

数据立方体格的图结构特性研究

王 洋,游进国,张 婷,张正凡

(昆明理工大学 信息工程与自动化学院,昆明 650500)

摘 要: 数据立方体是数据仓库的核心数据模型,其对应的数据立方体格因保留上卷下钻的语义关系而有利于查询和导航。目前对数据立方体内部结构特性尚未展开系统的研究。为此,将数据立方体格视为图数据,研究数据立方体格的结构特性和解析模型。分析结果表明,不同于随机网络和复杂网络的模型,数据立方体格在度分布、聚集系数、平均最短路径等方面具有不同的结构特性。根据上述特性进一步建立数据立方体格的解析模型。

关键词: 数据立方体格;复杂网络;度分布;平均最短路径;聚集系数

中文引用格式:王 洋,游进国,张 婷,等.数据立方体格的图结构特性研究[J].计算机工程,2017,43(2):68-73.

英文引用格式:Wang Yang, You Jinguo, Zhang Ting, et al. Research on Graph Structure Characteristics of Data Cube Lattice[J]. Computer Engineering, 2017, 43(2): 68-73.

Research on Graph Structure Characteristics of Data Cube Lattice

WANG Yang, YOU Jinguo, ZHANG Ting, ZHANG Zhengfan

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

[Abstract] Data cubes are the core data model of data warehouses. The corresponding data cube lattices facilitate querying and navigation for its preserving semantics of rolling-up and drilling-down. But the intrinsic structure characteristics of data cubes have not yet been systematically researched. To address this issue, this paper studies the structure and the analytical model of data cubes from the graph view. The experimental results show that data cube lattices have different structural characteristics in degree distribution, aggregation coefficient, average shortest path and so on, compared with random networks and complex networks. Further the data cube lattice analytical model is established by utilizing the intrinsic structure characteristics.

[Key words] data cube lattice; complex network; degree distribution; average shortest path; clustering coefficient

DOI:10.3969/j.issn.1000-3428.2017.02.012

0 概述

在数据仓库、联机分析处理(Online Analytical Processing, OLAP)的研究领域存在一种重要的数据模型:数据立方体^[1]。为提高数据仓库联机分析和决策的性能,Gray 等人提出数据立方体算子 CUBE,其泛化 group-by, cross-tab, sub-totals 等操作符,对 group-by 的属性(即维度)构成的视图或方体进行预先聚集计算和物化,从而避免查询时大量的聚合和连接操作。物化视图间存在计算依赖关系,从而形成偏序结构,即数据立方体格。文献[2]提出了商立方体的概念,其在保持数据立方体语义的前提下,采用 Cover Partition 技术将数据立方体中上确界相同的数据单元划分为等价类,对于每个等价类只保存

其上确界,进而实现数据立方体有效的压缩。国内的研究者在商立方体基础上做了进一步研究^[3-7],对商立方体进行了严谨的形式化描述^[3,4],提出了更为精简的立方体格结构^[5,6],并进行了 MapReduce 分布式计算^[7]。

商立方体利用单元间的等价关系对数据立方体进行语义压缩,极大地降低了冗余信息,但当面对海量数据分析时,数据立方体仍然非常庞大,数据的存储、管理及分析面临极大的挑战^[8]。东北大学宋杰等人针对大数据环境提出一种基于 Hadoop 分布式文件系统和 MapReduce 编程模型的分布式 MOLAP 系统 DOLAP(Distributed OLAP),其采用数据分块和线性化算法将维度和度量保存在分布式文件系统中,采用 MapReduce 编程模型实现 OLAP 操作^[9]。

