M1 - PLDAC 2018

Analyser l'évolution des sciences :

Analyse de l'évolution de domaines de recherche à partir d'archives bibliographiques

Encadrants : B. Amann, H. Naacke Contact : bernd.amann@lip6.fr

Contexte

Le projet proposé s'inscrit dans le projet EPIQUE [3] qui étudie l'évolution des domaines de recherche dans le temps lorsque des nouveaux domaines émergent et d'autres fusionnent ou s'estompent. L'objectif de ce projet consiste à étendre et à valider une nouvelle approche fondée sur la notion de *ensembles fréquents maximaux (maximum frequent itemset)* pour représenter des domaines de recherche présents dans une période de temps et pour caractériser l'évolution de domaines scientifiques entre les différentes périodes.

Travail à faire.

1) Analyser l'approche existante fondée sur les algorithmes FPGrowth [1] et FPMax pour le calcul des ensembles fréquents maximaux de termes (dénotés *MFI* pour *maximal FI*) sur la plateforme Spark.
2) Proposer une solution pour sélectionner parmi les ensembles de termes obtenus, ceux qui sont les plus représentatifs d'un domaine scientifique. Implanter la solution sur la plateforme de calcul parallèle Spark.
3) Proposer une solution pour comparer les domaines obtenus pour deux périodes de temps différentes. Le but est d'identifier l'évolution qui s'est produite entre deux périodes de temps. Illustrer la solution sur les données de HAL, Arxiv ou MedLine [2].

Prérequis : Programmation (Java / Scala), Bases de données

Ce projet s'adresse surtout aux étudiants qui sont attirés par les thématiques des UEs BDR, BI et TALN.

Références :

[1] FPGrowth Frequent Pattern Mining - RDD-based API
https://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html

[2] Medline dataset https://mbr.nlm.nih.gov/Download/

[3] EPIQUE Projet ANR http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start