基于最大 Code 码的极大完全子图算法

郭 平[1] 康艳荣[1] 史晓晨[2]

(重庆大学计算机学院 重庆 400044)[1] (东北大学机械工程与自动化学院 沈阳 110004)[2]

摘 要 本文通过引入极大 code 码,提出了一种寻找图的极大完全子图的算法 FMCSG,该算法用邻接矩阵表示图。在寻找极大完全子图时根据得到的 code 码及时剪掉非极大 code 码的子矩阵,从而减少对矩阵的遍历次数,提高了算法的效率。

关键词 极大完全子图,极大 code 码,图表示

An Algorithm Find Maximal Complete-subgraph by Max-code

GUO Ping[1] KANG Yan-Rong[1] SHI Xiao-Chen[2]

(School of Computer Science, Chongqing University, Chongqing 400044)[1]

(School of Mechanical Engineering and Automation, Northeastern University, Chongqing 110004)[2]

Abstract In this paper, we propose a new algorithm, called FMCSG, to find the Maximal Complete-Subgraph in graph by Max-code. The graph is represented as adjacency matrix in FMCSG. By this way, it can prune the corresponding matrix of non-Max-code for the purpose of reducing the search space and improving the efficiency of the algorithm.

Keywords Maxical complete-subgraph, Maxcode, Represent graph

1 引言

极大完全子图有着十分广泛的应用。在管理决策方面,一些管理事务问题如人员管理、运输调度等均可抽象为求解极大完全子图问题。在数据挖掘方面,关联规则的挖掘是数据挖掘研究的重要内容之一,利用极大完全子图来寻找频繁项集,可以在很大程度上减少对数据库的访问,提高 CPU 的利用率[2,3]。

本论文对极大完全子图问题的研究基于最大完全子图的研究成果。目前,对最大完全子图问题的研究已经取得了很大的成就,总体上分为两类:一类是求解最大完全子图的确定性算法,如上世纪 80 年代提出的深度优先列举算法[4]。另一类是求解最大完全子图的启发式方法,如顺序贪婪启发式算法[5]。本文的研究是基于顺序贪婪启发式算法的基本思想,融入了局部搜索方法以及 Washio Takashi, Kunio Nishinura 等人提出的 code 码[1]概念,通过引入图的邻接矩阵的极大 code 码,提出了一种新的寻找图的极大完全子图的算法——FMCSG(Finding Maximal Complete-Subgraph)。该算法用邻接矩阵表示图,通过寻找拥有极大 code 码的子矩阵来实现在图中对极大完全子图的搜索。搜索过程中由于及时剪掉了非极大 code 码的子矩阵,有效地减少对子矩阵集的扫描次数,提高了算法的效率。

2 图的极大 code 码表示

2.1 图的 code 码

邻接矩阵是图表示方法的一种常用方法[7]。本文用到的邻接矩阵定义如下:

定义 1 设 $G=(V,E)$ 是一个无向图,则 G 的邻接矩阵记为 $A=(a_{ij})_{n\times n}$,其中

$$a_{ij}=\begin{cases}1 & \text{若 } v_i \text{ 和 } v_j \text{ 之间有边相连且 } i\neq j\\ 0 & \text{若 } v_i \text{ 和 } v_j \text{ 之间无边相连或 } i=j\end{cases}$$

定义 2 设 $H=(V',E')$, $V'\subset V$, $E'\subset E$。如果 $\forall x,y\in V'$, H 中都有连接 x 与 y 的边,则称 H 是 G 的完全子图。如果不存在 G 的完全子图 M,使得 $V(H)\subset V(M)$,则称 H 为 G 的极大完全子图。

显然,图 G 的极大完全子图可以不止一个。在给出极大邻接矩阵的概念之前,我们先引入如下概念:

定义 3 设图 G 的邻接矩阵 X 为:

$$X(v_1,v_2,\cdots,v_n)=\begin{array}{c}\\ v_1\\ v_2\\ v_3\\ \vdots\\ v_n\end{array}\begin{pmatrix} 0 & x_{1,2} & x_{1,3} & \cdots & x_{1,n}\\ x_{2,1} & 0 & x_{2,3} & \cdots & x_{2,n}\\ x_{3,1} & x_{3,2} & 0 & \cdots & x_{3,n}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ x_{n,1} & x_{n,2} & x_{n,3}\cdots & & 0\end{pmatrix}$$

称 $\text{code}(X)=x_{1,2}x_{1,3}x_{2,3}x_{1,4}\cdots x_{n-2,n}x_{n-1,n}$ 为邻接矩阵 X 的 code 码。