基金项目:国家自然科学基金(61462050);云南省自然科学基金(KKSY201603016)。

作者简介:王 洋(1989—),男,硕士研究生,主研方向为数据仓库、数据挖掘;游进国(通信作者),副教授、博士;张 婷、张正凡,硕士研究生。

收稿日期:2016-01-18 **修回日期:**2016-03-03 **E-mail:**jgyou@126.com

密歇根大学及 Google 等研究机构提出的 MR-Cube 将数据立方体格划分为 reducer-friendly 区域和 reducer-unfriendly 区域,并针对 Top-k 等整体性度量函数计算在 MapReduce 编程模型上予以实现^[10]。文献[11]在 Hadoop 和 Spark 两种环境下对分层封闭立方体进行查询,比较了 2 种环境下的分布式查询效率。在云计算环境中,数据立方体格的划分及计算一直是一个研究热点。研究格结构中节点的分布规律有利于存储数据划分以及计算任务的划分,以达到较好的负载平衡和较小的通信开销。

不同以往对数据立方体的研究,本文强调从图的视角研究数据立方体及其格结构。复杂网络的度分布与其拓扑结构紧密相关,绝大多数复杂网络具有无标度性,其度分布满足幂律分布^[12]。文献[13]基于完全图的生成子图的思想,得到了生成随机图的一种新算法,并用数值实验验证了加边和去边生成的随机图的结构特性是相近的。现实网络的结构类型极为丰富,以往的研究更多关注了这些网络在统计意义上的共性问题,而实际网络的差异性也是一个值得注意的基本问题^[14]。

本文研究以商立方体(格)为代表的数据库立方体格,参考 Erdos-Renyi(ER)随机图、社交网络等经典网络解析模型,视格结构数据为图数据,使用实验统计的方法分析数据立方体格的结构特性(如度的分布、平均最短路径、聚集系数等特性),提出格结构数据的解析模型。

1 相关定义

定义 1(偏序关系 \leq) 设 R 是集合 A 上的一个二元关系,如果 R 具有自反性、反对称性和传递性,那么称 R 为一个偏序关系(或部分序,或半序关系);将集合 A 在偏序关系 R 下做成一个偏序集,记为 $\langle A, R \rangle$ 。

定义 2(上界和下界) 设 $\langle A, \leq \rangle$ 为偏序集 MA , A 中元素 a 称为 A 子集 M 的一个上界(下界),如果 a 满足对 M 中任意元素 m ,都有 $m \leq a(a \leq m)$ 。

定义 3(上确界和下确界) 设 $\langle A, \leq \rangle$ 为偏序集 MA , A 中元素 a 称为 M 的最小上界即上确界(最大下界即下确界),如果 a 是 M 的一个上界(下界),则对 M 的任意一个上界(下界) x ,有 $a \leq x(x \leq a)$ 。

定义 4(格) 给出一个偏序集 $\langle L, \leq \rangle$ 。如果对于任意 $a, b \in L$, L 的子集 $\{a, b\}$ 在 L 中都有上确界和下确界,则称 $\langle L, \leq \rangle$ 是一个格。

定义 5(数据立方体格) 数据立方体以基本表为元组集,以维度属性为坐标轴,不同维度属性(值)的交叉构成了多维空间,该空间的每个点根据上卷、下钻的计算依赖关系构成了数据立方体格。

定义 6(数据单元) 令 C 是在 R 上计算得到的数据立方体, $c = (a_1, a_2, \dots, a_n; m)$ 是数据立方体 C 中的一个数据单元,其中, a_i 是维度 $D_i \in D$ 的一个值(含 ALL 值), $1 \leq i \leq n, m \in M$, 并且当在 (a_1, a_2, \dots, a_n) 中,存在 $k(0 \leq k \leq n)$ 个不等于 ALL 的值时,称 c 是 k 维数据单元。

定义 7(基本元组集 BTS) 给定数据单元 $c \in C$, c 的基本元组集 $BTS(c) = \{t | t \in R \text{ 且 } t \leq c\}$, 即所有上卷到数据单元 c 的基本表元组的集合。

