What can we do with graph-structured data? - A data mining perspective -

Hiroshi Motoda*

Institute of Scientific and Industrial Research,
Osaka University
8-1, Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Recent advancement of data mining techniques has made it possible to mine from complex structured data. Since structure is represented by proper relations and a graph can easily represent relations, knowledge discovery from graph-structured data (graph mining) poses a general problem for mining from structured data. Some examples amenable to graph mining are finding functional components from their behavior, finding typical web browsing patterns, identifying typical substructures of chemical compounds, finding typical subsequences of DNA and discovering diagnostic rules from patient history records. These are based on finding some typicality from a vast amount of graph-structured data. What makes it typical depends on each domain and each task. Most often frequency which has a good property of anti-monotonicity is used to discover typical patterns. The problem of graph mining is that it faces with subgraph isomorphism which is known to be NP-complete. In this talk, I will introduce two contrasting approaches for extracting frequent subgraphs, one using heuristic search (GBI) and the other using complete search (AGM). Both uses canonical labelling to deal with subgraph isomorphism. GBI [6, 4] employs a notion of chunking, which recursively chunks two adjoining nodes, thus generating fairly large subgraphs at an early stage of search. It does not use the anti-monotonicity of frequency. The recent improved version extends it to employ pseudo-chunking which is called chunkingless chunking, enabling to extract overlapping subgraphs [5]. It can impose two kinds of constraints to accelerate search, one to include one or more of the designated subgraphs and the other to exclude all of the designated subgraphs. It has been extended to extract unordered trees from a graph data by placing a restriction on pseudo-chunking operations. GBI can further be used as a feature constructor in decision tree building [1]. AGM represents a graph by its adjacency matrix and employs an Apriori-like bottom up search algorithm using anti-monotonicity of frequency [2]. It can handle both connected and disconnected graphs. It has been extended to handle a tree data and a sequential data by incorporating to each a different bias in joining operators [3]. It has also been extended to incorporate taxonomy in labels to extract generalized subgraphs. I will show how both GBI and AGM with their extended versions can be applied to solve various data mining problems which are difficult to solve by other methods.

*Current address: AFOSR/AOARD, 7-23-17 Roppongi, Minato-ku, Tokyo 105-0032, Japan, e-mail: hiroshi.motoda@aoard.af.mil

References

1. W. Geamsakul, T. Yoshida, K. Ohara, H. Motoda, H. Yokoi, and K. Takabayashi. Constructing a decision tree for graph-structured data and its applications. *Fundamenta Informaticae*, 66(1-2):131–160, 2005.
2. A. Inokuchi, T. Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, 50(3):321–354, 2003.
3. A. Inokuchi, T. Washio, and H. Motoda. General framework for mining frequent subgraphs from labeled graphs. *Fundamenta Informaticae*, 66(1-2):53–82, 2005.
4. T. Matsuda, H. Motoda, and T. Washio. Graph-based induction and its applications. *Advanced Engineering Informatics*, 16(2):135–143, 2002.
5. P. C. Nguyen, K. Ohara, H. Motoda, and T. Washio. Cl-gbi: A novel approach for extracting typical patterns from graph-structured data. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 639–649, 2005.
6. K. Yoshida and H. Motoda. Clip : Concept learning from inference pattern. *Journal of Artificial Intelligence*, 75(1):63–92, 1995.