显然,同一个图可以用不同的邻接矩阵表示,因此其 code 码也不同。可以证明:

性质 1 假设图 G 为有 k 个顶点的完全图,X 为 G 的邻接矩阵。则 X 的 code 码中仅含字符"1"且"1"的个数为 $k(k-1)/2$。

定义 4 设 $H=(V',E')$ 是图 $G=(V,E)$ 的顶点个数为 k 的极大完全子图。如果 G 的邻接矩阵 X 的左上角 k 阶子矩阵与 H 的邻接矩阵相同,称 $code(X)$ 为图 G 的极大 code 码(Max-Code)。

由于图 G 的极大完全子图不止一个,因此 G 的极大 code 码也不止一个。记 G 的极大 code 码的集合为 $MCode(G)$。

定义 5 设 X 是图 G 的邻接矩阵。若 $code(X)\in Mcode$

郭 平 副教授,主要研究方向:人工智能、数据挖掘;康艳荣 硕士研究生,主要研究方向:数据挖掘。

(G)，称 X 是 G 的极大邻接矩阵。记 G 的极大邻接矩阵集合为 $Mmatrix(G)$。

定理 1 假设 H 是图 G 的有 k 个顶点的极大完全子图，且 $X\in Mmatrix(G)$。H 的邻接矩阵与 X 的左上角 k 阶子矩阵相同的充分必要条件是 $r_1\leqslant r<r_2$。这里 $r_1=k(k-1)/2$，$r_2=k(k+1)/2$，r 是 X 从第一位开始的连续“1”的个数。

2.2 极大 code 码及其性质

定义 6 设 g_1 和 g_2 是图 G 的子图，它们都有 k 个顶点且其中有 $k-1$ 个顶点相同。

$$X_k(v_1,v_2,\cdots,v_{k-1},v')=\begin{pmatrix}X_{k-1} & x_{1,k}\\ x_{k,1} & 0\end{pmatrix},Y_k(v_1,v_2,\cdots,v_{k-1},v'')=\begin{pmatrix}Y_{k-1} & y_{1,k}\\ y_{k,1} & 0\end{pmatrix}$$

分别是 g_1 和 g_2 的邻接矩阵。如果 $code(X_k)\geqslant code(Y_k)$，称 $k+1$ 阶矩阵：

$$\begin{pmatrix}X_{k-1} & x_{1,k} & y_{1,k}\\ x_{k,1} & 0 & z_{k,k+1}\\ y_{k,1} & z_{k+1,k} & 0\end{pmatrix}$$

为矩阵 X_k，Y_k 的连接，记为 $X_k\infty Y_k$。其中，

$$z_{k+1,k}=z_{k,k+1}=\begin{cases}1 & (v',v'')\in G\\ 0 & (v',v'')\notin G\end{cases}$$

例如，图 1 中的子图 g_1 和 g_2 对应的邻接矩阵分别是 X_4 和 Y_4。连接 X_4 和 Y_4 得到的 5 阶矩阵对应的图可以是 g_3 或 g_4，如图 2。若 G 中 v_4 和 v_5 之间有边相连，则 X_4 和 Y_4 连接得到的 5 阶矩阵对应的图是 g_3，否则得到 g_4。

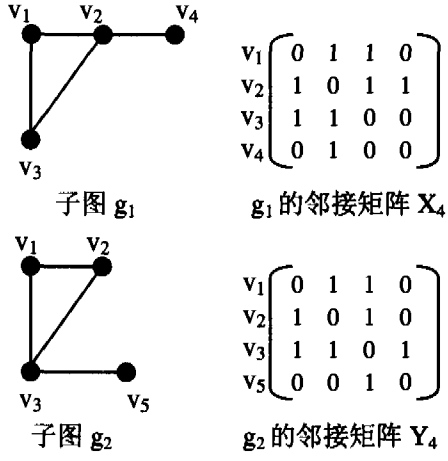

图 1 子图 g_1 和 g_2 及其邻接矩

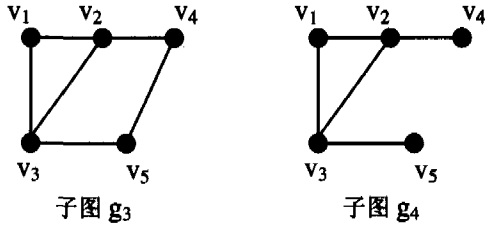

图 2 图的连接

定义 7 两个 k 阶邻接矩阵进行矩阵连接生成的 $k+1$ 阶矩阵所对应的 code 码称为 $k+1$ 阶候选 code 码，记它们的集合为 $Ccode(k+1)$。在 $Ccode(k+1)$ 中，全 1 的 code 码称为 $k+1$ 阶极大子 code 码，记它们的集合为 $Scode(k+1)$。

