

# Stage Recherche et Développement ESILV

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

Encadrants : Imen Ouled Dlala, Nicolas Travers  
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]. 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 sont d'un volume difficile 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.

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. De plus, l'aspect géodésique des données est un facteur important lié à la topologie des données [4, 5], tout autant que la fréquentation. Nous étudions donc l'adaptation des méthodes existantes pour améliorer l'efficacité de la fouille de motifs basé sur ces informations.

Le but de ce stage est donc double :

- 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.

### *Profil et Compétences attendues*

Étudiante ou étudiant de niveau M1/M2 en informatique (Master ou école d'ingénieurs).

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



### ***Lieu du stage***

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

### ***Période***

Stage de 5 mois à effectuer à partir de mars - début avril 2021 (900€ pour M2 – 592€ pour M1).

### ***Candidature***

Les candidat.e.s sont invité.e.s à nous envoyer un mail à [nicolas.travers@devinci.fr](mailto:nicolas.travers@devinci.fr) avec :

CV indiquant leurs expériences et compétences

Une lettre de motivation

Les bulletins de notes des deux dernières années.

- [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, pages 135-145, 2010
- [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) 34:355–393