

基于频繁模式树的频繁连通闭图集挖掘算法

刘 振, 杨路明, 彭佳扬

(中南大学 信息科学与工程学院, 湖南 长沙 410083)

摘 要:随着频繁模式挖掘的深入研究,图模型被广泛地应用于为各种事务建模,因此图挖掘的研究显得越来越重要。文中针对唯一标识的有向连通图模型,基于频繁模式树结构,改进了频繁模式增长算法挖掘频繁连通闭子图。使用生物代谢路径数据集的实验证明,这种算法能有效地挖掘出唯一标识的有向连通图集集中的频繁闭图集,一次运算可以挖掘出多个阈值的最大频繁子图集。这种算法适用于以唯一标识的有向连通图建模的网络或图集,可以应用到基于图简化模型的生物网络的子图挖掘任务中。

关键词:子图挖掘; 频繁模式树; 频繁模式增长; 频繁闭图集; 生物网络

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2009)05-0037-04

An Algorithm for Mining Connected Closed Frequent Subgraphs Based on FP-Tree

LIU Zhen, YANG Lu-ming, PENG Jia-yang

(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: With the deep study of the frequent pattern mining, graphs can be modeled for many transactions widely, and the study of graphs have become increasingly important. Based on FP-Tree, presents an improved FP-Growth algorithm, which can find the closed frequent connected subgraph from the model of unique labeled directed connected graph set. The experiment of biology metabolize pathway dataset demonstrated that the algorithm can get the closed frequent subgraph set effectively, and can get the max frequent subgraph sets of many different threshold by execute once. This algorithm can use for mining the network or graph set which can modeling by unique labeled, directed, connected graph. It can be applied to the subgraph mining in biological networks which is based on the simplification model.

Key words: subgraph mining; FP-Tree; FP-Growth; closed frequent subgraph set; biological networks

0 引 言

频繁模式的挖掘首先是由 Agrawal 等人提出来的^[1],在许多数据挖掘应用中有着重要作用和地位。由于图模型具有高度的一般性,可以方便地表达各种事务的状态,以及事务与事务之间的相互作用或联系,并且可以广泛地应用到化学信息学、生物信息学^[2]和 Web 分析等领域中。因此,频繁子图挖掘在频繁模式挖掘中有着重要的地位^[3]。频繁子图挖掘问题最早由 Inokuchi 等人提出^[4]。2001 年,Inokuchi 等人提出了 AGM 算法^[5],采用邻接矩阵模型,基于点增长进行图挖掘。Kuramochi 和 Karypis 提出的 FSG 算法^[6]采用

稀疏邻接列表数据结构,基于边增长的方式进行图挖掘。2002 年 Yan 和 Han 提出的 gSpan 算法^[7]采用 DFS(深度优先搜索)词典排序法和最小 DFS 代码这两种技术来实现频繁子图挖掘。王艳辉等人对上述算法进行了分析和研究^[8]。李玉华和唐德权等人分别对 FSG 和 FFSM 算法进行了研究^[9,10]。2004 年 Koyutürk 等人提出的 MULE 算法^[11]将生物网络中的顶点进行合并,对顶点进行唯一标识,采用基于 Apriori 思想的算法,把频繁项目集挖掘的基本思路应用到频繁连通图集挖掘中。2007 年李先通等人对 gSpan 算法深入研究后提出了新的 GraphGen 算法^[12],将子图同构问题转换为子树同构问题。

文中利用 Koyutürk 等人提出的对顶点唯一标识方法和 Han 等人提出的 FP-Tree 结构^[13],改进了 FP-Growth 算法,实现出了一种频繁连通闭图集挖掘算法。实验表明,文中的算法能有效地解决对顶点唯一标识图的频繁子图挖掘问题,并能应用到 Koyutürk 等

收稿日期:2008-08-19

基金项目:湖南省科学研究项目(08b040)

作者简介:刘 振(1983-),男,湖南长沙人,硕士研究生,研究方向为数据挖掘;杨路明,教授,博士生导师,主要研究领域为生物信息学、计算机理论、网络安全、计算机图形理论与算法。