定义 8(等价关系 \equiv) 当 u, v 满足 $BTS(u) = BTS(v)$, 则 u 和 v 等价,记为 $u \equiv v$ 。

定义 9(等价类) 数据立方体的等价类指所有具有等价关系的数据单元的集合。通过等价关系,可以将数据立方体划分为多个等价类。

定义 10(商立方体格) 设 C 为数据立方体的所有等价类的集合,对于所有 $c \in C, q$ 为 c 的上确界,则所有 q 的集合构成商立方体格。

例如,表 1 是一个基本数据表,每条记录分别表示了某时间、某产地和某产品的销量数据。

表 1 产品销量基本数据

时间	产地	产品	销量
1Q	Kunming	Mac	2
2Q	Beijing	Pad	7
2Q	Kunming	Pad	6

以“*”表示 ALL,基本元组集 $\{(2Q, *, Pad, 4)\}$ 与基本表元组 $\{(2Q, Beijing, Pad, 2)\}, \{(2Q, Beijing, Pad, 2)\}$ 分别具有上卷、下钻的语义关系,它们之间的偏序关系使之成为一个数据立方体格。不存在上卷、下钻的语义关系的数据单元处于数据立方体格中的同一个层次,因此,数据立方体格具有明显的层次结构,如图 1 所示。

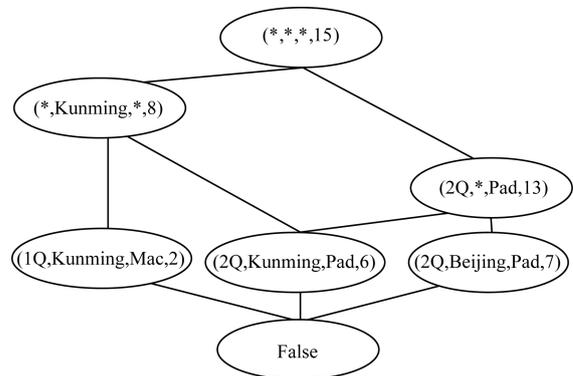


图 1 数据立方体格层次结构

将格中每一个数据单元看作一个结点,它们之间的偏序关系看作一条边,则可将该数据立方体格视为一个图并基于图的结构特性来分析数据立方体格。本文研究的图的结构特性主要为以下 3 点:

1) 度:结点的度指的是与该结点相连接的边的

数量,也就是与这个结点相链接的结点数量。度分布指的是整个网络中不同度值的结点数量分布或度为该值的结点存在的概率分布。

2) 聚集系数: 聚集系数代表了一个网络的聚集的程度大小。设一个网络中共有 K 个结点, 结点 i 与另外 N 个结点相连接, 那么这 N 个结点相互之间最多存在 $\frac{(N-1)N}{2}$ 条边, 而这 N 个结点相互之间存在的真实边数为 M 条, 那么结点 i 的聚集系数为实际边数与最大边数的比值, 也即 $CCF_i = \frac{2M}{(N-1)N}$ 。对于整个网络来说, 平均聚集系数为所有结点聚集系数的平均值, 即 $CCF_{avg} = \frac{1}{k} \sum_{i=1}^k CCF_i$ 。

3) 平均最短路径: 某两个结点之间的最短路径指的是连接这两个点之间的最短通路上的边的数量, 记为 d 。整个网络的平均最短路径则指的是任意两点间最短路径的平均值, 即平均最短路径 $L = \frac{1}{C_K} \sum d$ 。

2 数据立方体格结构特性的实验分析

第1组实验使用 SQL Server 2000 的经典示例数据库 Foodmart 生成商立方体格, 通过实验计算出其度分布、聚集系数、平均最短路径等结构特性, 并与 ER 随机图、社交网络等其他类型的网络进行对比分析。第2组实验对不同大小的商立方体格的结构特性进行对比, 进一步分析不同大小的数据立方体格之间是否具有相似的结构特性。

2.1 实验环境

实验使用的计算机配置: CPU 为 Intel Core i3 2.5 GHz, 内存为 4 GB, 硬盘为 320 GB; 使用的操作系统为 Windows7, 开发语言为 C++, 开发环境为 Microsoft Visual Studio 2012。

