Research Internship at l'ESILV

Optimization of Frequent Pattern Mining for Tourist Behavior Analysis

Hugo Alatrista-Salas, Sonia Djebali, Imen Ouled Dlala, Nicolas Travers

Keywords: Pattern Mining, Neo4j, Pregel

Description

Understanding the appreciation of visits made by tourists is a major issue in the tourism sector to anticipate trend evolutions as well as how they move across the territory. One approach to estimating this appreciation is based on the extraction of frequent patterns on a circulation graph, such as Graphlet extraction [1], k-decomposition [2], or cohesive structures like k-plexes [6]. Thus, tourism trends are extracted using their frequency of occurrence in a topological manner.

However, tourism data from experience-recommending platforms such as TripAdvisor or Google Maps results in large data graphs that become challenging to process with traditional data mining techniques. With a large number of places visited (millions) and an enormous number of user comments (billions), it is necessary to develop a new approach for scaling graph-based algorithms.

To this end, within the STARCS axis of DVRC, we have developed an exhaustive and scalable pattern extraction approach on a graph using Pregel [3]. This approach allows us to extract both the pattern topology and node properties, including geodesic information [4, 5, 7]. The extraction has been extended to complex patterns giving interesting perspectives of enhancement. We now wish to take this approach a step further by focusing on optimizing the mining process.

The internship has two main goals:

- Use a topological signature technique to mine patterns in a Neo4j database (in Pregel/Java).
- Improve the method to provide a heuristic adapted to the geodesic context.

Example of aggregated tourist propagation graph across the French territory:

- How can we identify significant propagation patterns?
- What are the characteristics of a pattern?
- Can we extract seasonality from different groups of patterns?

Profile and expected skills

M2 level students (Master or Engineering Schools).
Databases, Data Mining, graph DB (Neo4j, Cypher), Java, parallelism.

Location

De Vinci Research Center at ESILV at (École Supérieure d'Ingénieurs Léonard de Vinci ; Paris, la Défense).

Duration

6 months (from march or before- 900€/month).

Application

Send you CV, last grades (M1/M2), motivation letter and recommendation letters to:

- nicolas.travers@devinci.fr

[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.
[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. Conference on Knowledge Discovery & Data Mining, {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).