Abstract

Graph structure is useful for representing complex non-linear relationships among entities in a wide variety of information systems. As massive amount of non-linear structured data become increasingly available to the public in the applications such as traffic network, social network, biological network, and technological and information networks, graph mining research has been becoming an active research in recent decade. Graph Mining problems have many variants due to various restrictions. Literature surveys and some preliminary results of an on-going graph mining algorithm research are shared in this presentation. General challenges of graph mining problems, possible restricted variants and possible optimization techniques are discussed.

Related Work

qSpan: efficient canonical labeling to reduce redundancy AGM and FSG: Adapting frequent itemset mining algorithms to graph mining SPIN and GASTON: Mining and extending simple subgraphs (trees, paths) MULE and RNGV: DFS based - Unique labeled directed Graph, exact and inexact subgraph mining GraphGrep: mainly a graph indexing and searching algorithms using suffix paths Otminer: Restrictedly Embedded Ordered Tree Mining Algorithm

Challenges

The exponential nature of the number of subgraphs. Subgraph Isomorphism: Fro counting frequencies, it is necessary to check whether a given graph is a subgraph of another. In general form, subgraph checking is an NP-complete Canonical Labeling: To avoid redundancy while generating subgraphs, canonical labeling of graphs is necessary. Equivalent to subgraph isomorphism. Connectivity: Patterns of interest are generally connected, so it is necessary to only generate connected subgraphs. Weakly or strongly connected. However, disconnected patterns might be also meaningful.

Several Issues Under Study

- New canonical form design and properties of the canonical form
- Optimize existing or design new subgraph enumeration algorithm
- Optimize existing or design new subgraph detection algorithm
- These techniques are then integrated into an efficient algorithm to solve the graph mining problem at hand.

Aim of the project

The ultimate goal of the research is to lead to designing, implementing and evaluating novel and relatively efficient algorithms to mine frequent sub-graphs from graph databases under specially defined restrictions. This is an on-going research.

Research Plan

1. Extend the author’s previous tree mining research to graph mining.
2. Conduct literature review of existing graph mining algorithms
3. Develop Synthetic dataset generators
4. Obtain a small amount of real world dataset
5. Identify new graph mining problems
6. Understand and Optimize the new graph mining algorithms
7. Design and implement new graph mining algorithms
8. Investigate parallelizable components of Graph Mining Algorithms

References


Wikipedia


Acknowledgements

Jason Wang, Professor of Computer Science and Bioinformatics, NJIT for providing sample datasets and collaborating on the research.