人提出的生物网络的子图挖掘问题中,可以得到比 MULE 算法更多的结果。

1 相关知识和问题描述

图是一种可以用来表示实体集之间联系的数据结构。图由顶点集 V 和连接顶点对的边集 E 构成。每条边用顶点对 (v_i, v_j) 表示,其中 $v_i, v_j \in V$ 。可以给每个顶点 v_i 赋予一个标识 $l(v_i)$,代表实体的名字。类似地,每条边 (v_i, v_j) 也可以关联到一个标识 $l(v_i, v_j)$,描述实体对之间的联系。

定义 1 图 $G' = (V', E')$ 是另一个图 $G = (V, E)$ 的子图,如果它的顶点集 V' 是 V 的子集,并且它的边集 E' 是 E 的子集。子图关系记作 $G' \in sG$ 。

定义 2 给定一个图的集族 \sum ,子图 g 的支持度的定义为包含它的所有图所占的百分比,即

$$s(g) = \frac{|\{G_i \mid g \subseteq sG_i, G_i \in \sum\}|}{|\sum|}$$

定义 3 给定图的集合 \sum 和支持度阈值 minsup ,频繁子图挖掘的定义是找出所有使得 $s(g) \geq \text{minsup}$ 的子图 g 。

文中的讨论主要关注顶点唯一标识的有向连通图。这种图的描述如下:

(1) 一个图是连通的,如果图中每对顶点之间都存在一条路径;其中,路径是顶点的序列 $\langle v_1 v_2 \dots v_k \rangle$,使得序列中每对相邻的顶点 (v_i, v_{i+1}) 之间都有一条边。

(2) 一个图是有向的,如果它只包含有向边。

(3) 一个图是顶点唯一标识的,如果同一个图中所有的顶点的标识各不相同。

在子图挖掘实践中,完全频繁子图集中的频繁子图数量可能非常大。因此,只推导出较小的、具有代表性的结果集是有用的。下面给出频繁闭图集的定义。

定义 4 图集 X 是闭图集(Closed Subgraph Set),如果 X 中的任何一个子图的真超图都不具有和它相同的支持度计数。

定义 5 一个子图集是频繁闭图集(Closed Frequent Subgraph Set),如果它是闭的,并且它的支持度大于或等于最小支持阈值。

Koyutürk 等人开发的 MULE 算法得到的是最大频繁子图集,文中提出的算法求出的是频繁闭图集。

2 算法

FP-Growth 算法只进行两次数据库扫描,不使用候选集,减少了计算时空开销。尽管 FP-Growth 算

法有这些优点,但要将其运用到挖掘频繁连通闭图集问题也仍有一些问题需要解决:

① FP-Growth 算法用于频繁项集挖掘,不能直接用于频繁子图挖掘,且其挖掘出的频繁项集不是频繁闭项集;

② FP-Growth 算法通过一棵 FP-Tree 挖掘出频繁边集,无法保证结果图的连通性;

③ FP-Growth 算法要对单个路径 P 中的所有组合进行频繁判断,对于挖掘长路径会付出较大的时空代价。

文中的算法利用顶点唯一标识方法和 FP-Tree 结构,对 FP-Growth 算法做相应的改进,将提供频繁边集的数据压缩到一颗频繁模式树,然后将这种压缩后的数据库分组组成一组条件数据库,从中挖掘出以频繁连通边集形式来表示的频繁连通子图,从而解决挖掘频繁连通子图的问题。并对其运算结果做了进一步处理。具体算法如下:

算法 1

输入:图数据库 D ;最小频繁度阈值 ϵ 。

输出:频繁连通闭图集 S 。

(1) 调用 $\text{createFPtree}(D, \epsilon, t)$, 构造 FP 树 t ;

(2) 调用 $\text{miner}(t, \text{null})$ 挖掘 FP 树,生成初步挖掘结果集 I ;

(3) for each 频繁边集 γ_i 在结果集 I 中 {

(4) 调用 $\text{exchange}(\gamma_i, S)$; }

(5) 删除结果集 S 中重复的结果项;

(6) 将结果集 S 中没有的 1-频繁项插入到结果集 S 中;

