de la notion de métrique dans le support afin d'extraire des motifs dans des données numériques
- Amélioration de la pertinence et de la sémantique des approches de subspace clustering
- Etude des interactions entre le clustering et d'autres formes de motifs

6- Prospectives

- Axe « Masses de données » ou « données et connaissances »

7- Références

[1] Arthur Zimek, Ira Assent and Jilles Vreeken (2014) Frequent Pattern Mining Algorithms for

Data Clustering, C. C. Aggarwal, J. Han (eds.), Frequent Pattern Mining, Springer International

[2] Douglas Fisher, Data mining tasks and methods: Clustering: conceptual clustering, Handbook of data mining and knowledge discovery, Oxford University Press, Inc., New York, NY, 2002