05 Efficient Knowledge Discovery from Large-scale Graph

～Efficient Structure Mining for Large-scale Graph～

Abstract

Recent advances in information science have shown that linked data pervade our society and the natural world around us. Graphs have become increasingly important for representing complicated structures such as Wikipedia, Freebase, and various social networks. However, existing algorithms cannot handle large graphs efficiently, so fast algorithms are needed. We introduce two fast algorithms. They outperform previous algorithms in terms of both speed and quality. Our algorithms are fundamental to many applications. Our algorithms allow many applications to be processed more efficiently and will help to improve the effectiveness of future applications.

● Efficient graph based labeling algorithm [1]

Overview Find high PageRank nodes from large-scale graphs efficiently and exactly

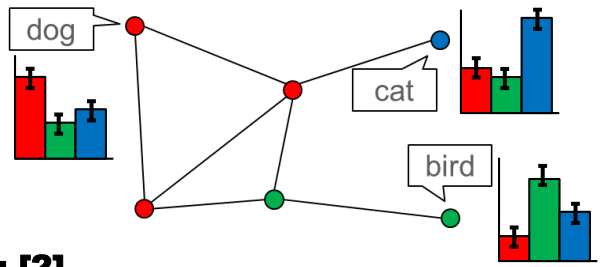
- Compute lower/upper labeling scores from the definition

$$F = \alpha SF + (1-\alpha)Y$$
$$= (1-\alpha)(Y + \alpha SY + \alpha^2 S^2 Y + \cdots)$$

F : nodes × sores matrix　α : constant parameter

S : adjacency matrix　Y : given score distribution

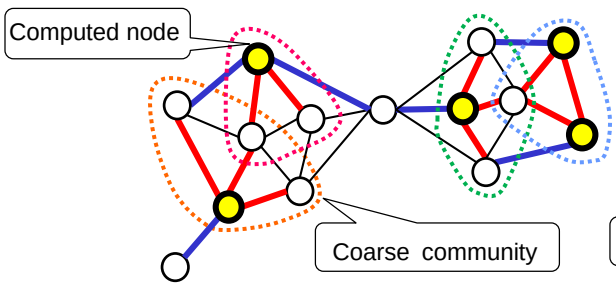
- Prune low score labels by using the bounds



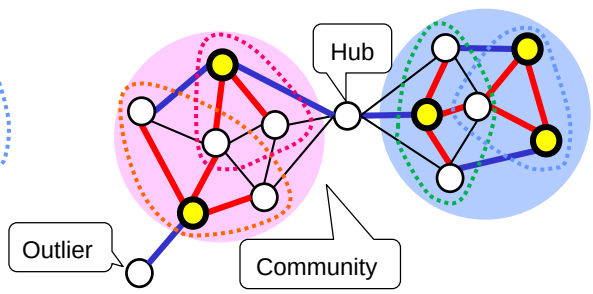
● Fast algorithm for graph clustering [2]

Overview Efficiently detect communities, hubs and outliers from large-scale graph data

- Find coarse communities by computing 2-hop away nodes



- Refine communities by finding nodes that are belong to multi-communities



Related works

[1] Y.Fujiwara,G.Irie, “Efficient Label Propagation”, In Proceedings of the 31th International Conference on Machine Learning (ICML 2014), 2014

[2] H.Shiokawa, Y.Fujiwara, M.Onizuka, “SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-scale Graphs”, In Proceedings of the VLDB Endowmen(PVLDB), Vol. 8, No. 11, 2015

Contact

Sotetsu Iwamura Distributed Computing Technology Project, NTT SIC

E-mail : iwamura.sotetsu(at)lab.ntt.co.jp

Machine Learning and Data Science Center