Intégration de préférences utilisateurs pour la fouille de données selon l'intérêt subjectif de l'utilisateur - 2021

Laboratoire d'accueil : IRISA (Rennes) - CNRS UMR 6074

Encadrement :

- Bouadi Tassadit, Cellier Peggy, et Termier Alexandre (prenom.nom@irisa.fr) IRISA-INRIA Rennes (LACODAM - SEMLIS)
- Crémilleux Bruno (bruno.cremilleux@unicaen.fr) GREYC - Université de Caen Normandie

Mots-clés : fouille de données interactive, mesures d'intérêt, préférences utilisateur

Contexte :

Les méthodes de fouille de données ont pour objectif d'assister l'utilisateur dans sa compréhension des données, en découvrant des modèles intuitifs, utiles, inattendus mais surtout intéressants pour ce dernier. L'importance de l'utilisateur dans ce processus d'exploration paraît donc évidente. Extraire des modèles présentant un intérêt pour l'utilisateur implique la prise en compte de ses attentes. Dans ce cadre, la notion de *préférence* a été définie : une préférence reflète une "*opinion*" d'un utilisateur sur un objet d'intérêt, ici un modèle (ex. une règle d'association, un itemset, un cluster, etc.).

Les préférences expriment des comparaisons sur un ensemble d'éléments, de choix ou dans notre cas des modèles. Les préférences peuvent être exprimées sous différentes formes : de manière quantitative, en indiquant des notes/scores grâce à des mesures d'intérêt (ex. mesures statistiques sur des règles [2], l'intérêt subjectif de l'utilisateur [3]), ou de manière qualitative [4], par des comparaisons par paires ou d'autres formalismes de préférences plus sophistiqués (par exemple, "Je préfère la règle d'association A à la règle d'association B"). L'approche qualitative est plus générale que l'approche quantitative. Les préférences apparaissent donc comme un moyen naturel pour classer les modèles. Pourtant la notion de préférence utilisateur est encore peu utilisée en fouille de données, alors qu'elle permet à l'analyste d'exprimer des requêtes de façon plus riche que les traditionnelles mesures d'intérêt [1]. Ce manque d'utilisation s'explique certainement par la difficulté pour l'utilisateur à expliciter la relation de préférence permettant de formaliser ses attentes. En effet, plus un modèle de préférence est expressif, plus il est difficile à représenter/acquérir et à intégrer dans le processus de fouille.

Travail à réaliser :

Dans ce stage, nous nous intéresserons plus particulièrement à l'intérêt subjectif de l'utilisateur [4,5]. La fouille de données selon l'intérêt subjectif de l'utilisateur consiste à rechercher des modèles surprenants par rapport à une connaissance du domaine telle qu'un a priori sur les données. Autrement dit, on maximise la préférence de l'aspect inattendu de l'information découverte par rapport à la connaissance du domaine.

Nous pensons intéressant et prometteur de combiner des préférences qualitatives (ex. relation d'ordre Pareto, préférences conditionnelles ou floues, etc.) à l'intérêt subjectif de l'utilisateur dans le processus d'explicitation de préférences. Les motifs obtenus auront ainsi un intérêt à la fois selon la connaissance du domaine (ici issue des données) et des préférences de l'utilisateur (*i.e.* extraction de modèles divers et représentatifs, en plus d'avoir un intérêt pour l'utilisateur).

Pour ce faire, un certain nombre de verrous sont à lever : (i) choix de la mesure d'intérêt subjectif ; (ii) choix du modèle de préférence qui va déterminer la représentation de l'utilisateur (compromis entre expressivité et complexité du modèle) ; (iii) intégration des préférences qualitatives dans le calcul de la mesure d'intérêt subjectif.

Le but ultime de ce stage est de construire un système capable d'expliciter les préférences de l'utilisateur à partir de motifs découverts dans les données selon un scénario de fouille interactive [6, 7] alternant phases de fouille de données et phases d'apprentissage. A partir d'une requête initiale de l'utilisateur, le système présente un premier ensemble de motifs : (1) l'utilisateur sélectionne certains de ces motifs, les désignant comme réellement intéressants pour lui ; (2) le système considère ces motifs comme des exemples de préférences de l'utilisateur et apprend alors ses préférences ; (3) une nouvelle collection de motifs est extraite en utilisant ces préférences mises à jour, celle-ci est présentée à l'utilisateur, et retour à l'étape (1).

Références :

[1] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38(3):9, 2006.

[2] Wilhelmiina Hämäläinen and Matti Nykänen. Efficient discovery of statistically significant association rules. In Proceedings of the 8th IEEE Int. Conf. on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy, pages 203–212, 2008.

[3] Tijl De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases, DAMI, 2011.

[4] Öztürké, Meltem, Alexis Tsoukiàs, and Philippe Vincke. "Preference modelling." Multiple criteria decision analysis: State of the art surveys. Springer, New York, NY, 2005. 27-59.

[5] Puolamäki, Kai, Oikarinen, Emilia, Kang, Bo, et al. Interactive visual data exploration with subjective feedback: an information-theoretic approach. Data Mining and Knowledge Discovery, 2020, vol. 34, no 1, p. 21-49.

[6] B. Crémilleux, M. Plantevit, and A. Soulet. Preference-based pattern mining. In 14th International Conference on Formal Concept Analysis, Rennes, France, 2017.
https://perso.liris.cnrs.fr/marc.plantevit/doku/doku.php?id=preferencebasedpatternminingtutorial#material

[7] M. Van Leeuwen. Interactive data exploration using pattern mining. In Interactive knowledge discovery and data mining in biomedical informatics, pages 169–182. Springer, 2014.

Période :

Stage de 5 à 6 mois à effectuer entre le 1er février et le 31 août 2021.

Gratification :

Selon règles en vigueur (environ 560 euros par mois.

Profil souhaité :

Etudiante ou étudiant en master informatique ou école d'ingénieur en informatique. De solides compétences en fouille de données et programmation ainsi qu'une ouverture sur les statistiques seront hautement appréciées.

French and/or English are the working languages.

Les candidates et candidats sont encouragés à nous contacter dès que possible.

Pour candidater :

Pour candidater, envoyer les documents suivants (exclusivement au format pdf) à tassadit.bouadi@irisa.fr, alexandre.termier@irisa.fr, peggy.cellier@irisa.fr et bruno.cremilleux@unicaen.fr :

- curriculum vitae ;
- lettre de motivation expliquant vos qualifications, expériences et motivation pour ce sujet ;
- relevé de notes de licence 3, de 1ère année de master et les notes de 2ème année de master disponibles ou équivalent pour les écoles d'ingénieurs ;
- noms de personnes pouvant fournir des informations sur vos compétences et votre travail.