FOG: Finding Outerplanar Graphs

Jan Ramon
Dept. of Computer Science
K.U.Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
Jan.Ramon
@cs.kuleuven.be

Tamás Horváth
Dept. of Computer Science III
University of Bonn and
Fraunhofer Institute AIS
Sankt Augustin, Germany
tamas.horvath
@ais.fraunhofer.de

Leander Schietgat
Dept. of Computer Science
K.U.Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
Leander.Schietgat
@cs.kuleuven.be

ABSTRACT

In this demo we will present FOG, a system that mines frequent outerplanar graphs. We argue that outerplanar graphs are an interesting class as they can be mined efficiently and are practically relevant for chemical applications. Our system provides several features that allow a user to ask detailed mining questions and visualize the fragments found.

Keywords

Frequent pattern mining, outerplanar graphs, molecules

1. INTRODUCTION

In recent years there has been an increased interest in frequent pattern discovery in large databases of graph structured objects. While the frequent connected subgraph mining problem for tree datasets can be solved in incremental polynomial time, it becomes intractable for arbitrary graph databases. Existing approaches have therefore resorted to various heuristic strategies and restrictions of the search space, but have not identified a practically relevant tractable graph class beyond trees.

In recent work [2] we defined the class of so called *tenuous outerplanar graphs*, a strict generalization of trees, developed a frequent subgraph mining algorithm for tenuous outerplanar graphs that works in incremental polynomial time, and evaluated the algorithm empirically on the NCI molecular graph dataset (containing over 250,000 molecules).

Since the submission of [2] we have further extended our implementation, the FOG system. In particular, we added a number of features which help the user to constrain the search, and allow for visualization of data and patterns (at least, when using a database of molecules).

The FOG system aims at both being a framework for developing further extensions and other graph mining algorithms, and being a practically useful tool for performing data mining in a number of domains. We will highlight both aspects in our demonstration.

The remainder of this proposal is structured as follows. First, we motivate our application in Section 2. Next, we outline the features of our system in Section 3. In Section 4 we then sketch the planned demonstration. Section 5 concludes.

KDD demo session August 2006 Philadelphia, PA USA

2. MOLECULAR DATABASES

In earlier work (see e.g. [2, 3]) we have argued that the class of outerplanar graphs is a relevant one for applications. In fact, in one of the popular graph mining data sets (the NCI data set), 94.3% of all elements are tenuous outerplanar graphs.

On the other hand, we have also shown that all outerplanar subgraphs which are frequent w.r.t. a database can be enumerated in incremental polynomial time. The FOG frequent outerplanar graph miner is therefore able to gain substantial efficiency by exploiting the outerplanar structure of molecules when applied to chemical databases, while at the same times still producing most frequent subgraphs (e.g. all frequent subgraphs that are not outerplanar are guaranteed to have a frequency lower than 6% on the NCI database).

As our experimental results on this domain are promising, molecules are easy to visualize, and the chemical domain is an important application area for the data mining community, we will present our demonstration using the NCI database, start from a number of concrete data mining questions and visualize the results.

3. SYSTEM FEATURES

FOG's main features are the following:

- *Investigating properties of the database.* The FOG system is able to make statistics on the complexity of the graphs included in the database. This is useful for graph mining research, as it allows to take several measures of graph complexity (such as outerplanarity, treewidth, ...) into account when evaluating a new approach.
- *Mining frequent subgraphs.* The user can ask the FOG system queries asking for the set of all frequent satisfying certain constraints. Such constraints can include monotonic and anti-monotonic constraints such as being subgraphs of another graph, having a specific graph as a subgraph, have a maximal number of biconnected components, etc. or can start from earlier selections of subsets of the database.

- *Association Rules.* Based on sets of frequent subgraphs, FOG can generated association rules predicting frequent extensions of subgraphs. Several notions of association rules for relational data have been proposed [1] and here too there is room for experimentation and extensions.

- *Visualizing examples and patterns.* If a visualization component is available (currently there is only one for molecules based on a publicly available molecule viewer in Java), the user can visualize patterns, association rules and examples.

- *Refining results.* FOG allows for reusing results from earlier mining tasks, such that time can be saved by formulating new questions as a refinement of earlier ones.

- *Efficiency.* FOG is an efficient system, both in theory and in practice, and both for time and space. If necessary, the system can run with the database entirely on disk and depending on the setting, only a small number of passes over the database are needed to mine relatively large frequent subgraphs.

4. DEMO OUTLINE

The demonstration will essentially consist of four parts:

- First, we will briefly present the NCI database we will use.

- Next, we present an overview of the graph analysis module, at the same time demonstrating the FOG features for analyzing the complexity of graphs in a database and arguing that molecular graphs are usually in relatively easy graph classes.

- We will then focus on the demonstration of the data mining features. For this, we will start from a number of specific data mining questions and try to provide insight in some major parts of the database (such as the HIV part with about 43000 molecules) from the resulting frequent patterns and association rules. As the NCI database fits in memory and the algorithm is sufficiently fast for simple questions, most questions will be answered online. For some larger mining tasks involving the full dataset and patterns up to size 30 (the size of a pattern equals the number of cycles in a graph and the number of vertices not in a cycle), we will present pre-computed results.

- Finally, we will give some more details on the design of the algorithm, downloading the system and ideas for future work.

5. CONCLUSIONS

We propose to demonstrate the FOG system. An earlier version (with only the frequent pattern mining functionality) of the C++ source code of the FOG system can be downloaded from `http://www.cs.kuleuven.ac.be/~janr/fog/readme.html` and the version used for the demo will be available for download at the time of the conference. The system runs under Linux (or cygwin for Windows) and requires less than 10Mb of memory (excluding the database).

We argued that this system illustrates contributions at the conference and is interesting for applications. We believe that a demo will also trigger a number of interesting discussions on implementation techniques and future directions for graph mining.

Acknowledgements

Jan Ramon is a postdoctoral fellow of the Fund for Scientific Research (FWO) of Flanders.

6. ADDITIONAL AUTHORS

Stefan Wrobel (Fraunhofer Institute AIS and Dept. of Computer Science III University of Bonn, Sankt Augustin, Germany, email: `stefan.wrobel@ais.fraunhofer.de`).

7. REFERENCES

[1] L. De Raedt and J. Ramon. Condensed representations for inductive logic programming. In D. Dubois, C. A. Welty, and M. Williams, editors, *Proceedings of 9th International Conference on the Principles of Knowledge Representation and Reasoning*, pages 438–446. AAAI Press, 2004.

[2] T. Horvath, J. Ramon, and S. Wrobel. Frequent subgraph mining in outerplanar graphs. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, August 2006. To appear.

[3] T. Horvath, J. Ramon, and S. Wrobel. Frequent subgraph mining in outerplanar graphs. In *Workshop on Mining and Learning with Graphs (MLG 2006)*, Berlin, Germany, September 2006. Submitted.