Discovering Frequent Embedded Subtree Patterns from Large Databases of Unordered Labeled Trees

Yongqiao Xiao, SAS Inc., USA
Jenq-Foung Yao, Georgia College & State University, USA
Guizhen Yang, University at Buffalo, State University of New York, USA

ABSTRACT

Recent years have witnessed a surge of research interest in knowledge discovery from data domains with complex structures, such as trees and graphs. In this paper, we address the problem of mining maximal frequent embedded subtrees which is motivated by such important applications as mining “hot” spots of Web sites from Web usage logs and discovering significant “deep” structures from tree-like bioinformatic data. One major challenge arises due to the fact that embedded subtrees are no longer ordinary subtrees, but preserve only part of the ancestor-descendant relationships in the original trees. To solve the embedded subtree mining problem, in this article we propose a novel algorithm, called TreeGrow, which is optimized in two important respects. First, it obtains frequency counts of root-to-leaf paths through efficient compression of trees, thereby being able to quickly grow an embedded subtree pattern path by path instead of node by node. Second, candidate subtree generation is highly localized so as to avoid unnecessary computational overhead. Experimental results on benchmark synthetic data sets have shown that our algorithm can outperform unoptimized methods by up to 20 times.

Keywords: data mining; data mining algorithms; maximal frequent subtrees; tree mining

INTRODUCTION

Tree mining has been applied to Web usage mining, mining semi-structured data and bioinformatics, and so forth. The focus of this article is on mining maximal frequent embedded subtrees from unordered labeled trees. As a motivating example, consider mining the Web logs at a particular Web site. Several types of traversal patterns — for example, sequential patterns and maximal traversal patterns — have been proposed to analyze the browsing behavior of the user (Lan, Bressan, & Ooi,
follows. In the next section the tree mining problem is formally defined, followed by a description of the TreeGrow algorithm. We then report the experimental results, related work is described, and we conclude the article with future work.

**PROBLEM STATEMENT**

A tree is defined as $T = \langle N, B, r, L \rangle$, where $N$ is the set of nodes, $B$ is the set of branches (directed edges), $r \in N$ is the root of the tree, and $L$ is the set of labels on the nodes. For each branch, where $n_i, n_j \in N$, we call $n_i$ the parent of $n_j$, and $n_j$ a child of $n_i$. On each node $n_i \in N$, there is a label $l_i \in L$. The labels in a tree could be unique, or duplicate labels are allowed for different nodes. In both cases, the label of a node is given by a function, $l : N \rightarrow L$. Without loss of generality, the labels are represented by positive integers.

**Paths and Root Paths**

A path is a sequence of connected nodes $n_i, n_{i+1}, \ldots, n_k$, where $n_i, n_{i+1} \in B (1 \leq i < k)$, and $k$ is the number of nodes on the path. A node $u$ is called an ancestor of another node $v$ if there is a path from $u$ to $v$ in the tree. Correspondingly, $v$ is called a descendant of $u$. A path starting from the root node is called a root path. Each root path in a tree can be uniquely identified by the last node on the path in the tree. In Figure 1, node $n_3$ represents the root path $<n_1, n_2, n_3>$, and the labels on the path are $<1,2,3>$.

**v-path and v-node**

A v-path (virtual path) is a sequence of nodes $n_i, n_{i+1}, \ldots, n_k$, where $n_i$ is an ancestor of $n_{i+1} (1 \leq i < k)$, that is, two
Related Content

Discovery of Anomalous Windows through a Robust Nonparametric Multivariate Scan Statistic (RMSS)
[www.igi-global.com/article/discovery-anomalous-windows-through-robust/75614?camid=4v1a](www.igi-global.com/article/discovery-anomalous-windows-through-robust/75614?camid=4v1a)

Classification of Failures in Photovoltaic Systems using Data Mining Techniques
[www.igi-global.com/chapter/classification-of-failures-in-photovoltaic-systems-using-data-mining-techniques/150220?camid=4v1a](www.igi-global.com/chapter/classification-of-failures-in-photovoltaic-systems-using-data-mining-techniques/150220?camid=4v1a)
An Evolutionary Misclassification Cost Minimization Approach for Medical Diagnosis
www.igi-global.com/chapter/evolutionary-misclassification-cost-minimization-approach/25759?camid=4v1a

Adoption of Dual Iris and Periocular Recognition for Human Identification
www.igi-global.com/chapter/adoption-of-dual-iris-and-periocular-recognition-for-human-identification/139406?camid=4v1a