2.2 数据立方体格与其他网络图结构的特性对比

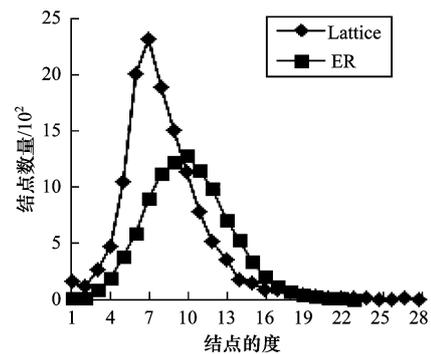
从 Foodmart 数据库中随机抽取其中 10 000 条元组, 利用格结构构造算法^[2]生成商立方体格 Lattice_10k。接着利用斯坦福大学网络分析平台 SNAP^[15]基于 ER 随机图模型生成一个重连概率 $p=1$ 的随机图。生成方法如下: 随机图是由规则网络经过重连生成的, 随机图则对规则网络的每一条边以概率 p 选择另外的结点进行重连。当 p 逐渐增大, 重连的边会逐渐增多, 当 $p=1$ 时, 原规则网络所有的边都会重连。最后加上 Facebook 社交网络数据^[15]形成第1组实验的数据, 具体如表2所示。

表2 第1组实验数据

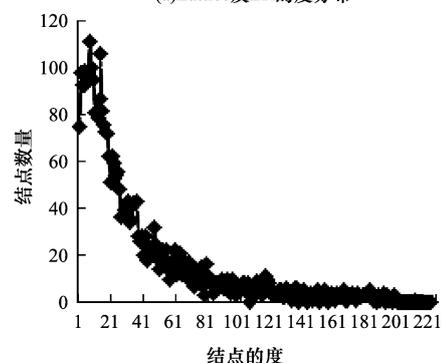
图结构	结点数	边数
商立方体格	13 230	53 424
ER 随机图	10 000	50 000
社交网络数据	4 000	80 000

分别对表2中的3种不同类型的图数据计算其度分布、聚集系数、最短路径等结构特性, 实验结果分别如图2~图4所示, 本文对数据立方体格、ER 随机图、Facebook 社交网络分别以 Lattice, ER, Facebook 简称。

图2是3种不同类型网络结构的度分布情况, 其中横轴表示结点的度值, 纵轴表示度为该值的结点的数量。图2(a)中展示的分别是 Lattice 和 ER 的度分布情况。经对比发现, Lattice 的度分布并不同于 ER 的度分布, Lattice 的度分布类似于泊松分布, 先急剧跃升, 达到一个峰值之后再指数衰减, 度的分布在某个小范围内达到最大, 而在其他范围内则很小。图2(b)中所展示的是 Facebook 的度分布情况, 可以看出, Facebook 中度很小的结点占到了绝大多数, 而度很大的点只占了很少的一部分, 其度分布服从长尾分布。因此, Lattice 的度分布也不同于 Facebook 的度分布。



(a) Lattice及ER的度分布



(b) Facebook的度分布

图2 Lattice, ER, Facebook 社交网络的度分布

图3是3种不同类型的网络结构的聚集系数分布图, 其中横轴表示结点的度值, 纵轴表示度为该值的所有结点的平均聚集系数。图3(a)中展示的是 Lattice 和 ER 聚集系数的分布情况, 根据计算, Lattice 的平均聚集系数为 0.004 3, ER 的平均聚集系数为 0.001 1。这2种网络结构都具有很小的聚

集系数,但 ER 中聚集系数的大小分布均匀、随机,而 Lattice 在平均聚集系数较小的情况下,度比较小的结点的聚集系数相对度很大的结点来说聚集系数还要明显大一些,并不同于 ER 的相对均匀的分布。而图 3(b)中所展示的 Facebook 的平均聚集系数为 0.522 5,符合小世界网络聚集系数比较大的特点,而 Lattice 的平均聚集系数很小,与 Facebook 也是完全不同。

