基于属性和关系的 OLAP 算法研究

盛玉晓,王童童,李盛恩
(山东建筑大学 计算机科学与技术学院,山东 济南 250101)

摘 要:随着社会网络的兴起,尤其是 OLAP 概念的提出,人们提出了很多对 OLAP 研究的算法,其中对图聚类算法的研究也引起了人们的广泛关注。但是这类算法大多数只是关注节点的属性或者节点之间的关系,而很少同时考虑到节点的属性和它们之间的关系。文中从这两个方面考虑的同时,在划分节点的时候更考虑到了要划分的节点在整个组中的紧密性,把网络中的模块化运用到节点的划分中,使得划分的结果更具有现实意义,而且很好地把 Q 函数的理论应用到社区的划分过程中,更加注重了单个节点对整个社区划分的影响,使得划分之后的各个子社区内部关系更加紧密。

关键词:属性;结构;OLAP;模块化
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)06-0099-04
doi:10.3969/j.issn.1673-629X.2014.06.025

Research on OLAP Algorithm Based on Attributes and Relationships

SHENG Yu-xiao, WANG Tong-tong, LI Sheng-en
(College of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

Abstract: With the rise of social networks, especially the concept of OLAP has been proposed, people make a lot of research on the OLAP algorithm, in which the graph clustering algorithm also attracts widespread attention. However, most of these algorithms are only concerned about attributes or relationships between nodes, rarely taking into account both the node's attributes and the relationships between them. In this paper, consider this problem, meanwhile even taking the problem of the tightness of the node which is in the division group into account, and also use the idea of the modular in the network on partitioning procedure, so that the results of partitioning will have a more realistic significance. And the Q-function theory is so well applied to the process of dividing the community that emphasize on the impact of individual nodes for the entire community, making each child after division within the community much closer.

Key words: attribute; structure; OLAP; modularity

0 引 言

近年来,随着电子信息技术的发展,大量的数据信息随之产生,而且这些信息正潜移默化地影响着人们的生活,如 facebook、web、博客等,而且这些都可以划归到社会网络这一范畴。社会网络定义为:社会网络是一个明确表示人,团体,组织,计算机以及其他一些实体之间关系的名词[1]。简而言之,社会网络可以被抽象成一个复杂的图,其中节点表示社会网络中被研究的对象,边表示对象之间的关系。从而可以把对社会网络的研究等价于对现实社会中的大型图数据集的研究,对图数据集的研究的关键是图聚集算法,即聚类算法。

1 聚类算法简介

聚类算法可以帮助人们在社会网络中发现社区,聚类算法就是将物理或抽象的对象,按照对象间的相似性进行区分和分类的过程。张敏等[2]提出社区发现是一个聚集的过程,社区是图中一些节点的集合,在同一个社区中联系相对稠密,不同社区之间的联系相对稀疏[3]。

聚类算法可以分为下面几种。第一种聚类算法是基于划分的聚类算法,这类划分算法的主要思想是:给定需要划分的社区的数目,运用一些诸如相似、不相似等参数和一个划分社区的标准来达到要划分的目的[4]。这类划分算法中比较有代表性的算法是 k-

收稿日期:2013-08-30 **修回日期**:2013-12-05 **网络出版时间**:2014-02-24
基金项目:国家自然科学基金资助项目(61170052)
作者简介:盛玉晓(1987-),女,山东日照人,硕士研究生,研究方向为数据库;李盛恩,博士,教授,研究方向为数据库。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0922.052.html

means 聚类算法,这个算法是一个基于距离的聚类算法,在这个算法中距离是一个重要的评判标准,如果两个对象之间的距离越接近,就代表它们的相似度越高[5]。在这个算法中,每一次划分都是由距离相互靠近的对象组成的,而且这个算法对于初始点的选取比较重要。当然其他一些相似性的参数也可以在划分算法中用到,如 Jaccard 系数等。