由极大子 code 码和候选码的概念，我们可得到如下性质：

性质 2 若图 G 的 $Ccode(k+1)$ 中没有极大子 code 码，那么 $Scode(k)$ 中每个元素对应的 k 阶邻接矩阵是图 G 的极大完全子图对应的邻接矩阵。

证明：对 $Scode(k)$ 中的任一元素，设它对应的 k 阶邻接矩阵所对应的图为 H。根据 $Scode(k)$ 的定义，H 是 G 的完全子图。我们来证明 H 还是 G 的极大完全子图。

若 H 不是 G 的极大完全子图，则存在 G 的极大完全子图 M，使得：

$H\subset M$ 且 $|V(H)|<|V(M)|$

从 M 中删除若干个顶点得图 M'，使得：

$H\subset M'$ 且 $|V(H)|+1=|V(M')|=k$

显然，M' 是 G 的有 $k+1$ 个顶点的完全子图。记 M' 的邻接矩阵为 X，则有：X 的 code 码全 1，即：$code(X)\in Ccode(k+1)$。这与 $Ccode(k+1)$ 中没有极大子 code 码相矛盾。所以 H 是 G 的极大完全子图。

[原命题得证]

显然，我们还可以得到如下性质：

性质 3 若图 G 中只有一个 k 阶矩阵对应的 code 码在 $Scode(k)$ 中，那么 $Ccode(k+1)$ 中没有极大子 code 码。

3 FMCSG 算法

3.1 算法的理论基础

给定一顶点序列 $V=v_1v_2\cdots v_n$，引入两个函数：

$$Head(V)=v_1,Tail(V)=\begin{cases}v_2v_3\cdots v_n & |v|>1\\ \phi & |v|=1\end{cases}$$

定义 8 给定图 $G=(V,E)$，$|V|=n$。称图 $G'=(V',E')$ 为图 G 的逆导出子图，其中 V' 是这样一种点的集合：设 v_0 是 G 中顶点度数最大的点（若顶点度数最大的点不唯一，任选其一），$V'\leftarrow\{v_0\}$，且将 v_0 称为 V' 中的核心点。对于 $\forall v_i\in V$，如果 v_i 和 v_0 之间有边相连，$V'\leftarrow\{v_i\}\cup V'$；$E'$ 是这样的集合：$\forall v_i,v_j\in V'$，如果 $(v_i,v_j)\in E$，则 $(v_i,v_j)\in E'$。

定义 9 给定图 $G=(V,E)$，$G'=(V',E')$ 是 G 的逆导出子图。称 $G'_c=(V'',E'')$ 为图 G 的逆导出补图，其中 $V''=(V-V')\cup\{v_i\mid\forall v_i\in V',v_j\in V-V'$，有 $(v_i,v_j)\in E\}$；E'' 是这样的集合：$\forall v_i,v_j\in V''$，如果 $(v_i,v_j)\in E$，则 $(v_i,v_j)\in E''$。

由定义 8 和定义 9，我们有：

定理 2 设 G' 和 G'' 分别是图 G 的逆导出子图和逆导出补图，则有：$G=G'\cup G''$。

定理 3 一个图可以由若干个图的逆导出子图构成。

证明：由定理 2，图 G 可以由其逆导出子图 G' 及其逆导出补图 G'_c 构成。而 G'_c 又可以由其逆导出子图 G'' 及其逆导出补图 G''_c 构成，依次类推，直到得到的逆导出补图为空，即：

$G=G'\cup G''\cup\cdots\cup G'''\cdots$

[定理得证]

由定理 3，每个逆导出子图的所有极大完全子图构成的集合就是原图的所有极大完全子图集。

定义 10 设 F_2 是图 G 的逆导出子图中的核心点 v_0 与其他各点 $v_i(0<i\leqslant n)$ 两两组合构成的顶点序列的集合，$F_2=\{v_0v_1,v_0v_2,\cdots,v_0v_m\}$，二元关系 $\equiv_{head}$，称为 Head 关系，规定为 $X\equiv_{head}Y$，当且仅当 Head(X)=Head(Y) 且 $X,Y\in F_2$。

