Available online at www.sciencedirect.com
SCIENCE DIRECT
Electronic Notes in Theoretical Computer Science
ELSEVIER Electronic Notes in Theoretical Computer Science 127 (2005) 77–87
www.elsevier.com/locate/entcs

The Gaston Tool for Frequent Subgraph Mining

Siegfried Nijssen and Joost N. Kok

LIACS, Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
snijssen@liacs.nl

Abstract

Given a database of graphs, structure mining algorithms search for all substructures that satisfy constraints such as minimum frequency, minimum confidence, minimum interest and maximum frequency. In order to make frequent subgraph mining more efficient, we propose to search with steps of increasing complexity. We present the GrAph/Sequence/Tree extractiON (GASTON) tool that implements this idea by searching first for frequent paths, then frequent free trees and finally cyclic graphs. We give results on large molecular databases.

Keywords: Frequent Subgraphs, Data Mining

1 Introduction

In recent years data mining of structures such as graphs, trees, molecules, XML documents and relational databases has attracted a lot of research. Especially the idea of discovering all frequent substructures has recently led to a large number of specialized algorithms for mining paths, trees and graphs. Frequent substructures give interesting information about the database. This information can be used in many different ways, for example for classification. As an example, Figure 1 shows molecular fragments that are frequent in an HIV inhibitor database, but not frequent in another database, thus providing features that distinguish between the two databases; the top leftmost fragment is a major constituent of AZT, a nucleoside reverse transcriptase inhibitor, and is used in many drugs for anti-HIV treatment. Our tool computes these fragments completely automatically.

1571-0661/$ – see front matter
doi:10.1016/j.entcs.2004.12.039

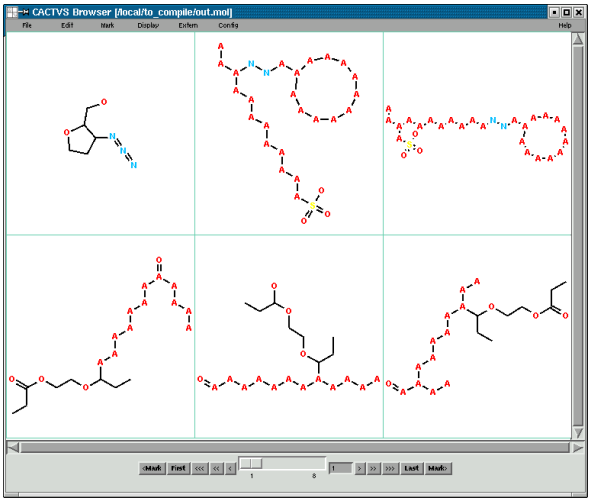

Fig. 1.

Experiments with molecular databases reveal that the largest numbers of frequent substructures in such databases are actually free trees. Free trees are much simpler structures than general, cyclic graphs, and efficient algorithms for them exist. Therefore, we integrate a frequent path, tree and graph miner into one tool called Gaston in order to gain efficiency. The main challenge in the development of the Gaston tool is how to split up the discovery process into several phases. Ideally, the tool should behave like a specialized free tree miner when faced with free tree databases, but should also be able to deal with graph databases efficiently.

Our homepage for frequent structure mining http://hms.liacs.nl/ contains further references, overview papers and the source code of our tool.

2 Foundations

We will only briefly discuss the mathematical preliminaries: the definitions are similar to those used in other papers concerning frequent structure mining, for example [2,13,14,15]. A labeled graph G consists of a finite set of nodes V, a set of edges $E \subseteq V \times V$ and a labeling function $\ell : V \cup E \to \mathcal{L}$ that assigns labels from $\mathcal{L}$ to all edges and nodes. We only consider undirected graphs, i.e., (v_1, v_2) is the same edge as (v_2, v_1). When graph G is subgraph isomorph with graph G', we denote this with $G \subseteq G'$.

We assume that a database $\mathcal{D}$ consists of a collection of graphs. The frequency of a graph G in $\mathcal{D}$ is defined by $\mathit{freq}(G, \mathcal{D}) = \#\{G' \in \mathcal{D} | G \subseteq G'\}$, the support of a graph is given by $\mathit{support}(G, \mathcal{D}) = \mathit{freq}(G, \mathcal{D})/|\mathcal{D}|$. The task