分层聚类算法是另一种比较普遍的聚类算法,这类聚类算法试图将数据组织成一种分层的结构形式,其中又分为凝聚的和分裂的这两种分层聚类方法[6]。凝聚的算法主要采用自底向上的策略,首先每一个对象作为一个类,根据某种度量,将较小的类合并为一个较大的类,直到所有对象都合并到一个类中或者达到了某个终止条件。相反,分层聚类算法采用的是自上而下的策略,首先所有的对象被归并到一个类中,然后根据某种度量,把这一个类划分为更小的类,直到每一个研究对象自成一类或者达到某个终止条件。与基于划分的聚类算法不同的是分层聚类算法不需要基于划分的聚类算法中的聚集数目 K 作为参数,但是分层聚类算法也有一个明显的缺点就是终止条件必须是具体指定的。

光谱聚类算法主要基于部分相连这一概念。这类算法主要通过基于连接或权重的拉普拉斯矩阵实现。如果被研究的社会网络中包含 K 个不相连的社区结构(也被称为部分相连),那么特征值为 0 所对应的特征向量就是这 K 个不相连社区的指标向量[7-8]。如果这个社会网络中只包含相互连接的社区结构,那么可以通过最小的那一个特征值所对应的特征向量来得到 K 个社区结构。

当然还有其他一些算法用于图聚类的分析,如有些算法利用节点的度、聚类中的一些参考系数等来实现社会网络的研究,虽然这些方法也可以得到比较理想的结果,但是其过程是不可控的。而且综合上述几种算法可以很容易地得出:这几种算法在图聚集的过程中只考虑了社会网络中的关系问题,而没有进一步考虑到节点内部所含有的属性。所以提出基于属性和关系的 OLAP 算法是非常有必要的。虽然 K-SNAP 算法[9-10]在进行图聚集的时候从节点的属性和节点之间的关系都进行了研究,但是还有许多可以改进的地方。K-SNAP 算法首先按照用户选择的属性进行分组,然后通过计算找出需要进一步划分的组和与该组联系最紧密的组,进行划分时,如果该组与联系紧密的组有边,那么把该组中的节点划分出去,否则,留在原组内。这种方法在一定程度上能根据用户的意愿产生理想的分组,但是这种分组方法没有考虑到组内的模块化问题,如图 1 所示。

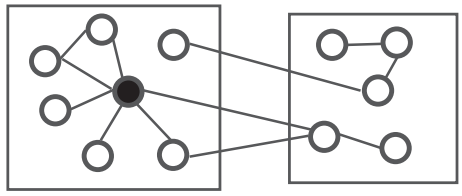

图 1 简单社区结构示例

其中大矩形框内表示将要进行划分的元组,用 G_a 表示该组,小矩形框内表示与 G_a 联系最紧密的元组,用 G_b 表示该组,根据 K-SNAP 算法的划分原则,在 G_a 划分时,图中全黑色标示的节点要被划分到新的组中,然而可以很容易地看出该节点是当前组中与该组内其他节点联系最紧密的节点,如果把此节点划分出去,无疑会破坏掉元组内节点之间的紧密性,从而影响了聚集结果的实用性。根据以上的问题,文中在 K-SNAP 算法的基础上解决了这一个问题,使得在划分的同时更好地考虑到组内节点之间的紧密性问题,而解决这一问题可以用质量函数模块度即社会网络中的模块性很好的解决。

2 算法相关理论

2.1 基于属性和关系的算法思想

图是由顶点和它们之间的关系构成的,但是顶点之间的关系是多种多样的,如在一个网络结构中,人与人之间就可以有不同的关系,那么就需要把数据转换成直观易懂的方式,在下面给出图的定义。