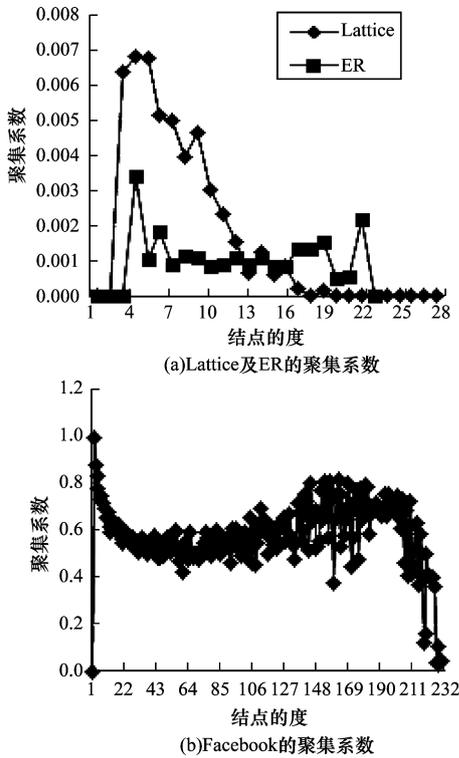


图 3 Lattice,ER,Facebook 社交网络的聚集系数

图 4 是 3 种不同类型的网络结构的最短路径分布情况,其中横轴表示路径长度(跳数),纵轴表示路径长度为该长度的结点对的数量。图 4(a)中展示的是 Lattice 和 ER 的最短路径分布情况,计算得出 Lattice 的平均最短路径为 7.21,ER 的平均最短路径为 4.26。两者均具有较小的平均最短路径,但存在于 Lattice 中两点间的路径最大值为 15,其维度数也为 15,即 Lattice 中两点间最长路径与其维度数可能存在一定的关系,而 ER 中的两点间最长路径则与结点的数量等数据有关。图 4(b)中展示的是 Facebook 的最短路径分布情况,其平均最短路径为 3.7,网络中两点间最长路径为 6,满足小世界网络理论,符合现实社交网络的真实情况,显然,Lattice 的最短路径分布情况不同于 Facebook 的最短路径的分布规律。

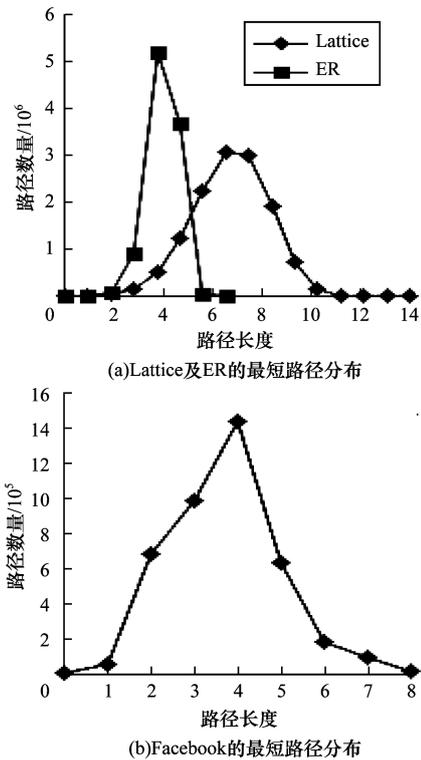


图 4 Lattice,ER,Facebook 社交网络的最短路径分布

2.3 不同大小数据立方体格结构特性对比

为进一步说明数据立方体格独有的结构特性,本文进行了第 2 组实验。第 2 组实验数据除了第 1 组实验中生成的高立方体格 Lattice-10k,再另外从 Foodmart 数据库中随机抽取 20 000 条基本元组,生成了不同结点数的商立方体格 Lattice-20k。具体的实验数据如表 3 所示。

