DOCTORAL THESIS

Interactive Search and Summarization on Hierarchical Graphs

ZHU, Xuliang

Date of Award:
2023

The hierarchical graph is a data structure to represent entities and general-certain relationships between entities, which has been widely used in real applications, such as ImageNet, disease ontology, Wikipedia categories, ACM computing classification system, and so on. Due to the massive terminologies and complex structures in a large hierarchical graph, it is challenging to resolve hierarchical graph analytics problems algorithmically, even with the help of leveraging human intelligence, such as object categorization, library classification, labeling, etc. This thesis focuses on two crucial problems of hierarchical graph analytics: data summarization and interactive search. Data summarization uses a small-sized answer to give a direct and human-friendly overview of the hierarchical graph data being analyzed. It is useful for understanding and visualization. Interactive search leverages human intelligence to categorize target labels in a hierarchy, which has applications in image classification, product categorization, and database search.

First, we study the data summarization on hierarchical tree. We motivate and formulate our kWTS-problem as selecting a diverse set of k nodes to summarize a hierarchical tree T with weighted terminologies. To depict diverse summarization and important vertices, we design a summary score function for capturing vertices' diversity coverage and structure correlation. To efficiently tackle it, we first propose an efficient greedy tree summarization algorithm GTS. It solves the problem with (1 − 1/e)-approximation guarantee. Although GTS achieves quality-guaranteed answers approximately, but it is still not optimal. To tackle the problem optimally, we further develop a dynamic programming algorithm OTS to obtain optimal answers for kWTS-problem in O(nhk^3) time, where n, h are the node size and height in tree T. The algorithm complexity and correctness of OTS are theoretically analysed. In addition, we propose a useful optimization technique of tree reduction to remove useless nodes with zero weights and shrink the tree into a smaller one, which ensures the efficiency acceleration of both GTS and OTS in real-world datasets.

Next, we study the data summarization on hierarchical DAGs. Similar to kWTS-problem, we propose a problem of finding k representative vertices to summarize a hierarchical DAG called kDAG-problem. The studied kDAG-problem is theoretically proven to be NP-hard. To efficiently tackle it, we propose a greedy algorithm with an approximation guarantee, which iteratively adds vertices with the large summary contributions into answers. To further improve answer quality, we propose a subtree extraction based method, which is proven to guarantee achieving higher-quality answers. In addition, we develop a scalable algorithm k-PCGS based on candidate pruning and DAG compression for large-scale hierarchical DAGs. We further study a new query-dependent summarization problem kQDAG-problem that could support query and summarize the subgraph under the query vertex. To tackle it, we propose an efficient algorithm k-PCGS+ based on the index of tight upper bound and threshold. We also prove the efficiency in theoretical time complexity analysis. Extensive experimental results on real-world datasets show the effectiveness and efficiency of our proposed algorithms on both hierarchical tree and DAG datasets.

Last, we study the interactive search on hierarchical graphs. Interactive search leverages human intelligence to categorize target labels in a hierarchy, which is useful for image classification, product categorization, and database search. However, many existing interactive graph search studies aim at identifying a single target optimally, and suffer from the limitations of asking too many questions and not being able to handle multiple targets. To address these two limitations, we study a new problem of budget constrained interactive graph search for multiple targets called kBM-IGS-problem. Specifically, given a set of multiple targets T in a hierarchy and two parameters k and b, the goal is to identify a k-sized set of selections S, such that the closeness between selections S and targets T is as small as possible, by asking at most a budget of b questions. We theoretically analyze the updating rules and design a penalty function to capture the closeness between selections and targets. To tackle the kBM-IGS-problem, we develop a novel framework to ask questions using the best vertex with the largest expected gain, which provides a balanced trade-off between target probability and benefit gain. Based on the kBM-IGS framework, we first propose an efficient algorithm STBIS to handle the SingleTarget problem, which is a special case of kBM-IGS. Then, we propose a dynamic programming based method kBM-DP to tackle the MultipleTargets problem. To further improve efficiency, we propose two heuristic but efficient algorithms, kBM-Topk and kBM-DP+. kBM-Topk develops a variant gain function and selects the top-k vertices independently. kBM-DP+ uses an upper bound of gains and prunes disqualified vertices to save computations. Experiments on large real-world datasets with ground-truth targets verify both the effectiveness and efficiency of our proposed algorithms.