

Post-Doctorant ESILV

Titre : Fouille de motifs fréquents pour l'analyse de comportement touristique circulatoire

Encadrant : Nicolas Travers

Equipe référente : Gaël Chareyron, Sonia Djebali, Imen Ouled Dlala

Mots-clés : Pattern Mining, Neo4j, Pregel

Description

L'appréciation des visites effectuées par les touristes est un enjeu majeur dans le monde du tourisme afin d'anticiper les évolutions de tendances, mais aussi la manière dont ils circulent sur le territoire. Une approche permettant d'estimer cette appréciation est de reposer sur l'extraction de motifs fréquents sur un graphe de circulation, comme l'extraction de Graphlet [1], k-decomposition [2], ou encore les structures cohésives comme les k-plex [6]. Ainsi, les tendances touristiques sont extraites grâce à leurs fréquences d'apparition de manière topologique.

Toutefois, les données touristiques provenant de sites prescripteurs d'expérience, tels que TripAdvisor, donnent lieu à des volumes difficiles à intégrer dans les techniques traditionnelles de fouille de données. En effet, avec un grand nombre de lieux visités (millions), et un nombre énorme de commentaires laissés par les utilisateurs (milliards), il est nécessaire de développer une nouvelle approche pour le passage à l'échelle d'algorithmes basés sur les graphes. De plus, la compréhension du territoire se fait également par l'analyse du graphe à plusieurs échelles.

Pour ce faire, au sein du groupe digital du DVRC, nous travaillons sur le développement en Pregel [3] de différentes approches existantes pour pouvoir définir la meilleure stratégie de fouille de motifs tout en les adaptant à des problématiques de résumer de graphes à différentes échelles. De plus, l'aspect géodésique des données est un facteur important lié à la topologie des données [4, 5, 7], tout autant que la fréquentation. Nous étudions donc l'adaptation des méthodes existantes pour améliorer l'efficacité de la fouille de motifs basée sur ces informations [8, 9, 10, 11].

Ce postdoctorat est financé grâce à un projet lié à la Métropole Européenne de Lille (MEL2). Dans ce cadre, nous collaborons avec la Métropole et l'ADULM pour les questions métiers, mais surtout avec le laboratoire EIREST de Panthéon-Sorbonne sur la recherche en géographie liée au tourisme. Un autre postdoctorant sur ces aspects a déjà débuté ses travaux et sera en interaction avec le candidat. Le but étant d'apporter de la transdisciplinarité dans ces travaux en intégrant les questions métiers du tourisme au sein même des stratégies de fouille de motifs. Nous accorderons aussi une importance à l'interprétation des résultats pour traduire les motifs pour la compréhension des comportements touristiques sur le territoire Lillois. Pour finir, les données pourront être complétées avec des informations multidimensionnelles disponibles au sein du projet MEL2 (Booking, Airbnb, Observatoire du tourisme à la MEL, etc.).

Le but de ce postdoc est donc multiple :

- Intégrer des méthodes de fouille de motifs dans une base de données Neo4j (en Pregel/Java) ;
- Améliorer une méthode pour donner une heuristique de fouille adaptée au contexte géodésique ;
- Traduction des motifs fréquents sur un comportement touristique à différentes échelles ;
- Développement sur la plateforme [eKhonsou](#) avec données multidimensionnelles.

Profil et Compétences attendues

Niveau Doctorat en informatique.

Connaissances en bases de données, Data Mining, BD graph (Neo4j, Cypher), Java programmation répartie, node.js.

Lieu du Postdoc

Laboratoire de recherche De Vinci Research Center au sein de l'École Supérieure d'Ingénieurs Léonard de Vinci ; Paris, la Défense.

Intervention à Lille avec l'équipe de la MEL et l'EIREST pour les revues de projet.

Période & Salaire

12 mois, 35k€ brut annuel (+ mutuelle/titre restaurant/intéressement)

Début prévu en juin 2023 au plus tôt

Candidature

Les candidat.e.s sont invité.e.s à nous envoyer un mail à nicolas.travers@devinci.fr avec :

CV indiquant leurs expériences et compétences

Une lettre de motivation & recommandation

Rapports des rapporteurs de la thèse

- [1] XIAOWEI CHEN and JOHN C. S. LUI. Mining Graphlet Counts in Online Social Networks. In TKDD, pages 1–38, 2018.
- [2] Lijun Chang, Jeffrey Xu Yu, Lu Qin, Xuemin Lin, Chengfei Liu, Weifa Liang, Efficiently Computing k-Edge Connected Components via Graph Decomposition. In SIGMOD, 2013
- [3] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: A System for Large-Scale Graph Processing. In SIGMOD 2010, pages 135-145
- [4] A. Wu, M. Garland, J. Han. Mining Scale-free Networks using Geodesic Clustering. In KDD, 2004
- [5] A. Bendimerad, A. Mel, J. Lijffijt, M. Plantevit, C. Robardet, T. De Bie. SIAS-miner: mining subjectively interesting attributed subgraphs. Data Mining and Knowledge Discovery (2020)
- [6] A. Conte, T. De Matteis, D. De Sensi, R. Grossi, A. Marino, L. Versari. D2K: Scalable Community Detection in Massive Networks via Small-Diameter k-Plexes. KDD 2018, London, UK, August 19-23, 2018.
- [7] R. Espejo, G. Mestre, F. Postigo, S. Lumbreras, A. Ramos, T. Huang, E. Bompard. Exploiting graphlet decomposition to explain the structure of complex networks: the GHuST framework. Scientific Reports (2020).
- [8] Carlos H. C. Teixeira, Alexandre J. Fonseca, Marco Serafini, Georgos Siganos, Mohammed J. Zaki, and Ashraf Aboulnaga. 2015. Arabesque: a system for distributed graph mining. SOSP '15, NY, USA, 425–440.
- [9] Jonas Fischer and Jilles Vreeken. 2021. Differentiable Pattern Set Mining. KDD '21, New York, NY, USA, 383–392.
- [10] Francisco Neves, Anna Finamore, and Rui Henriques. 2021. Efficient discovery of emerging patterns in heterogeneous spatiotemporal data from mobile sensors. In MobiQuitous 2020, New York, USA, 158–167.
- [11] Lin Hu and Lei Zou. 2022. A GPU-based Graph Pattern Mining System. CIKM '22. New York, USA, 4867–4871.