表 3 第 2 组实验数据

数据立方体格	结点数	边数
Lattice-10k	13 230	53 424
Lattice-20k	31 154	138 403

分别计算这两组数据的度分布、聚集系数、平均最短路径等结构特性,实验结果如图 5 所示。图 5(a)展示的是 Lattice-10k 的度分布图,图 5(b)则展示的是 Lattice-20k 的度分布图。经对比可发现,两图中曲线均是先急剧跃升再指数下降,类似于泊松分布。通过计算得出,Lattice-10k 的每个结点的平均度为 8.1,而 Lattice-20k 的每个结点的平均度为 8.9,可见数据立方体格度分布与结点数的多少并无太大关系。图 5(c)展示了 2 个不同大小的商立方体格的聚集系数,它们均具有比较小的平均聚集系数。图 5(d)展示了不同大小的商立方体格的最短路径分布,计算得出 Lattice-20k 的平均最短路径

为 7.65, 两结点之间最短路径的最大值为 15, 在结点数量大幅增加之后, 数据立方体格的平均最短路

径并没有明显的变化, 而两结点之间最短路径的最大值均为其维度数 15。

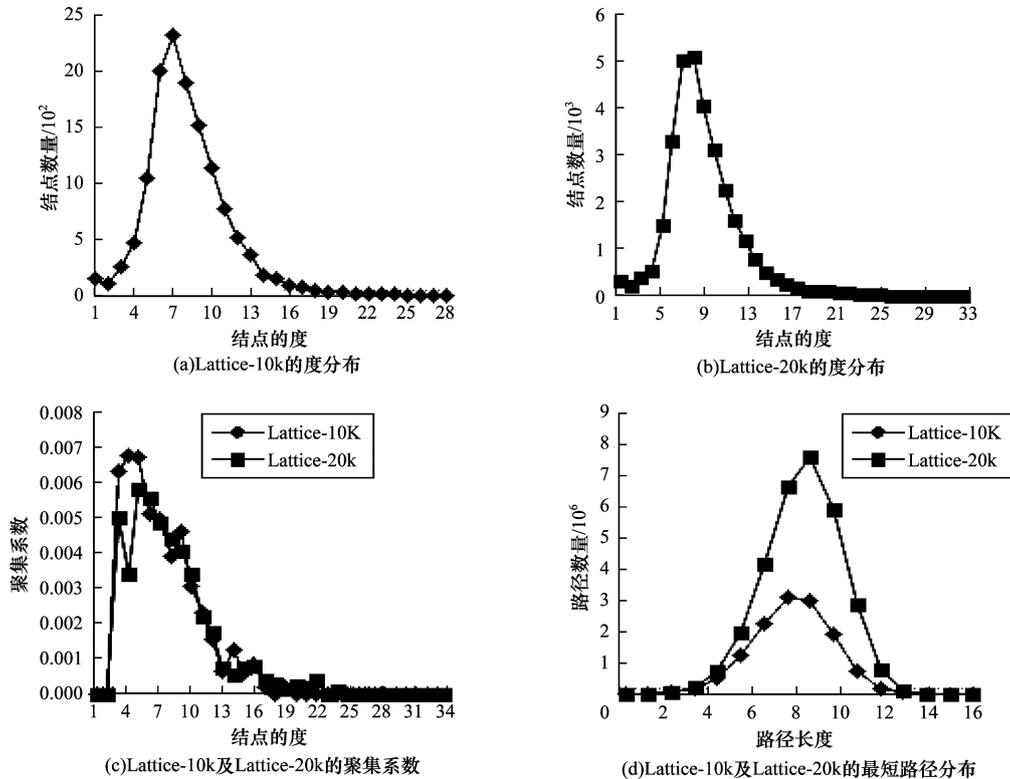


图 5 不同大小商立方体格的结构特性

3 数据立方体格解析模型

根据实验分析可知, 数据立方体格与 Facebook 社交网络等复杂网络具有不同的结构特性。实际典型复杂网络具有较大的聚集系数和较小的平均最短路径, 同时度分布服从长尾分布, 而数据立方体格具有较小的聚集系数和较小的平均最短路径, 度分布则类似于泊松分布。数据立方体格与随机网络也具有不同的结构特性, 因为数据立方体格的产生机制与随机网络是不同的。首先, 在格结构数据的产生过程中, 结点之间并不是按照一个随机的概率进行链接, 而是某些不同的结点之间存在着偏序关系, 依照偏序关系产生边; 其次, 同一层次之间的数据单元之间不存在偏序关系, 也即格结构数据具有一个较为清晰的层次结构; 最后, 格结构具有自己特有的规则结构, 即每两个点对具有上确界和下确界。

