Summary on Graph Mining Techniques

May 26, 2015

1 A general review on graph mining techniques

1. Graph mining can be used to mine spatial-temporal data
2. Graphs have been used in biology, chemistry, social networks, communications and etc. It can capture the relations between objects far beyond flattened representations.
3. Graphs are a set of nodes joined by a set of lines (i.e., undirected graphs) or arrows (directed graphs)
    - Planar: can be drawn with no 2 edges crossing
    - Non-planar:
        - Bipartite: if the graphs are non-planar and vertex set can be partitioned into S and T such that every edge has one end in S and one end in T.
        - Complete: if the graphs are non-planar and each node is connected to every other node
4. Connected means it is possible to get from any node to any other node by following a sequence of adjacent nodes
5. Cycle occurs when there is a path that starts from a particular node and returns to that same node
6. The edges can be directed, undirected, and weighted. Note that the weight means the cost associated to traverse the edge and it is used in graph-related algorithms, such as the MST.
7. For undirected G, the degree is the number of edges incident to the node. For directed G, indegree is the number of edges coming into the node and outdegree is the number of edges going out of the node
8. Graphs representations
    - Adjacency list
    - Adjacency matrix, use 0 and 1 as entries.

- Incidence matrix, i.e., use -1, 0, 1 as entries.

9. The edit distance between two graphs is the shortest or the least cost sequence of elementary graph edit operations that transform one graph into the other. The elementary edit operations include rotation, substitution, deletion, insertion of vertices and edges.

10. For an approximate graph matching problem, the cost includes $C_{vd}, C_{vi}, C_{vs}, C_{es}$, which means the costs associated with vertex deletion, vertex insertion, vertex substitution, and edge substitution. Note that C_{ed}, C_{ei} are assumed to be included in the costs of the corresponding vertex deletions and insertions. The edit distance between two graphs G_1 and G_2 is the least cost approximate graph matching from G_1 to G_2.

11. The average path distance is the average distance between any two entities, i.e., the average length of the shortest path connecting each pair of entities (edges are unweighted and undirected).

12. Clustering coefficient is a measure of how clustered, or locally structured, a graph is. In another words, it is an average of how interconnected each entity's neighbors are.

13. Frequent subgraph discovery stems from searching for frequent items in association rules discovery. Apriori-based methods have been developed for FSM. Testing for graph isomorphism is needed as

    - Candidate generation step: to decided whether a candidate has been generated
    - Candidate pruning step: to check whether $k-1$-subgraphs are frequent
    - Candidate counting step: to check whether a candidate is contained within another graph

2 A survey of frequent subgraph mining algorithms

1. Structured data and semi-structured data are naturally suited to graph representations.

2. Frequent subgraph mining (FSM) is the essence of graph mining. The aim is to extract all frequent subgraphs in a given data set, whose occurrence counts are above a specified threshold.

3. The straightforward idea behind FSM is to grow candidate subgraphs, in either a breadth-first or a depth-first manner, and then determine if the identified candidate subgraphs occur frequently enough in the graph data set.

4. Two types of FSM, i.e., graph transaction based FSM, and single graph based FSM.

5. A simple graph is un-weighted and un-directed with no loops and no multiple links between any two distinct nodes. The graphs used in FSM are assumed to be labelled simple graphs. A labelled graph indicates that both vertices and edges are labelled.

6. Free tree indicates an undirected graph that is connected and acyclic.

7. Subgraph isomorphism are divided into exact matching or error tolerant matching. Most FSM algorithms adopt exact matching.

8. FSM techniques are divided into two types

    - Apriori-based approaches: in a generate-and-test manner using a breadth-first search (BFS) strategy to explore the subgraph lattices of a given data set.
    - Pattern growth-based approaches: in a depth-first search (DFS) strategy. Each subgraph is extended recursively until all frequent supergraphs are discovered.
    - While subgraph isomorphism is NP-complete, subtree isomorphism can be solved in $O(\frac{k^1.5}{log(k)} \times n)$ time.
    - Frequent graph mining
        - Inexact FGM (for subgraphs, as compared with FTM for subtrees)
            * SUBDUE uses the minimum description length principle to compress the graph data and a heuristic beam search method is used to narrow down the search space. The scalability is an issue as the run time does not increase linearly with the size of the input graph. In addition, it tends to discover only a small number of patterns.
            * GREW aims to find connected subgraphs which have many vertex-disjoint embedding in single large graphs.
        - Exact FGM
            * It can be applied to graph transaction based mining or single graph based mining.
            * It is guaranteed to find all frequent subgraphs in the input data.
            * For graph transaction based minng
                · BFS: tends to be more efficient as it allows for pruning of infrequent subgraphs. Representative algorithms include AGM, AcGM, FSG, gFSG, and DPMine

    - DFS: requires less memory usage in exchange for less efficient pruning. Representative algorithms include MoFa, gSpan, ADI-Mine, FFSM, and GASTON. gSpan is arguably the most frequently cited FSM algorithm.
- SUBDUE tends to find only small sizes patterns, consequently it may miss interesting larger patterns. FFSM and GASTON cannot be used in the context of directed graphs. gSpan can be used for directed graphs with some minor changes.

3 Graph Mining Chapter

1. Graphs can be used to model complex structures. Many graph search algorithms have been developed in chemical informatics, computer vision, video indexing, and text retrieval.

2. The discovery of frequent substructures usually consists of two steps, i.e., generate frequent substructure candidates and frequency check. The second step involves a subgraph isomorphism test whose computational complexity is NP-complete.

3. Two types of FSM, i.e., Apriori-based, and pattern-growth approaches.

4. Social network analysis, from a data minig perspective, is also called link analysis or link mining.