ÉLÉMENT DE PORTFOLIO 04

Vidéo

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : EPIQUE : Extracting Meaningful Science Evolution Patterns from Large Document Archives

URL de l'élément : `http://nextcloud.iscpif.fr/index.php/s/KBJk5MEyyoweQcp`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cette video a été produite dans le cadre du projet ANR interdisciplinaire EPIQUE [1] (2017-2021) porté par l'équipe BD et avec trois partenaires académiques : l'Institut des Systèmes Complexes (ISC-PIF), l'IRISA de Rennes et une UMR en philosophie des sciences de Paris 1 (IHPST). Elle est un des résultats importants de la thèse de Ke LI (2018-2021) [1]. Ce travail donné lieu à trois publications internationales dont une revue.

Le projet Epique a permis à l'équipe BD d'acquérir de nouvelles compétences en fouille de données textuelles et ainsi de s'ouvrir à de nouveaux axes de recherche mais en les abordant, de manière différente, à travers un prisme "BD" qui cible en priorité la gestion efficace des données accompagnée de langages hautement expressifs pour les manipuler. Il a ouvert des perspectives intéressantes qui sont actuellement explorées dans la thèse de Hamed Rahimi (démarrée en 2021).

3 PRÉSENTATION DE CET ÉLÉMENT

Le projet Epique part du constat qu'il existe une demande croissante d'outils pratiques pour explorer l'évolution de la recherche scientifique publiée dans des archives bibliographiques telles que le Web of Science (WoS), arXiv, PubMed ou ISTEX. Mettre en évidence des modèles d'évolution à partir de ces archives documentaires est un besoin exprimé dans de nombreux cas d'usage, par exemple en histoire des sciences ou dans la gouvernance de l'activité scientifique.

Le projet Epique vise à reconstruire l'évolution des sciences dans le temps, en utilisant des méthodes quantitatives pour analyser des grands corpus textuels de publications scientifiques. Il s'agit de produire une *cartographie* des sciences au fil du temps. Les domaines, représentés par des mots-clés, sont positionnés par période croissante : des plus anciennes en haut de la carte aux plus récentes en bas de la carte. Pour représenter le changement progressif du contenu d'un domaine, des liens connectent les domaines similaires appartenant à deux périodes consécutives. La structure de la carte est un graphe complexe : il contient des motifs d'évolution qu'il s'agit de mettre en évidence. Une des nouveautés du projet est de rendre les cartes **interactives** afin d'augmenter les possibilités d'exploration et à terme découvrir des nouvelles formes d'évolution des sciences. Un modèle pour caractériser des motifs d'évolution a été défini ainsi qu'un langage déclaratif permettant d'interroger des graphes d'évolution. Le langage a l'avantage d'être composable afin de pouvoir rechercher des motifs en exprimant successivement plusieurs requêtes.

Ce travail a fait l'objet d'une publication dans la revue Big Data Research [3] et une autre dans le workshop BigVis@EDBT [2].

De plus, nous avons abordé un problème de performance soulevé lors de l'alignement temporel des domaines scientifiques représentés par des vecteurs. Un algorithme de calcul parallèle et distribué de similarité entre les domaines a été conçu et publié dans la conférence IEEE BigData [4].

Actuellement nous poursuivons les travaux sur l'analyse de l'évolution des sciences en nous appuyant sur les avancées récentes dans les domaines des modèles de langues et des embeddings de graphe. Les données textuelles des documents scientifiques sont combinées avec des graphes de citation pour représenter l'influence des communautés scientifiques sur les évolutions thématiques.

1. http ://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start

Enfin, notre expertise en gestion efficace de grands volumes de données est mise à profit pour concevoir des solutions d'optimisation de chaînes de traitement manipulant à la fois des masses de données textuelles et des grands graphes.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Ke Li. *Exploring Topic Evolution in Large Scientific Archives with Pivot Graphs*. Theses, Sorbonne Université, June 2021.

[2] Ke Li, Hubert Naacke, and Bernd Amann. Exploring the Evolution of Science with Pivot Topic Graphs. In *International Workshop on Big Data Visual Exploration and Analytics BigVis at EDBT 2020*, Copenhague, Denmark, March 2020.

[3] Ke Li, Hubert Naacke, and Bernd Amann. An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science. *Big Data Research*, 26 :100247, November 2021.

[4] Hubert Naacke, Ke Li, Bernd Amann, and Olivier Curé. Efficient similarity-based alignment of temporally-situated graph nodes with Apache Spark. In *High Performance Big Graph Data Management, Analysis, and Mining*, IEEE Int'l Conference on Big Data, Los Angeles, United States, December 2019.