定义 1:图 $G=(V,\square)$。

其中,V 代表图中顶点的集合,每一个顶点都有与之相对应的属性,在这里把属性定义为:$A=\{a_1,a_2,\cdots,a_t\}$,然后可以根据属性值的不同来区分顶点;$\square=\{E_1,E_2,\cdots,E_r\}$($\sum_i E_i = V\times V$)是图中不同类型的边集合,如在一个社交网络中,两个人可以是同学关系,可以是朋友关系,也可以两种关系同时具备。那么对于特定的图 G 有如下定义:$V(G)$ 表示图 G 中顶点的集合,$A(G)$ 用来表示图 G 中属性的集合,$\square(G)$ 用来表示图 G 中边的类型的集合,$E_i(G)$ 表示在图 G 中关系类型为 E_i 的点集合。

如果设定图 G 的聚集数目为 k,那么对于这 k 个分组有如下定义。

定义 2:$\Phi=\{g_1,g_2,\cdots,g_k\}$。

对于每一个分组必须满足以下定义:

(1) $\forall g_i\in\Phi, g_i\subseteq V(G)$ 并且 $g_i\neq\varnothing$;

(2) $\bigcup_{g_i\in\Phi} g_i = V(G)$;

(3) $\forall g_i, g_i\in\Phi(i\neq j), g_i\cap g_j=\varnothing$。

算法在初始时按照属性划分,如果用户选中的属性中有 n 个不同的值,那么就会有 n 个不同的组,一般

情况而言 $n>1$,所以经过第一步的属性划分,再按照结构划分时,需要计算适合划分的元组。对于两个元组 g_i 和 g_j,如果它们之间存在一条边的类型为 E_t,那么可以计算出参与该种类型边的顶点的个数为:

$$p_{E_t,g_j}(g_i)=\{u \mid u \in g_i, \exists v \in g_j,(u,v) \in E_t\}$$

很容易可以计算出两个元组中参与此类型的顶点的个数,然后可以得到两个组中参与此种类型的边的顶点个数的百分比:

$$p_{i,j}^t=\frac{\mid p_{E_t,g_j}(g_i)\mid+\mid p_{E_t,g_i}(g_j)\mid}{\mid g_i\mid+\mid g_j\mid}$$

通常定义如果上式中的值大于 0.5,那么这两个组之间的关系是紧密的,否则的话就是稀疏的[10]。

通过上面的定义对每个组分别计算它与每个组之间联系的参与节点的个数,然后找出需要划分的元组。

假设 Φ_A 是一个图的划分,它包含其中 k 个不同的元组,然后计算

$$\delta_{E_t,g_j}(g_i)=\begin{cases}\mid p_{E_t,g_j}(g_i)\mid & p_{i,j}^t \leqslant 0.5\\ \mid g_i\mid-\mid p_{E_t,g_j}(g_i)\mid & p_{i,j}^t > 0.5\end{cases}$$

$$\Delta(\Phi_A)=\sum_{g_i,g_j\in\Phi_A}\sum_{E_t\in R}(\delta_{E_t,g_i}(g_j)+(\delta_{E_t,g_j}(g_i))$$

可以很容易得出 $\Delta(\Phi_A)\geqslant 0$,并且它的值越小,表示越接近理想的分组[10]。

上面这一个问题已经被证实为 NPC 问题,只能用启发式的算法得到相近的结果,而且用这种方法在划分时没有考虑到元组内部的连接性,所以打算用网络结构中 Newman 等[11-12]提出的基于全局的模块度函数 Q 来得到比较好的划分结果,用来解决上述问题中节点的紧密性问题。

2.2　Q 函数

在一个社会网络中存在 n 个节点,定义一个矩阵 $\boldsymbol{A}_{ij}$,该矩阵用于表示顶点之间的关系,如果顶点 i 和顶点 j 之间存在边,那么在这个矩阵中相应的位置表示 i 和 j 之间存在的边的条数,如果不存在边,那么相应位置的值为 0,通过这个矩阵可以很容易地计算出每一个相应顶点 i 的度 k_i。然后把这个社会网络根据结构划分为两个社区,对于这个社会网络中的节点 i,如果这个节点在划分后属于社区 1,那么令 $s_i=1$,否则 $s_i=-1$。