其中,构造 FP-Tree 的算法参看文献[13],这里不再赘述。

由于在数据挖掘过程中,可能会存在较长的模式,而对于这种长模式的情况,FP-Growth 算法尝试求出其所有可能的组合。对于一个长度为 n 的模式,其所有可能的组合数为 $2^n - 1$ 。文中的算法引入了一个新的概念高频闭组合 β : 设所有组合形成的集合为 A ,所有 β 形成的集合为 $B, B \subseteq A$; 对于某个组合 $a \in A$,且在集合 A 中不存在一种组合 b ,使得 $a \subset b$ 且 a 和 b 的频繁度相同,则 $a \in B$ 。

例: AS, 19; SS, 17; SA, 17; AG, 15 为一条路径 P , 如图 1 所示, 频繁度阈值为 15 的高频闭组合为 $\{AS, SS, SA, AG: 15\}$, $\{AS, SS, SA: 17\}$, $\{AS: 19\}$ 。其中 AS, SS, SA, AG 均为图中的边标识。如此求出的组合数为 3, 而所有可能的组合数是 15。这样,即使是针对较长的单条路径 P 进行频繁判断,效率也会很高。

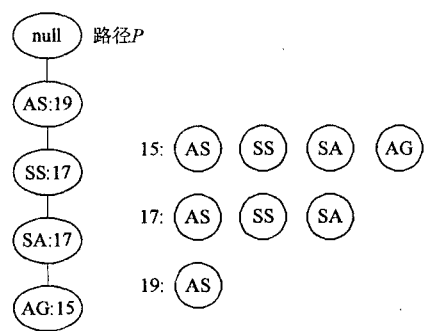


图 1 路径 P 和它的三个高频闭组合

初步挖掘结果集算法如算法 2 所描述:

算法 2 $\text{miner}(t, \text{null})$

输入:FP 树 t ;模式 α

输出:生成初步挖掘结果集 I

- (1) if t 含单个路径 P then
- (2) for 路径 P 中节点的每个高频闭组合(记作 β)
- (3) 产生模式 $\beta \cup \alpha$,其 $\text{support} = \beta$ 中节点的最小频繁度;
- (4) else for each α_i 在 t 的头部{
- (5) 产生一个模式 $\beta = \alpha_i \cup \alpha$,其 $\text{support} = \alpha_i$.

support;

(6) 构造 β 的条件模式基,然后构造 β 的条件 FP-树 Tree_β ;

(7) if $\text{Tree}_\beta \neq \emptyset$ then

(8) 调用 $\text{miner}(\text{Tree}_\beta, \beta)$;

获得初步挖掘结果集 I 之后,需要对 I 进行进一步处理。主要工作包括:

- ① 将挖掘算法前期产生的结果集(频繁边集)转换成频繁连通子图集;
- ② 删除结果集中重复的结果项;
- ③ 补充 1-频繁子图到结果集中。

算法 3 描述了将频繁子图集转换成频繁连通子图集的算法:

算法 3 $\text{exchange}(\gamma_i)$

输入:频繁边集 γ_i 。

输出:频繁连通闭图集 S 。

- (1) 频繁子图 $\gamma_i = \{e_1, e_2, e_3, \dots, e_n\}, B = \{e_i\}, C = \gamma_i - B, e_i \in \gamma_i$;
- (2) for each $e_j, e_j \in C$ {
- (3) if C 中有边 e_x 与 e_i 连通, $e_i \in B$ then
- (4) $B = B + \{e_x\}, C = \gamma_i - B$;
- (5) else{
- (6) $B \rightarrow S$;
- (7) if $C! = \emptyset$ then
- (8) 对 C 调用 $\text{Exchange}()$; }

将频繁子图集转换成连通频繁子图集的过程中,可能会产生一些重复的连通边集,需要将其删除。由于文中算法是基于 FP-Growth 算法的,某些 1-频繁边可能不在结果集中,因此需要将这些频繁边插入到结果集中。

下面使用真实的生物网络数据进行实验,来验证文中的算法。

3 实验

从 KEGG 代谢路径数据库中选取了生物的不同规模的代谢路径进行了多次实验。实验证明,文中的算法能高效地找到符合条件的频繁模式。KEGG 代谢路径数据库中的数据模型见文献[11]和[14]的介绍。实验数据如表 1 所示。