根据数据立方体格的独特的结构特性, 发现其解析模型具有如下特征:

1) 度的分布

在以上的实验中可以看出数据立方体格的度分布类似于泊松分布。为了证实这一点, 本文首先用泊松分布对该数据立方体格的度分布进行曲线

拟合。

已知泊松分布的概率表达式如式(1)所示。

$$P(X) = \frac{\theta^x}{x!} e^{-\theta} \quad (1)$$

其中, θ 表示期望值。将图中结点的度值分别与其对应的概率相乘, 然后求和, 可以计算出: $\theta = \sum \text{NodeDegree} \times P(\text{Degree})$, 进而绘制出泊松分布的曲线。如图 6 所示, 其中, 纵轴 $P(\text{Degree})$ 代表不同度值结点出现的概率; Poission 曲线表示计算得出的泊松分布曲线; Lattice-All 表示 Lattice-10k 的度分布曲线, 可以看到与 Lattice-All 曲线基本拟合。

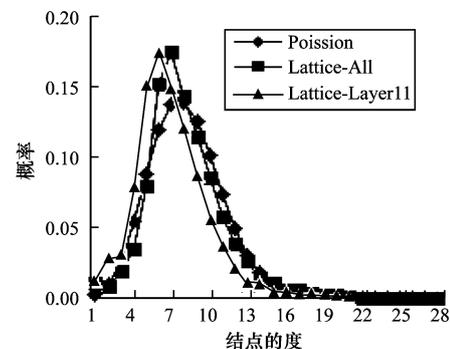


图 6 数据立方体格的度分布及模拟泊松分布

对数据立方体格做更深层次的研究,由数据立方体格的定义可以知道数据立方体格具有明显的层次结构。选取 Lattice-10k 数据集,通过实验统计每一层次的结点的度分布,实验结果表明每一层次内的结点的度分布均服从泊松分布。为简化起见,本文仅选取数据立方体格中结点数量最多的第 11 层来进行展示。如图 6 所示,Lattice-Layer11 曲线表示的是第 11 层结点的度分布,可以看到其度分布也类似于泊松分布。根据泊松分布的可加性,进一步证实了数据立方体格的度分布类似于泊松分布。

2) 聚集系数

由于数据立方体格具有明显的层次结构,并且根据实验得出在这种层次结构中,中间层结点的数量远大于两边层结点的数量,也就是“两头小、中间大”。而数据立方体格结点之间的边是由偏序关系产生的,同一层的结点具有相同的不为“ALL”值的维度数,因此同一层的结点之间不存在泛化(特化)关系。在一个数据立方体格中,处于同一层的任意两结点之间不存在边,而都存在上确界和下确界,即数据立方体格中存在较多的格,而只有较少的三角形。取任意一个结点 i ,其聚集系数计算公式为:

$$CCF_i = \frac{2M}{(N-1)N} \quad (2)$$

其中, N 为结点 i 的度; M 为与结点 i 连接的所有结点之间实际存在的边数。已知整个数据立方体格的平均聚集系数的计算公式为:

$$CCF_{avg} = \frac{1}{K} \sum_{i=1}^K CCF_i \quad (3)$$

其中, K 为数据立方体中总结点数量。如图 1 所示,结点 i 总是作为同一层中某两个结点的上确界(下确界),即结点 i 与某一层之间很多结点相连接,而同一层之间任意两结点不存在边,因此, M 的值较小。根据式(2)可以看出结点 i 具有较小的聚集系数,进而结合式(3)得出数据立方体格具有较小的平均聚集系数的结论。