定理 4 Head 关系为等价关系[9]

类似地，我们定义如下极大完全子图的等价关系。

定义 11 设 $G=G_1\cup G_2\cup\cdots\cup G_m$，其中 $G_i(1\leqslant i\leqslant m)$ 是逆导出子图，且

$\sum=\{v_0v_1\cdots v_k \mid (v_0v_1\cdots v_k)$是 G_i 的极大完全子图的顶点序列，v_0 是核心点，$1\leqslant i\leqslant m\}$

二元关系 $\equiv_{head}$，称为 MaxCompleteSubgraph 关系，规定为 $X\equiv_{head}Y$，当且仅当 $Head(X)=Head(Y)$且 $X,Y\in\sum$。

定理 5 MaxCompleteSubgraph 关系为等价关系。

由 MaxCompleteSubgraph 可以将∑划分为多个等价类，其中每个逆导出子图对应一个等价类。设逆导出子图为 G'，其核心点为 v_0，当等价类中有 $X\in\sum$且 $Head(X)=v_0$，用 $\varphi_c(v_0)$表示等价类。

定理 6(局部划分定理) 设 v_0 是 G 的逆导出子图 $G'=(V',E')$的核心点，以下两个步骤可以求得 v_0 表示的等价类 $\varphi_c(v_0)$：

步骤 1：将 G'中含有极大子 code 码的 k 阶子矩阵中的顶点序列放入 $\varphi_c(v_0)$，再由 k 阶子矩阵生成 $k+1$ 阶子矩阵，生成候选 code 码，找到极大子 code 码，$\varphi_c(v_0)\leftarrow$极大子 code 码对应的顶点序列。

步骤 2：如果 $\varphi_c(v_0)$中某一顶点序列中的所有顶点均出现在另一顶点序列中，则删除该顶点序列，直到不存在这样的顶点序列为止，得到一个集合簇 $\varphi_c(v_0)$。

证明：因为等价类 $\varphi_c(v_0)$中的顶点序列的元素的第一个是 v_0，由定理 5 可知，第二个元素应该是 V'中不等于 v_0 的某一个元素 y。如果该元素作为 $\varphi_c(v_0)$中某个完全子图的顶点序列的第 2 个元素，该序列形如 $X=xy\cdots$，则存在序列 $Z\subseteq\varphi_c(v_0)$，使得 $Tail(X)=y\cdots\subseteq Z$。

假设 $Tail(X)=y\cdots$序列中的元素不能构成完全子图，则 $X=xy\cdots$中的元素也不能构成完全子图，出现矛盾，则序列 $Tail(X)=y\cdots$的元素构成以 y 开头的完全子图。这就证明了步骤 1 的正确性。

步骤 2 则去掉了 $\varphi_c(v_0)$中不是极大完全子图的序列。

[定理得证]

例如，图 3 中给定 $G=(V,E)$，它的逆导出子图 G_1，其核心点为 A，$\varphi_c(A)=\{ABHI,ADEF,ABD,AHG,AGF\}$，再在 G 的逆导出补图 G_2 中找到其逆导出子图 G_{11}，其核心点是 J（度数最大的顶点不唯一，任选其一），$\varphi_c(J)=\{HIJ\}$，接着在 G_2 的逆导出补图 G_3 找到其逆导出子图 G_{111}，其核心点是 C，$\varphi_c(J)=\{BCD\}$，由于 G_3 的逆导出补图为空，因此停止寻找。此时我们得到图 G 的所有极大完全子图{ $ABHI,ADEF,ABD,AHG,AGF,HIJ,BCD$}。

