Mining frequent subgraphs from 'easy' classes

Jan Ramon[1], Department of computerscience, K.U.Leuven

1. Introduction

Recently, there is an increasing interest in mining structured data. Several frequent subgraph mining systems have been proposed. However, these usually consider general graphs. One can show that frequent subgraph mining for general graphs can not be performed in output-polynomial time.

In practice however, data usually does not consist of arbitrary graphs but has a much simpler structure. In this paper, we first summary recent results on efficient mining of outerplanar graphs. Next, we outline a number of ideas and open questions.

2. Mining outerplanar graphs

Outerplanar graphs are graphs that can be drawn in a plane without crossing edges in such a way that all vertices are adjacent to the outer face. In [1], we point out that 94.5% of the molecules stored in the NCI database (containing 250251 molecules) can be represented as outerplanar graphs. Moreover, only a few blocks contain more than 10 diagonals. We propose a new matching operator inspired by the chemical application called BBP subgraph isomorphism. We then show that given a set of t-tenuous outerplanar graphs, we can enumerate all frequent patterns under BPP subgraph isomorphism with polynomial delay (one needs per pattern $O(k^3m)$ time with k is the size of the pattern and m is the number of examples). Experiments show that our algorithm is also efficient in practice.

3. Future work

We can extend the above results in several ways. First, one can obtain similar results in the case where homomorphism is used as matching operator [2]. Second, one can improve the implementation by reducing the number of passes over the database and by remembering embeddings or coverage lists (but this is memory-expensive). Third, it is natural to look at larger classes of graphs that can be efficiently mined. In literature many graph classes that are common in practical applicatins have been show to satisfy favorable complexity propoerties. In each of these cases, one can investigate whether the candidate generation and/or the pattern frequency counting can happen in incremental polynomial time.

References

[1] T. Horvath, J. Ramon, and S. Wrobel. Frequent subgraph mining in outerplanar graphs. In *Proc. of KDD'2006*, pp. 197–206, Philadelphia, PA, August 2006.

[2] Jan Ramon. Efficient mining of frequent outerplanar graphs. In *Proc. of ILP'2006 – short papers*, pp. 170–172, 2006.

[1]Jan Ramon is a post-doctoral fellows of the Fund for Scientific Research (FWO) of Flanders