Stage M2 - 2020

Interrogation de motifs d'évolution temporelle des domaines scientifiques

Contacts, lieu et durée

- Bernd AMANN bernd.amann@lip6.fr Hubert NAACKE hubert.naacke@lip6.fr
- Equipe Bases de Données du LIP6

- 6 mois à partir de 1er mars 2020

Contexte

L'analyse thématique de corpus scientifiques permet de mieux comprendre les avancées effectuées dans les différents domaines de la science [1]. Au fil des années et des découvertes, les domaines scientifiques évoluent progressivement. Certains sujets prennent de l'ampleur dans un domaine jusqu'à devenir un domaine à part entière. Inversement d'autres sujets peuvent être délaissés dans certains domaines. Analyser la façon dont les domaines scientifiques se forment et évoluent dans un corpus qui rassemble des articles scientifiques sur plusieurs années ou décennies (comme par exemple HAL, arXiv ou Web of Science) permet aux historiens et philosophes de valider leurs modèles d'évolution de la science, aux scientifiques de mieux positionner leurs contributions dans le contexte historique, aux financeur de mieux prédire des évolutions technologiques etc…

Par exemple, on peut se poser les questions suivantes pour mieux comprendre l'engouement autour de l'IA : Est-ce que l'A est la même aujourd'hui que dans les années 2000 ? En quoi est-elle différente ou similaire ? Quels sont les domaines se rapportant à l'IA ? Est-ce que l'IA est une thématique transversale à plusieurs domaines ou plutôt au cœur de certains domaines ?

Objectifs généraux

Ce stage s'inscrit dans le projet ANR EPIQUE [2] dirigé par l'équipe Bases de Données laboratoire LIP6. Dans ce projet nous avons conçu un workflow d'analyse de grands corpus de documents scientifiques. Ce workflow est une application implanté en Scala avec la plate-forme SPARK, et utilise la méthode LDA (Latent Dirichlet Allocation) pour représenter un domaine scientifique (topic) sous forme d'un ensemble de **termes** pondérées.

Les documents sont tout d'abord regroupés par *périodes de temps successives* définies par des fenêtres de temps glissantes (par exemple des fenêtres de 3 ans qui se déplacent de 2 ans). Puis pour chaque période, les domaines scientifiques sont extraits avec la méthode LDA. Ensuite les domaines des périodes consécutives sont comparés afin de relier les domaines les plus similaires. On obtient ainsi des **graphes d'évolution** qui permettent de comprendre l'évolution des domaines : à partir d'un domaine d'une certaine année, on peut suivre son évolution au cours des années suivantes et savoir quels sont les termes qui apparaissent, demeurent ou disparaissent progressivement. **L'objectif principal du stage est de concevoir une solution pour interroger et visualiser ces graphes d'évolution.**

Travail à réaliser :

- Comprendre le modèle d'évolution reliant des domaines entre les périodes basé sur la similarité entre les domaines. Comprendre ce qu'est un graphe d'évolution. Recenser les caractéristiques structurelles d'un graphe qui peuvent être utiles pour indiquer dans quelle mesure un graphe évolue : taille, nombre d'année, degré, termes apparus ou disparus, ...

- Etudier des méthodes pour interroger les graphes d'évolution : recherche de topics par contenu, extraction de sous-graphes par des motifs d'évolution, filtrage par des relations d'évolution entre topics, etc.
- Mise en œuvre sous forme de notebooks Python/Scala et Spark
- Expérimentation avec des graphes d'évolution extraits de grands corpus comme ISTEX (Wiley, Nature, Elsevier) et Arxiv, en collaboration avec un doctorant.

Pour la réalisation et la validation expérimentale, le stagiaire aura à sa disposition la plateforme de calcul Apache Spark du LIP6 (déployée sur 16 machines totalisant plus de 300 cœurs et 800GB de RAM).

Prérequis :

Très bonne expérience en programmation : Python, Java/Scala, Apache Spark, notebook

Références bibliographiques:

[1] D. Chavalarias, P. Cointet. Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields, Plos One 2013, https://doi.org/10.1371/journal.pone.0054847
[2] EPIQUE Projet ANR http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start