表 1 选用不同的频繁阈值挖掘不同规模的代谢路径结果比较(丙氨酸-天门冬氨酸盐)

实验集大小 (点,边)	频繁 阈值	1-频繁 项数	频繁子图数 (MULE 算法)	频繁子图数 (文中的算法)	文中算法 的时间(s)
40(692,2123)	6	27	14	55	1.141
	13	10	6	10	0.485
	15	6	3	5	0.469
55(952,2978)	8	32	22	87	2.062
	15	16	11	19	0.703
	18	9	5	8	0.64
70(1227,3845)	9	39	17	117	3.86
	15	23	19	54	1.203
	20	18	10	18	0.844

从实验结果可以看出,文中的算法可以在很短的时间内求出比最大频繁子图集更多数量的频繁子图。这是因为,① 文中算法利用 Koyutürk 等人提出的对顶点唯一标识方法不会产生子图同构,大大简化了图挖掘问题;② 利用 FP-Tree 结构并对 FP-Growth 算法做相应的改进,保留了 FP-Tree 结构和 FP-Growth 算法的优点;③ 利用引入高频闭组合的概念,运算得到的结果是频繁连通闭图,不仅可以提高运算效率,还可以得到比最大频繁连通子图集更多的频繁连通子图。以某次运算为例:

例:对于天门实验集 40 进行结果检验,当频繁度阈值为 13 时,最大频繁连通子图集为 6 个频繁子图 (SS, SG, GH, GS), (AS, SA, SS), (AB, AG), (GH, AG), (DG), (AB2), 见图 2, 其中 SS, SG, GH, AS, AG 等均为图的边标识,而图中 aG, pA, aS, pB 等均为图的顶点标识,边 SG 表示由顶点 aS 指向顶点 aG 形成的边 $\langle aS, aG \rangle$ 。后面的描述都是以此为例。

文中算法得到的频繁连通闭图集为 (13:SS,

$SG, GS, GH), (13: DG), (13: AB2), (14: AB, AG), (15: AS, SA, SS), (15: GH, AG), (16: GH), (17: AB), (17: AG), (19: SS)$, 见图 3。

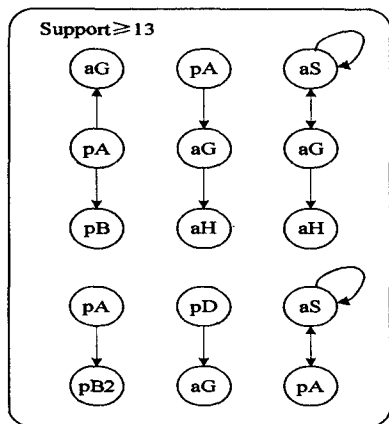


图 2 最大频繁连通子图集(阈值 13)

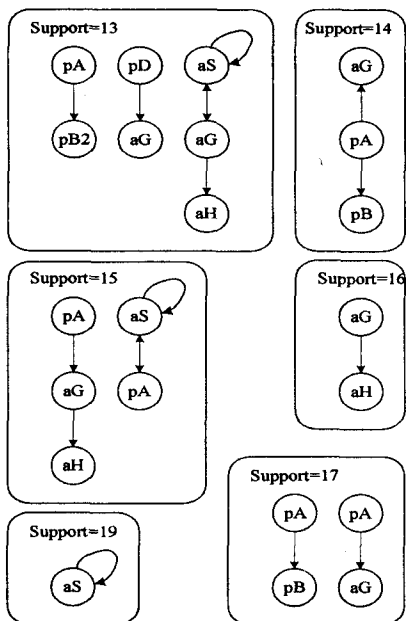


图 3 文中算法的结果集(阈值 13)

当频繁度阈值为 15 时,最大频繁连通子图集为 $(AS, SA, SS), (GH, AG), (AB)$, 见图 4; 文中算法得到的频繁连通闭图集为 $(15: AS, SA, SS), (15: GH, AG), (16: GH), (17: AB), (17: AG), (19: SS)$, 见图 5。

