A Method of Improving the Efficiency of Mining Sub-structures in Molecular Structure Databases

Haibo Li[1], Yuanzhen Wang[1], and Kevin Lü[2]

[1] Department of Computing Science, Huazhong University of Science and Technology, Wuhan 430074, China
lihaibo.wh@gmail.com
[2] BBS, Room76, Tin Building, Brunel University, Uxbridge, UK UB8 3PH

Abstract. One problem exists in current substructure mining algorithms is that when the sizes of molecular structure databases increase, the costs in terms of both time and space increase to a level that normal PCs are not powerful enough to perform substructure data mining tasks. After examining a number of well known molecular structure databases, we found that there exist a large number of common loop substructures within molecular structure databases, and repeatedly mining these same substructures costs the system resources significantly. In this paper, we introduce a new method: (1) to treat these common loop substructures as some kinds of "atom" structures; (2) to maintain the links of the new "atom" structures with the rest of the molecular structures, and to reorganize the original molecular structures. Therefore we avoid repeat many same operations during mining process and produce less redundant results. We tested the method using four real molecular structure databases: AID2DA'99/CA, AID2DA'99/CM, AID2DA'99 and NCI'99. The results indicated that (1) the speed of substructure mining has been improved due to the reorganization; (2) the number of patterns obtained by mining has been reduced with less redundant information.

1 Introduction

There have been several efficient substructure mining algorithms by now, such as AGM [1], FSG [2], gSpan [5] and Gaston [4] etc. But as the size of the molecular structure database increasing, the costs in terms of both time and space are increasing so greatly that these algorithms can process no longer. For example, on normal PCs, all of the above algorithms can process the Predictive Toxicology database (PTE), which contains 340 molecular structures. But for the database consisted of 422 confirmed active compounds from AIDS antiviral screen database, only FSG, gSpan and Gaston can do mining. When aiming at the whole AIDS antiviral screen database, which contains 42689 compounds, only gSpan and Gaston can accomplish mining. Finally, for the whole NCI database which contains all 250,251 compounds, none of the above four algorithms can process substructure mining on PCs, excepting Gaston running on SMP servers.

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 180–184, 2007.

To solve this problem, we need to reduce redundant information obtained in mining progress besides to improve performance of substructure mining algorithm. After examining a number of well known molecular structure databases, we found that there are a large number of common loop substructures in molecular structures. We can reduce redundant information greatly in molecular structure databases. If we didn't break these loops, we can avoid perform many same operations during mining process.

The specific method introduced in this paper is: (1) we regard most common loops in molecular structures as some kinds of "atom" structure; (2) we consider common edges and vertexes between loops in condensed cyclic structures as some kinds of "bond" edges. Finally, we maintain the vertexes and edges which are not in any loops. According to these rules, we reorganize molecular structure databases to new ones.

After the reorganization, the number of candidate substructures generated during mining will be decreased, and most of these candidate substructures are tree structures and will spend less time to do graph isomorphism testing. The efficiency of mining will be improved greatly. Finally, we won't get many redundant frequent substructures in the mining result. The performance testing proves the conclusion.

The remaining of this paper is arranged as following. Section 2 takes some statistics and analysis on loops in chemistry molecular structure databases to confirm the regenerating method is feasible. Section 3 introduces the algorithm of reorganizing molecular structures based on atomizing of loops. In section 4, we give the performance testing result on various databases. And section 5 draws a conclusion.

2 Statistics on Loops in Molecular Structure Database

In this paper, we'll mine substructures in four molecular structure databases: AID2DA'99/CA, AID2DA'99/CI, AID2DA'99 and NCI'99 [4]. The sizes of these databases are list in Table 1. The loops' shapes are list in Table 2 to Table 5.

Table 1. Molecular structure databases to be mined

Database	Number of compounds
AID2DA'99/CA	422
AID2DA'99/CM	1081
AID2DA'99	42689
NCI	250251

Table 2. Loops in AID2DA'99/CA

Loops	Freq.	Cumulative Freq.
6-edge loops	76.3%	76.3%
5-edge loops	21.3%	97.6%
7-edge loops	1.2%	98.8%
3-edge loops	0.7%	99.5%
4-edge loops	0.06%	99.56

Table 3. Loops in AID2DA'99/CM

Loops	Freq.	Cumulative Freq.
6-edge loops	74.9%	74.9%
5-edge loops	22.6%	97.5%
7-edge loops	1.2%	98.7%
4-edge loops	0.5%	99.2%
3-edge loops	0.2%	99.4%