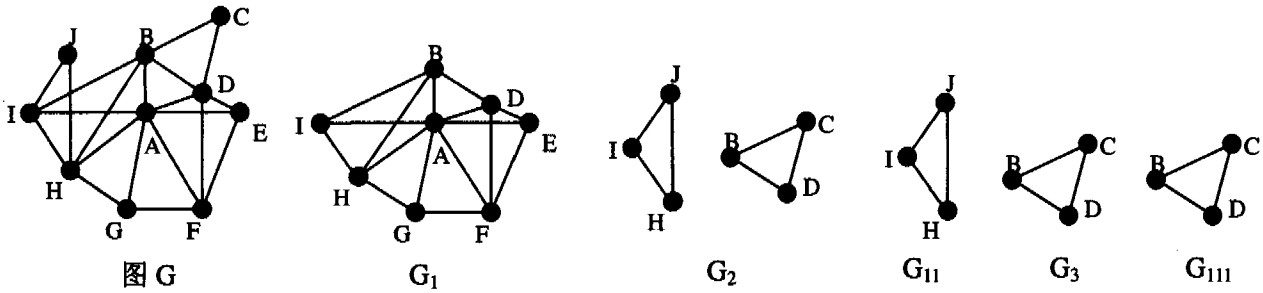

图 3 G 的极大完全子图的寻找过程

下面的定理 7 保证了算法 FMCSG 的完全性：

定理 7 v_0 是 G 的逆导出子图 $G'=(V',E')$的核心点，$\forall G'$的极大完全子图 $G_1'=(V_1',E_1')\subset G'$，都有 $v_0\in V_1'$。

证明(反证法)：假设 $\exists G'$的极大完全子图 $G_2'=(V_2',E_2')\subset G'$，其中 $v_0\notin V_2'$。设 $V_2'=\{v_1,v_2,\cdots,v_i\}(0<i<n)$，因为 G_2'是极大完全子图，那么 $\forall v_m,v_n\in V_1'(0<m\leqslant i,0<n\leqslant i)$都有 $(v_m,v_n)\in E_2'$，又根据图的逆导出子图的定义，$\forall v_m\in V_1'(0<m\leqslant i)$都有 $(v_m,v_0)\in E'$，再根据完全图的定义，知由 $V''_2=V'_2\cup v_0$，$E''_2=E'_2\cup\{(v_m,v_0)\mid\forall v_m\in V_1'(0<m\leqslant i)\}$构成的图 $G_2''=(V''_2,E''_2)$是一个完全图，有极大完全子图的定义知 G'_2 不是极大完全子图，与假设矛盾。原命题成立。

[定理得证]

3.2 FMCSG 算法

根据 3.1 节的理论基础，我们给出 FMCSG(Finding Maximal Complete Subgraph)算法如下。显然，前述的理论结果保证了算法的正确性。

算法 FMCSG

输入：图 G；

输出：Nset；// Nset 代表极大完全子图的顶点序列集，初始值为空；

(1)$G_0\leftarrow G$；

(2)求出图 G_0 的逆导出子图 G'与逆导出补图 G'_c；

(3)if G'_c 为空，退出；

(4)$k\leftarrow 1$；

将 G'的核心点 v_1 对应的一阶矩阵分别与 G'中其他顶点对应的一阶矩阵相连接，生成二阶矩阵；

(5)for $k=2$ to n do // n 表示图的顶点个数；

将 G'中含有极大子 code 码的 k 阶矩阵中的顶点序列放入 $\varphi_c(v_1)$；

由 k 阶矩阵两两相连接，生成 $k+1$ 阶矩阵；

生成候选 code 码，找到极大子 code 码；

$\varphi_c(v_1)\leftarrow$极大子 code 码对应的顶点序列；

(6)如果 $\varphi_c(v_1)$中某一顶点序列中的所有顶点均出现在另一顶点序列中，则删除该顶点序列，直到 $\varphi_c(v_1)$中不存在这样的顶点序列为止；

(7)$G_0\leftarrow G'_c$，转(2)；

设图 G 的结点数为 n，可以证明算法 FMCSG 的时间复杂度为：

$T(n)=T(n-1)*T(n-1)+O(n^2)$。

结论 最大完全子图问题是一个著名的组合优化问题，它是最早被证明的 NP-完全问题之一。由于理论研究及应用的需要，人们仍致力于寻找一些可行的算法。这些算法，无论是确定性算法还是启发式方法，其寻找最大完全子图的过程大部分都是先寻找极大完全子图集，再从中找出顶点个数最多的作为最大完全子图。因此，研究有效的极大完全子图算法既具有理论意义也具有应用价值。

本文通过对 Code 码和 Max-Code 码的研究，得到了一些有价值的结论。给出了图的逆导出子图分解和局部划分定

(下转第 200 页)

图 6　聚类结果

(2)支持平台、数据展现:适用于多种操作系统和多种数据源。通过将操作推向数据库,提升数据库和数据仓库的投资回报率。

(3)用户界面:通过连接节点的表示形式,模型在可视编程环境中被确定。

(4)数据准备:设置了全部的数据挖掘过程,包括大量的数据准备功能,不需要通过 SQL Sever 来处理数据库。

(5)模型发布:Clementine Solution Publisher 使分析人员能够抽出全部的数据挖掘过程。发布模型和升级模型既容易也经济。Clementine 也可将模型输出到 C、SQL 语言,通过编程来实现应用。