3) 平均最短路径

在数据立方体格中,结构的层数与其维度数总是相同的。对于一个数据立方体格,其结点的维度数为 d ,则该结构共有 d 层,令自下至上分别为第 0,1, ..., $d-1$ 层。取任意两个结点 u, v ,设其分别位于第 k 层和第 m 层,计算 u 到 v 的最短路径。取最坏的情况,假设每条路线都需要经过最上方的结点或者最下方的结点,则 u 到 v 的最短路径长度(跳数)应当不大于:

$$\min\{(d-k) + (d-m), (k+m)\} \quad (4)$$

可以计算出当 $k+m=d$ 时,式(4)具有最大值 d ,即 u 到 v 的最短路径的长度小于等于其维度数 d 。因此,数据立方体格中任意两点的最短路径的最大值总是小于等于其维度数 d ,使得数据立方体格具有比较小的平均最短路径,并且不同大小、相同维度的数据立方体格的层数总是相同或者相近,对于维度

相同、结点数量相差很大的数据立方体格,其平均最短路径并没有明显随着结点数的增多而显著增加。

4 结束语

本文首先对数据立方体格的概念进行简单介绍,然后利用现实数据集生成一个数据立方体格。通过实验分析该数据立方体格的结构特性,并对其在结构特性方面与 ER 随机图以及现实社交网络的差异,进而构建对应的解析模型。

下一步将根据格结构数据的度分布特性,在分布式计算环境下设计更合理的格结构划分方法,提高格结构数据分布式计算的负载平衡和通信效率。

参考文献

- [1] Han Jiawei, Kamber M, Pei Jian. 数据挖掘:概念与技术[M]. 3版. 范明,孟小峰,译. 北京:机械工业出版社,2012.
- [2] Lakshmanan L, Pei Jian, Han Jiawei. Quotient Cubes: How to Summarize the Semantics of a Data Cube[C]// Proceedings of the 28th International Conference on Very Large Data Bases. San Francisco, USA: Morgan Kaufmann, 2002: 778-789.
- [3] 向隆刚,龚健雅. 一种高度浓缩和语义保持的数据立方[J]. 计算机研究与发展, 2007, 44(5): 837-844.
- [4] 师志斌,高献卫,刘忠宝. 一种包含属性蕴含语义的数据立方体结构[J]. 小型微型计算机系统, 2014, 35(5): 1005-1009.
- [5] 李盛恩,王珊. 封闭数据立方体技术研究[J]. 软件学报, 2004, 15(8): 1165-1171.
- [6] Li Hongsong, Huang Houkuan. PMC: Select Materialized Cells in Data Cubes[J]. Journal of Computer Science and Technology, 2006, 21(2): 297-305.
- [7] 冷芳玲,鲍玉斌,于戈,等. 基于 MapReduce 的封闭数据立方[J]. 计算机研究与发展, 2011, 48(S3): 232-238.
- [8] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.
- [9] 宋杰,郭朝鹏,王智,等. 大数据分析的分布式 MOLAP 技术[J]. 软件学报, 2014, 25(4): 731-752.
- [10] Nandi A, Yu Cong, Bohannon P, et al. Distributed Cube Materialization on Holistic Measures[C]// Proceedings of the 27th IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2011: 183-194.
- [11] 张志朋. 格结构数据分布式存储研究[D]. 昆明:昆明理工大学, 2015.
- [12] 崔爱香,傅彦. 复杂网络演化模型分析[D]. 成都:电子科技大学, 2010.
- [13] 黄斌,吴春旺,郑丰华,等. 复杂网络中随机图模型研究[J]. 计算机工程与科学, 2014, 36(7): 1377-1383.
- [14] 周涛科学网博客. 复杂网络机遇和挑战(十大问题)[EB/OL]. (2013-08-25). <https://blog.sciencenet.cn/blog-3075-719543.html>.
- [15] Leskovec J. SNAP: Stanford Network Analysis Project[EB/OL]. (2014-02-01). <https://snap.stanford.edu/>.