Research Proposal

Sampling Informative Patterns From Large Networks

Mostafa H. Chehreghani and Albert Bifet and Talel Abdessalem

Telecom ParisTech

Contact: mostafa.chehreghani@gmail.com

During the last decade or so, the amount of data that is generated and becomes publicly available is rapidly growing. This makes it impossible to extract useful information from this huge amount of data manually without using automatic tools and algorithms. Furthermore, in many applications, such as Bioinformatics, the world wide web, social and technological and communication networks, data are usually represented with graphs. This makes mining and analyzing large networks practically interesting, while also a challenging research area, due to high computational cost involved in processing graph data. A key task in mining large networks is *frequent pattern discovery*.

The set of all frequent patterns that are extracted from a graph dataset can be huge. A technique recently proposed for obtaining a compact, informative and useful set of patterns is *output sampling*, where a small set of frequent patterns is randomly chosen [1]. However, existing algorithms work only in the *transactional setting*, where the database consists of a collection of relatively small graphs. Furthermore, they sample patterns uniformly or based on some simple criteria such as *support*. In this project will try to address these issues. First, we will extend the sampling framework to the *single network setting* where the database is a large single graph and counting supports of patterns is more complicated [2]. Second, we will propose sampling techniques that are based on more interesting/informative measures or those that are specific to large single networks, e.g., the product of the size with the support [3] and network compressibility [4]. Third, we will extend the framework to the streaming and/or semi-streaming settings. Here, the main challenge is to count supports of several patterns that are candidates for the samples, during one or a few passes over the stream. It seems that given a fixed number of possible passes over the stream, there is a *trade-off* between the sample size on the one hand and accuracy of frequency counting on the other hand. An in-depth investigation of this trade-off for different networks will be interesting. We may also require to propose measure that are specific to the streaming/semi-streaming settings.

References

[1] Mohammad Al Hasan, Mohammed J. Zaki: Output Space Sampling for Graph Patterns. PVLDB 2(1): 730-741 (2009).

[2] Toon Calders, Jan Ramon, Dries Van Dyck: All normalized anti-monotonic overlap graph measures are bounded. Data Min. Knowl. Discov. 23(3): 503-548 (2011).
[3] Floris Geerts, Bart Goethals, Taneli Mielikäinen: Tiling Databases. Discovery Science 2004: 278-289.
[4] Diane J. Cook, Lawrence B. Holder, Surnjani Djoko: Knowledge Discovery from Structural Data. J. Intell. Inf. Syst. 5(3): 229-248 (1995).