

The subgraph isomorphism problem in graph data mining and its application in frequent subgraph discovery

Pham Thi Lien

Abstract: This thesis investigates the subgraph isomorphism problem within the context of graph data mining, with a specific focus on its application to frequent subgraph discovery. Subgraph isomorphism, an NP-complete problem, serves as a core computational bottleneck in identifying recurring patterns across large graph datasets. This study systematically analyzes existing algorithmic approaches, including Ullmann's algorithm, VF2, and advanced filtering techniques based on canonical labeling and graph invariants, to evaluate their efficiency and scalability. The research proposes an optimized framework that integrates pruning strategies and indexing methods to reduce the search space during isomorphism testing. The primary application is the detection of frequent subgraphs, which are critical for characterizing common structural motifs in domains such as bioinformatics, social network analysis, and chemoinformatics. Experimental results on both synthetic and real-world datasets demonstrate that the proposed method significantly improves runtime performance while maintaining accuracy. The findings contribute to advancing graph mining techniques, offering a robust solution for extracting meaningful patterns from complex relational data.

Keywords: subgraph isomorphism, graph data mining, frequent subgraph, subgraph detection