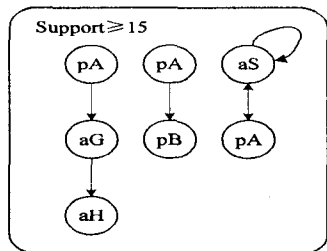


图 4 最大频繁连通子图集(阈值 15)

从图中不难看出,文中算法求出的给定阈值的频

繁连通闭图集,不仅包括了最大频繁子图,而且求出了更多的大于给定阈值的频繁项,还给出了输出的最大频繁项的阈值大小。比较图 3 和图 5 不难发现,图 5 的结果在图 3 中都存在,图 5 的结果集是图 2 结果集的子集。算法在阈值为 15 时得到的结果就是阈值为 13 时得到的结果中所有频繁度大于等于 15 的频繁项。比较图 2 和图 4 可以发现最大频繁连通子图集不存在这种特点。由以上比较可知,只求最大频繁连通子图集的算法一次运算只能求出关于一个阈值的最大频繁项集,而文中的算法不仅求出了关于该阈值的最大频繁项集,并且求出了大于此阈值的所有阈值的最大频繁连通子图集,同时给出了所有频繁子图的频繁度。

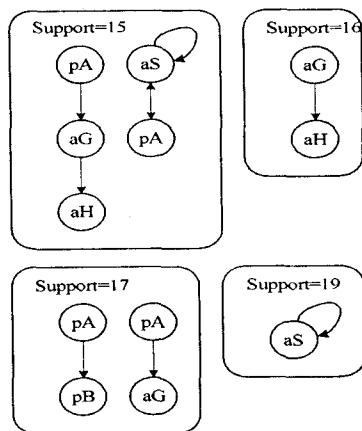


图 5 文中算法的结果集(阈值 15)

4 结束语

提出了一种 FP-Growth 算法的改进算法,用于挖掘顶点唯一标识的有向连通图模型中的频繁连通闭图集,并能应用到 Koyutürk 等人提出的生物网络的子图挖掘问题。通过引入高频闭组合的概念,该算法在 FP-Growth 算法的基础上提高了效率。通过生物实验数据发现,相对于 MULE 算法,文中的算法能在有效的时间内得到更多有意义的结果。

参考文献:

- [1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]//In: Proceedings of the ACM SIGMOD International Conference Management of Date. Washington: [s. n.], 1993: 207-216.
- [2] 杨炳儒,胡健,宋威.生物信息数据挖掘技术的典型应用[J].计算机工程与应用,2007,43(2):18-19.
- [3] Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions[J]. Data Mining Knowl Discov, 2007, 15(1): 55-86.

(下转第 44 页)

表 1 返回结果

本系统部分完成输出结果	1	中国电信	www.chinatelecom.com.cn	7
	2	中国海洋石油公司	www.cncooc.com.cn	6
	3	网易	www.163.com	7
	4	微软中国有限公司	www.microsoft.com/china/	7
	5	海尔集团	www.haier.com	5
	6	金山在线	www.kingsoft.com	6
	7	瑞星反病毒查询网	www.rising.com.cn	6
	8	中国上市公司查询网	www.cnlist.com	4
	9	中国移动公司	www.chinamobile.com	7
	10	中国国际股份有限公司	www.airchina.com.cn	6
百度输出结果	1	IT 公司查询网	www.seeitoo.com	4
	2	全国 IT 公司速查手册	www.iteer.net/modules/xdirectory/	5
	3	中国 IT 公司猎头公司	www.hrm.net.cn	5
	4	互联网实验室	www.chinalabs.com/lisi	2
	5	IT 公司速查手册	www.newwww.com	2
	6	IT 世界网	It.com.cn	5
	7	IT 公司	www.chinassb.com/nv/nv2501	0
	8	翻译公司	www.translation-company.net.cn/gzff.htm	0
	9	3691 网	www.3691.com/computer/105.htm	0
	10	IT 公司	www.texindex.com.cn/company/netail	0

