SkyGraph: An Algorithm for Important Subgraph Discovery in Relational Graphs*

Apostolos N. Papadopoulos, Apostolos Lyritsis, and Yannis Manolopoulos

Data Engineering Research Lab.
Department of Informatics, Aristotle University
54124 Thessaloniki, Greece
{apostol,lyritsis,manolopo}@delab.csd.auth.gr

Graph mining is gaining importance due to the numerous applications that rely on graph-based data. Some example applications are: (i) analysis of microarray data in bioinformatics, (ii) pattern discovery in social networks, (iii) analysis of transportation networks, (iv) community discovery in Web data. Existing pattern discovery approaches operate by using simple constraints on the mined patterns. For example, given a database of graphs, a typical graph mining task is to report all subgraphs that appear in at least s graphs, where s is the frequency support threshold. In other cases, we are interested in the discovery of dense or highly-connected subgraphs. In such a case, a threshold is defined for the density or the connectivity of the returned patterns. Other constraints may be defined as well, towards restricting the number of mined patterns. There are three important limitations with this approach: (i) there is an on-off decision regarding the eligibility of patterns, i.e., a pattern either satisfies the constraints or not, (ii) in the case where the constraints are very strict we risk an empty answer or an answer with only a few patterns, and (iii) in the case where the constraints are too weak the number of patterns may be huge.

Towards dealing with the previous limitations, we address the problem of incorporating preferences in the pattern discovery process and we propose the SkyGraph algorithm which is based on min-cut computations. Each subgraph is seen as a record containing two attributes: (i) the order (number of vertices) and (ii) the edge connectivity. The importance of a discovered subgraph increases as both the order and the edge connectivity increase. Therefore, the best possible subgraphs (termed skyline subgraphs) are the ones that are maximized both in order and edge connectivity. To the best of the authors' knowledge, this is the first work studying the skyline problem in the process of knowledge discovery. The performance of the proposed technique is evaluated by using real-life as well as synthetically generated random graphs.

References

1. Papadopoulos, A.N., Lyritsis, A., Manolopoulos, Y.: Skygraph: An algorithm for important subgraph discovery in relational graphs. *Data Mining and Knowledge Discovery* 17(1), 57–76 (2008)

* This is an extended abstract of an article published in the *Data Mining and Knowledge Discovery* journal [1].