如果网络中的边的条数为 m,那么 m 的值可以用 $m=\frac{1}{2}\sum_i k_i$ 表示,可以推出顶点 i 和 j 之间存在边的概率为:$k_ik_j/2m$,模块度的值 Q 为落在相同群落所有顶点的总和 $\boldsymbol{A}_{ij}-k_ik_j/2m$,因为这个结果仍然为矩阵,所以用对称矩阵 $\boldsymbol{B}_{ij}$ 来表示 $\boldsymbol{A}_{ij}-k_ik_j/2m$。此时模块度的值可以表示为:

$$Q=\frac{1}{4m}\sum_{ij}(\boldsymbol{A}_{ij}-\frac{k_ik_j}{2m})(s_is_j+1)=\frac{1}{4m}\sum_{ij}(\boldsymbol{A}_{ij}-\frac{k_ik_j}{2m})s_is_j$$

网络的边数 m 和矩阵 $\boldsymbol{A}$ 的值为固定值,所以此公式可写成如下矩阵形式:

$$Q=\frac{1}{4m}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{s}$$

其中,$\boldsymbol{s}$ 为列向量,它的元素为 s_i,然后对矩阵 $\boldsymbol{B}$ 研究发现此矩阵的每一行和每一列的值都为 0,那么矩阵 $\boldsymbol{B}$ 总有一个特征值为 0,其所对应的特征向量为(1,1,…,1)。

在这里将列向量 $\boldsymbol{s}$ 用矩阵 $\boldsymbol{B}$ 的特征值和特征向量表示,即 $\boldsymbol{s}=\sum_{i-1}^{n}a_iu_i$,其中 $a_i=\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{s}$,把 $\boldsymbol{s}$ 代入 Q 得:$Q=\frac{1}{4m}\sum_{i-1}^{n}(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{s})^2\beta_i$,其中 β_i 为矩阵 $\boldsymbol{B}$ 的特征值。为了使网络结构的特征值达到最大,可以通过 $\boldsymbol{s}$ 值的设定来达到想要的结果。即使 s_i 和 $\boldsymbol{u}_i$ 的值尽可能平行,但是因为 s_i 的值只能为 1 或-1,所以可以通过另一种途径来使得 Q 的值达到最大,即如果 $\boldsymbol{u}_i$ 的值为正,令 s_i 为 1,否则为-1。

3　算法实现

3.1　算法设计步骤

算法的设计主要分为两部分:属性划分和结构划分。首先初始化数据,根据用户所选择的属性进行属性划分,然后选择需要划分的组作为当前组,此时根据模块划分的依据,初始化当前组的数据,计算每一个节点的度,接着建立当前组的邻接矩阵,根据邻接矩阵求出矩阵 $\boldsymbol{B}$,计算矩阵 $\boldsymbol{B}$ 的特征值,再计算出最大特征值所对应的特征向量,最后根据特征向量把节点进行分类,但是为了使程序中的 Δ 值尽可能小,同时又不破坏原组中节点的紧密度,此时只把和原组联系最密切且出现在分组 2 中的节点划分出去。

算法步骤如下:

输入:图 G;属性集 $A\subseteq\wedge(G)$;关系类型为 E 的关系集合:$R=\{E\}\subseteq\Upsilon(G)$;产生的聚集数目 K;

输出:聚集图。

1:根据属性分组,初始化数据结构;

2:找出需要划分的组作为当前组 Φ,计算与之联系最紧密的组 Φ_c;

3:while($|\Phi|<=k$) do;

4:建立 Φ 的邻接矩阵,计算各个顶点的度;

5:求出矩阵 $\boldsymbol{B}$,并计算矩阵 $\boldsymbol{B}$ 的特征值;

