CS698 Report:
A survey of algorithms for finding repeated subgraphs in a single graph.

Asish Ghoshal

December 21, 2013

Abstract

Subgraph mining is an important technique for knowledge discovery in complex data modeled by graphs. Since a lot of real world data can be modeled using graphs, subgraph mining techniques can be used to gain important insights into data. Therefore subgraph mining techniques find lots of use in areas like bioinformatics where it is used for studying protein-protein interaction networks, chemoinformatics where example applications include predicting properties from molecular structure, social network analysis where it is used for detecting communities among other things and a plethora of other areas. Traditionally, subgraph mining has been studied from the context of finding repeated subgraphs in a graph database while there has been relatively little work done in the area of finding repeated subgraphs in a single graph. In this report we present a brief survey of algorithms for finding repeated subgraphs in a single graph.

1 Introduction

Many real world datasets are naturally modeled using graphs e.g. social networks, the internet, the webgraph of the World Wide Web (WWW), protein-protein interaction networks among others. An important tasks in such domains is to gain important insights into the relationships among various entities by detecting recurring substructures. Subgraph mining deals with the problem of detecting repeated subgraphs in a single large graph or in a graph dataset.

Subgraph mining is an important problem because repeated substructures can be important predictors of some global properties of the graph. For instance in protein-protein interaction networks where vertices represent proteins and edges represent interactions between them, repeated substructures might correspond to protein complexes or functional modules [1] and understanding these interactions might improve our understanding of various diseases. Similarly, molecular structure of a given substance can be modeled as a undirected labeled graph where nodes correspond to atoms and edges correspond to the chemical bond between atoms. In such a network identifying repeated substructures can help in identifying toxic substances [2].

Subgraph mining has traditionally been formulated as finding frequent subgraphs in a graph dataset comprising of multiple graphs. In such a setting subgraphs whose occurrence count is more than a specified threshold are emitted. The occurrence count of a subgraph is defined as the number of graphs which have a subgraph isomorphic to the given subgraph. Some popular approaches in this area are *gSpan* [3], *FFSM* [4], *SUBDUE* [5] etc. However in some applications we might need to detect frequent subgraphs in a single graph like the internet or a single protein-protein interaction network or a social network. Detecting subgraphs in multiple graphs is slightly easier than in a single graph because algorithms operating on graph datasets take advantage of the downward closure property of the occurrence count. Downward closure property, which stipulates that if a graph is frequent then all of its subgraphs will also be frequent, is used to prune the search space [6]. When counting number of occurrences of a subgraph in a single graph two different counting methods are allowed: in the first case subgraphs are considered different if they do not overlap i.e. they have no common edge while in the second method two subgraphs are considered different as long as they overlap but have at least on distinct edge. The second method of counting doesn't give rise to the downward closure property. So most approaches of subgraph mining in a single graph do not allow for overlapping subgraphs.

One of the major challenges in subgraph mining is that algorithms need to solve instances of the subgraph isomorphism problem, which states that given two graphs $\mathbf{G}$ and $\mathbf{H}$ as input, and one must determine whether $\mathbf{G}$ contains a subgraph that is isomorphic to $\mathbf{H}$. Subgraph isomorphism is known to be *NP-complete* while its specialization graph isomorphism, in which we are interested in knowing if $\mathbf{G}$ is isomorphic to $\mathbf{H}$, is in *NP* but neither known to be in either *P* or *NP-complete* [7]. Thus, frequent subgraph mining is inherently hard. Thus, there are two classes of algorithms based on whether the subgraph isomorphism detection is exact or approximate.

In this report we compare different approaches for frequent subgraph detection in single graphs based on the size of the graph that the algorithm can handle, whether overlap between subgraphs is allowed or not, the application domain where the algorithms are used and finally the performance of algorithms. The rest of the document is organized as follows: we first present relevant definitions that are pertinent to the in-depth understanding of the frequent subgraph mining problem and various algorithms, then we summarize various algorithms for subgraph mining and provide a comparison of the various algorithms before finally concluding.

2 Definitions

Definition 1. Labelled Graph: *A labeled graph can be represented by a 4-tuple,* $G = (\mathbf{V}, \mathbf{E}, \mathbf{L}, \ell)$*, where* $\mathbf{V}$ *is the set of vertices,* $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ *which can either be directed or undirected,* $\mathbf{L}$ *is a set of labels and* $\ell : \mathbf{V} \cap \mathbf{E} \rightarrow \mathbf{L}$ *is a function assigning labels to the vertices and the edges.*

Definition 2. Graph Isomorphism*: An isomorphism is a bijective function* $\mathbf{f} : \mathbf{V}(G) \rightarrow \mathbf{V}(G')$ *such that* $\forall\, v \in \mathbf{V}(G)$, $\ell_G(v) = \ell_{G'}(f(v))$.

3 Frequent subgraph mining in a single graph

4 Conclusion

References

[1] J. Wang, M. Li, Y. Deng, and Y. Pan, "Recent advances in clustering methods for protein interaction networks," *BMC genomics*, vol. 11, no. Suppl 3, p. S10, 2010.

[2] A. Srinivasan, R. D. King, S. Muggleton, and M. J. Sternberg, "The predictive toxicology evaluation challenge," in *IJCAI (1)*, pp. 4–9, Citeseer, 1997.

[3] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pp. 721–724, IEEE, 2002.

[4] G. Cong, L. Yi, B. Liu, and K. Wang, "Discovering frequent substructures from hierarchical semi-structured data.," in *SDM*, 2002.

[5] L. B. Holder, D. J. Cook, and S. Djoko, "Substucture discovery in the subdue system.," in *KDD Workshop*, pp. 169–180, 1994.

[6] R. Zou and L. B. Holder, "Frequent subgraph mining on a single large graph using sampling techniques," in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pp. 171–178, ACM, 2010.

[7] M. R. Gary and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness.* WH Freeman and Company, New York, 1979.