表 2 搜索关键字列表

搜索关键字	IT 公司	数据挖掘	MFC 基类	CAR
编号	1	2	3	4

实验中关键技术主要使用了 VC++ 6.0 中几个 MFC 类: HTTP 类、ChttpConnection 类、ChttpFile 类等。HTTP 类是在查询模块中应用, 利用 ChttpConnection 类中的方法 OpenRequest 发送一个 GET 命令, 此时 ChttpFile 类被创建, 然后利用 ChttpFile 类的几个成员函数就可以和搜索引擎响应了。为了获得 Url 信息,

可以利用 CString 类的成员函数 find 读取文件中的字符串。最后就可生成网络图, 主要包括两张表一是节点表, 另一是有向边表。

4 结束语

在分析网络结构的基础上, 介绍了网络结构挖掘中 HITS 算法和 PageRank 算法模型, 并针对其弱点提出了改进方案。主要是对每个网页定义这三个参数: PageRank、Authority、Hub, 并进行分析与优化, 为此建立一个原型系统来验证所提出的解决方案的有效性, 为改进算法提供实践证明, 在今后的工作中, 可以考虑根据这些思想进一步改进算法。

参考文献:

- [1] Kosala R, Blockeel H. Web Mining Research: A Survey[J]. SIGKDD Explorations, 2000, 2(1): 1-15.
- [2] 杨 鲲, 孟 波. 一种基于 XML 的 Web 数据挖掘方法[J]. 计算机应用, 2003, 23(6): 160-164.
- [3] 陈 莉, 焦李成. Internet/Web 数据挖掘研究现状及最新进展[J]. 西安电子科技大学学报, 2001, 28(1): 114-119.
- [4] 邓 英, 李 明. Web 数据挖掘技术及工具研究[J]. 计算机工程与应用, 2001(20): 92-94.
- [5] 冉 丽, 何毅舟, 许龙飞. 基于 Web 结构挖掘的搜索引擎作弊检测方法[J]. 计算机应用, 2004(10): 160-162.
- [6] 宋建康, 张礼平. Web 结构挖掘算法探讨[J]. 华东理工大学学报, 2003(10): 537-540.
- [7] 刘丽珍. 网络结构挖掘的关键分析[J]. 计算机应用研究, 2003(5): 116-118.
- [8] Spertus E. Parasite: Mining Structural Information on the Web[C]// The Sixth International WWW Conference. Santa Clara, USA: [s. n.], 1997.

(上接第 40 页)

- [4] Tan Pang-Ning, Steinbach M, Kumar V. 数据挖掘导论[M]. 范 明, 等译. 北京: 人民邮电出版社, 2006.
- [5] Inokuchi A, Washio T, Okada T, et al. Applying the apriori-based graph mining method to mutagenesis data analysis[J]. Comput Aided Chem, 2001(2): 87-92.
- [6] Kuramochi M, Karypis G. Frequent subgraph discovery[C]// In: IEEE Intl Conf Data Mining. California, USA: [s. n.], 2001: 313-320.
- [7] Yan X, Han J. gSpan: Graph-based substructure pattern mining[C]// In: IEEE Intl. Conf. Data Mining (ICDM'02). Maebashi City, Japan: [s. n.], 2002: 721-724.
- [8] 王艳辉, 吴 斌, 王 柏. 频繁子图挖掘算法综述[J]. 计算机科学, 2005, 32(10): 193-196.
- [9] 李玉华, 罗汉果, 孙小林. 一种基于 Apriori 思想的频繁子图

- 发现算法[J]. 计算机工程与科学, 2007, 29(4): 84-87.
- [10] 唐德权, 夏幼明, 张丽英. 基于图的数据挖掘算法研究[J]. 云南师范大学学报, 2007, 27(5): 30-34.
- [11] Koyutürk M, Grama A, Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks[J]. Bioinformatics, 2004, 20(1): 200-207.
- [12] 李先通, 李建中, 高 宏. 一种高效频繁子图挖掘算法[J]. 软件学报, 2007, 18(10): 2469-2480.
- [13] Han J W, Pei J, Yin Y W. Mining frequent patterns without candidate generation[C]// In: ACM-SIGMOD. Dallas, TX, New York: ACM Press, 2000: 1-12.
- [14] Krishnamurthy L, Nadeau J, Özsoyoğlu G, et al. Pathways database system: an integrated system for biological pathways[J]. Bioinformatics, 2003, 19(8): 930-937.