6:计算最大特征值所对应的特征向量;

7:对节点进行分类;

8:查找与 Φ_c 联系紧密的所有节点,如果节点所对应的 $s_i=1$,把此节点留在原组,否则,划分到新的组中;

9:end while;

10:go to 2;

11:输出所有的聚集图

3.2 实验结果及分析

实验用到的数据集是比较具有权威性的 DBLP 数据集,该数据集主要用来提供计算机科学领域的搜索服务,用来存储发表文献的标题、作者、发表日期等信息。在初始分组时,用文献[13]中提出的方法按照作者发表的论文数目对作者进行了分组,并根据实际测试,得出初始分组如表 1 时,得到的结果最适合。

然后用初始得到的分组,按照 K-SNAP 算法和文中提出的算法对 DBLP 数据集聚集的结果在图 2 中展示,因为文中提出的算法在考虑两个组之间的关系的时候,更从全局的模块性角度进行分析,从而使得组内的关系更加密切,使得用户在 drill-down 或者 roll-up 的时候得到更多有用的信息。

表 1 分组划分示意图

分组	发表论文数目
HP	>5
P	≥3
LP	<3

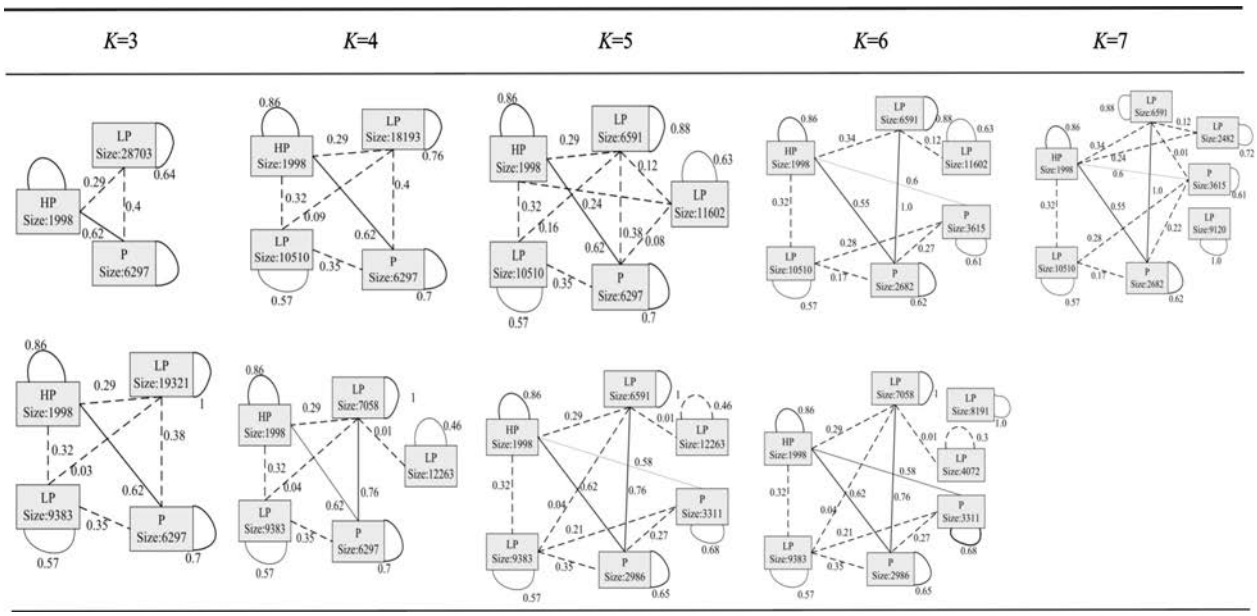

图 2 K-SNAP 算法与文中算法结果对比图

研究人员通过对 DBLP 数据集研究发现,近年来合作发论文的作者呈逐年增长的趋势,而这一点正可以说明作者之间的联系更加紧密,而在分组的时候充分考虑到组内部的结构性问题是非常有必要的[14]。K-SNAP 算法聚集结果是稀疏的,而且随着 K 值的增加,其节点间的连接度下降更快,而引起这一结果的主要原因是 K-SNAP 算法在划分的过程中没有考虑节点之间的连通性。而文中的算法却充分考虑了将要划分的元组的每一个节点在组内的关系,从而能很好地解决这一个问题。