(6)聚类方法采用层次聚类方法。层次方法的特点在于,一旦一个步骤(合并或分裂)完成,它就不能撤消。其优点是不用担心组合数目的不同选择,计算代价会较小。其缺点是它不能更正错误的决定。

4.2　DBMiner 具有如下特点

(1)融入了多种知识发现思想;系统功能较齐全;系统提供了一种较完整的知识发现语言 DMQL。

(2)支持平台、数据展现:适用于多种操作系统。使用 excel 作数据展现。

(3)用户界面:提供了直观的图形用户界面,可视化的数据浏览工具。

(4)数据准备:直接通过 Microsoft SQL Sever OLAP Manager 实现访问多种关系数据库系统。

(5)模型发布:提供一个开放式体系结构,能够轻松实现与流行数据库和前端工具的集成。

(6)能处理千兆级的大型数据库。

(7)聚类算法采用 K-Means 算法,K-Means 算法的不足之处在于它要多次扫描数据库,此外,它只能找出球形的类,而不能发现任意形状的类。还有,初始质心的选择对聚类结果有较大的影响。

总之,在选择数据聚类工具时需要考虑很多因素,应根据特定的应用需求加以选择。国外许多行业已经大量利用数据挖掘工具来协助其业务活动,国内在这方面的应用还处于起步阶段,研究和学习国外先进的数据挖掘工具是很有必要的。

参 考 文 献

1　Han Jiawei, Kamber M. Data Mining : Concepts and Techniques [C]. Mongan Kaufmann publishers, 2000. 225～278

2　Alsabti K, Ranka S, Singh V. An eficient k-means clustering algorithm[A]. In: IPPS-98, Proc. of the First Workshop on High Performance Date Mining[C]. Orlando, Florida, USA, 1998

3　Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm forDiscovering Clustersin Large Spatial Databaseswith Noise. In: Proc. 2nd International Conference on Knowledge Discovery an dData Mining[C]. Portland, OR, 1996. 226～231

4　Wang I-IX, Zaniolo C. Database System Extensions for Decision Support: the AxL Approach[A]. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery[C]. 2000. 11～20

5　张雪英. 国外先进数据挖掘工具比较分析. 计算机工程, 2003(9): 1～3

6　马庆国. 管理统计:数据获取、统计原理、SPSS 工具与应用研究[M]. 北京:科学出版社, 2002

7　Zhang T, et al. BIRCH: An efficient data clustering method for very large databases. In: Proc. of the ACM SIGMOD Int' l Conf. on Management of Data. Montreal: ACM Press, 1996. 73～84

8　王实, 高文. 数据挖掘中的聚类方法. 计算机科学, 2000(4): 42～45

(上接第 190 页)

理。基于这些理论研究,提出了一个基于 Max-Code 寻找极大完全子图的算法——FMCSG (Finding Maximal Complete Subgraph)。该算法在搜索极大完全子图的过程中,通过剪掉非 Max-Code 码对应的子矩阵,有效地减少对子矩阵集的遍历次数,提高了算法的搜索效率。文中给出了 FMCSG 算法复杂性的基本估计,并证明了算法的正确性和完备性,即 FMCSG 算法能够找出全部的极大完全子图。

参 考 文 献

1　Washio T, Kunio N, et al. Complete Mining of Frequent patterns from Graphs: Mining Graph Data. Maching Learning, 2003, 50 (3): 321～354

2　Symth P, Goodman R M. An Information Theoretic Approach to Rule Induction from Database. IEEE Trans. Knowledge Data Eng., 1992, 4(4): 301～316

3　Chen A L, Tang C J, et al. An improved algorithm based on maximum clique and FP-tree for mining association rules. Journal of Software, 2004, 15(8): 1198～1207

4　Loukakis E, Tsouros C. A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographically. Computing, 1981, 27: 249～266

5　Kopf R, Ruhe G. A computational study of the weighted independent set problem for general graphs. Foundations of Control Engineering, 1987, 12(4): 167～180

6　Bomze I M, Pelillo M, Stix V. Approximating the maximum weight clique using replicator dynamics. IEEE Trans. Neural Networks, 2000, 11(6): 1228～1241

7　Biggs N. Algebraic graph theory. Cambridge University Press, Cambridge, England, 1974

8　Adleman L M. Molecular computation of solutions to combinatorial optimization. Science, 1994, 226: 1021～1024

9　Sun S L. Algebra Structure. China University of Science and Technology Press, Hefei, China, 1990