Analyse de l'évolution des sciences :

Construction de topics à partir de publications scientifiques

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

Contexte

L'évolution des connaissances scientifiques est directement liée à l'histoire de l'humanité. Les archives documentaires et bibliographiques comme le « Web Of Science » (WoS) ou PubMed représentent des sources fécondes pour l'analyse et la reconstruction de cette évolution.

Le projet proposé part des travaux visant à représenter l'évolution de topics au fil des ans. Pour chaque année, un ensemble de topics est calculé. L'évolution d'un topic est représentée par les liens entre des topics de l'année n et les topics similaires de l'année suivante. Les solutions actuelles sont limitées au traitement de corpus de taille moyenne et à une utilisation non interactive. Notre objectif est de développer des solutions performantes pour générer les topics et les liens de similarité. En particulier, le projet utilisera les technologiques récentes (Apache Spark et Flink) pour le calcul parallèle sur des données complexes et volumineuses.

Objectifs

L'objectif de ce stage est d'étudier la construction de topics à partir d'une base de documents scientifiques. On dispose de la base de données médicale MedLine [2] contenant 11 millions de documents publiés de 1975 à 2015. On connait l'année d'un document et la liste des termes qu'il contient. Autrement dit, on connait les relations (terme, document) et (document, année).

1) **Calcul des topics**. Etudier l'algorithme parallèle FPGrowth [1] construisant des topics définis par des ensembles fréquents de termes. Etendre l'algorithme pour calculer seulement les ensembles maximaux (un ensemble est maximal s'il n'est contenu dans aucun autre ensemble). Implanter la solution en scala.

2) Impact de la **segmentation temporelle** des documents sur le calcul des topics. Initialement, les documents sont découpés par tranches selon leur année de publication. Pour chaque tranche de temps les topics sont calculés. Il s'agit d'étudier l'impact de la taille d'une tranche (ex. tranche d'un an ou de 2 ans) sur les topics obtenus. On se pose deux questions (i) Est-ce que des tranches plus ou moins grandes aboutissent à des topics plus ou moins similaires ? (ii) Est-ce que la taille des tranches a une influence sur l'évolution des topics ? Proposer une méthode pour l'investigation de chaque question. Apporter des réponses empiriques en expérimentant sur la base Medline.

Prérequis : java

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