图 3 中展示了两种方法聚集结果的密度关系,其中 A 代表文中算法表示的结果,横轴表示组内密度值,纵轴表示划分次数。

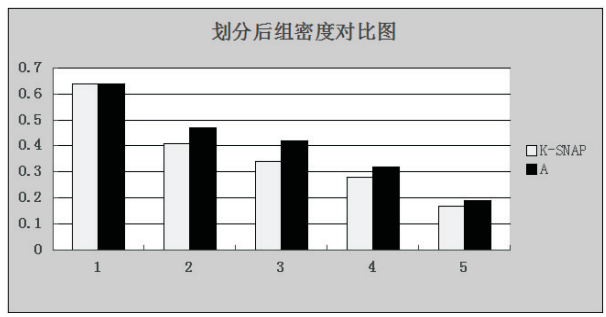

图 3 划分后组密度对比图

4 结束语

文中首先总结了几种比较常用的 OLAP 聚集算法,并在这几种算法的基础了提出了一种新的算法,该算法在考虑了节点属性和节点之间的联系的同时,还充分考虑了节点在整个划分组中的联系性的问题,从而可以让用户在 drill-down 或 roll-up 时得到更多有用的信息。

参考文献:

[1] Özyer T,Rokne J,Wagner G,et al. The influence of technology on social network analysis and mining[C]//Proc of lecture notes in social networks. [s. l.]:Springer,2013.

[2] 张 敏,于 剑. 基于划分的模糊聚类算法[J]. 软件学报,2004,15(6):858-869.

[3] 李晓佳,张 鹏,狄增如,等. 复杂网络中的社团结构[J].

(下转第 106 页)

运行结束后,寄存器中的值就是所要求的 CRC 码。

现假设输入数据 11100000 101,观察图 6,对图 6 的变量含义进行解释。CRC 校验结果为 1100100100110010(C932H),计算结果正确。由图 6 可知,采用 4 个 clk 周期即可完成 CRC 的计算。与串行算法(需 11 个 clk 周期)相比,运算速率提升了 2.75 倍,而且数据位数越多,运算速率提升越明显。

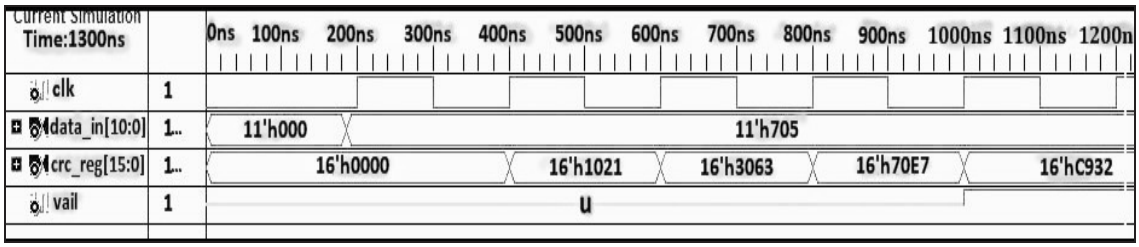

图 6 CRC 仿真图

3 结束语

串行算法电路简单,但存在校验时间长的缺点,并行算法校验时间短,但通常要求数据序列的长度为 8 的整数倍。文中提出了串并结合的算法,既提高了编码速率,又解决了数据序列长度非 8 的整数位的问题。文中推导出的算法具有通用性,可以根据不同的生成多项式推导出相似的实现算法,以运用在各种通信协议中。

参考文献:

[1] Nair R, Ryan G, Farzaneh F. A symbol based algorithm for hardware implementation of cyclic redundancy check (CRC)[C]//Proc of 1997 VHDL international user's forum. Washington, DC, USA: IEEE Computer Society, 1997.

[2] 宋富新,朱晓明,马小社. CRC 编码的并行算法与软件实现[J]. 电子科技,2007(11):62-65.

[3] 张德云,尹勇生,刘志文,等. 面向 USB 应用的 CRC 编解码电路的设计与实现[J]. 合肥工业大学学报(自然科学版),2005,28(3):292-295.

[4] 刘 峰. 超高频 RFID 读写器的研究与实现[D]. 天津:南开大学,2009.

[5] 蒋安平. 循环冗余校验码(CRC)的硬件并行实现[J]. 微电子学与计算机,2007,24(2):107-109.

[6] 朱荣华. 一种 CRC 并行计算原理及实现方法[J]. 电子学报,1999,27(4):143-145.

[7] 赵玉红. 循环冗余校验的实现方法[J]. 雷达与对抗,2006(4):25-27.

[8] 姚 威. 循环冗余校验码并行算法的研究与实现[J]. 计算机与数字工程,2006,34(9):112-114.

[9] 860 MHz-930 MHz class 1 radio frequency identification tag radio frequency & logical communication interface specification candidate recommendation, version 1.0.1[S]. [s.l.]: Auto-ID Center, 2002.

[10] EPC™ radio-frequency identity protocols class-1 generation-2 UHF RFID protocol for communications at 860 MHz-960 MHz version 1.1.0[S]. [s.l.]: EPC Global, 2005.

[11] Information technology automatic identification and data capture techniques-radio frequency identification for item management air interface-part 6: parameters for air interface communications at 860-960 MHz[S]. 2003.

[12] 樊昌信,张甫翊. 通信原理[M]. 北京:国防工业出版社,2005.

(上接第 102 页)

复杂系统与复杂性科学,2008,5(3):19-42.

[4] 张 聪,沈惠璋. 复杂网络中社团发现的快速划分算法[J]. 系统工程,2011,29(4):93-98.

[5] 山玉段,徐 勇,安利平. 一种复杂网络中社团划分的新算法[J]. 系统工程,2012,30(2):120-123.

[6] 程学旗,沈华伟. 复杂网络的社区结构[J]. 复杂系统与复杂性科学,2011,8(1):57-70.

[7] 骆志刚,丁 凡,蒋晓舟,等. 复杂网络社团发现算法研究新进展[J]. 国防科技大学学报,2011,33(1):47-52.

[8] 杨 博,刘大有,LIU Jiming,等. 复杂网络聚类方法[J]. 软件学报,2009,20(1):54-66.

[9] Zhang Ning, Tian Yuanyuan, Patel J M. Discovery-driven graph summarization[C]//Proc of ICDE. [s.l.]: IEEE, 2010:880-891.

[10] Zhou Yang, Cheng Hong, Yu J X. Graph clustering based on structural/attribute similarities[J]. Proceedings of the VLDB Endowment, 2009, 2(1):718-729.

[11] Elmacioglu E, Lee D. On six degrees of separation in DBLP-DB and more[J]. ACM SIGMOD Record, 2005, 34(2):33-40.

[12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2):026113.

[13] Tian Yuanyuan, Hankins R A, Patel J M. Efficient aggregation for graph summarization[C]//Proc of SIGMOD'08. New York, NY, USA: ACM, 2008:567-580.

[14] 解 冶,汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学,2005,2(3):1-12.

基于属性和关系的OLAP算法研究

作者： 盛玉晓， 王童童， 李盛恩， SHENG Yu-xiao， WANG Tong-tong， LI Sheng-en
作者单位： 山东建筑大学 计算机科学与技术学院,山东 济南,250101
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(6)