软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3):732−743 [doi: 10.13328/j.cnki.jos.005170]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

透析计算:面向 OLGP 的 InfoNetCube 高效物化*

刘光明[1], 任 艳[2], 李 川[1], 杨 宁[1], 唐常杰[1]

[1](四川大学 计算机学院,四川 成都 610065)

[2](工业和信息化部电子第五研究所,广东 广州 510610)

通讯作者: 李川, E-mail: lcharles@scu.edu.cn

摘 要: 信息网络数据立方(InfoNetCube)的计算是进行信息网络在线分析处理的基础.然而,不同于传统的数据立方,信息网络数据立方由多个子方体格组成,每个方体格中任意方体(cuboid)的任意单元格都包含一个主题图(或称图度量),因而空间开销较传统数据立方大 2 个数量级以上.如何快速、高效地进行信息网络数据立方的部分物化,是极具挑战的研究课题.提出了基于透析计算思想的信息网络立方物化策略,通过主题图度量在信息维和拓扑维上反单调性运用,提出了基于透析计算的空间剪枝算法,快速透析掉不可能命中的子图度量、方体单元、方体乃至方体格.实验结果表明,所提出的基于透析计算的部分物化策略可以对信息网络方体进行有效剪枝,算法较基于基本方体的部分物化策略运行时间平均降低 75%.

关键词: 信息网络;部分物化;透析计算

中图法分类号: TP311

中文引用格式: 刘光明,任艳,李川,杨宁,唐常杰.透析计算:面向 OLGP 的 InfoNetCube 高效物化.软件学报,2017,28(3):732−743. http://www.jos.org.cn/1000-9825/5170.htm

英文引用格式: Liu GM, Ren Y, Li C, Yang N, Tang CJ. Dialysis computing: Efficient InfoNetCube materialization strategy for OLGP. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):732−743 (in Chinese). http://www.jos.org.cn/1000-9825/5170.htm

Dialysis Computing: Efficient InfoNetCube Materialization Strategy for OLGP

LIU Guang-Ming[1], REN Yan[2], LI Chuan[1], YANG Ning[1], TANG Chang-Jie[1]

[1](College of Computer Science, Sichuan University, Chengdu 610065, China)

[2](The Fifth Electronic Research Institute of Ministry of Industry and Information Technology, Guangzhou 510610, China)

Abstract: Calculation of the information network data cube (InfoNetCube) is the foundation of information online analytical processing. However, different from the traditional data cube, InfoNetCube consists of multiple lattices in which each cuboid contains a topic graph (or graph measurement), thus the storage consumption overhead is two orders of magnitude more than that of traditional data cube. How to materialize the specified cuboids or lattice rapidly and efficiently in the information network is a quite challenging research issue. In this paper, a novel InfoNetCube materializing strategy for information network is proposed based on dialysis computing. By leveraging the anti-monotonicity of topic graph measurement in the information and topology dimensions, a dialysis based space pruning algorithm is constructed to rapidly dialysis out the hidden sub graph, cuboids and lattices. Experimental results show that the proposed partial materialization algorithm outperform the cube based partial materialization strategy, saving almost 75% aggregation time.

Key words: information network; partial materialization; dialysis computing

信息网络(information network)和信息网络在线分析处理(InfoNetOLAP)是 Han 和 Yu 等人在 EDBT 2009

* 基金项目: 国家自然科学基金(61103043); 国家科技支撑计划(2012BAG04B02)

Foundation item: National Natural Science Foundation of China (61103043); National Science and Technology Support Ministry (2012BAG04B02)

收稿时间: 2016-07-31; 修改时间: 2016-09-14; 采用时间: 2016-11-01; jos 在线出版时间: 2016-11-29

CNKI 网络优先出版: 2016-11-29 13:35:11, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1335.012.html

会议和 SIGMOD 2010 会议的 Tutorial 上正式提出和倡导的新研究方向,是对现实空间中海量、多维、复杂结构数据和问题更具一般性的抽象[1,2].通常,这种网络可建模为顶点代表实体,边描述实体间关系的大图[3,4].除了图中潜在的拓扑结构外,顶点及边通常包含多种属性,组成了所谓的信息网络.

虽然当代网络的研究已经存在了数十年[5],但均未同时考虑网络拓扑结构与节点、链接的多维属性这两个方面的因素同时纳入考虑的视野.因此,在有效地管理、查询和汇总这些数据方面存在相当大的技术差距,越来越需要为了缩短这些技术差距而为信息网络提出一些特殊方法.信息网络以及信息网络在线分析处理正是在这样的背景下应运而生[1,2,6−8].

在线分析处理(OLAP)作为传统数据仓库技术的重要组成部分,以其在商业领域深入、广泛地应用价值激起了大量关于多维数据模型和数据立方体的研究工作[6−12].在信息网络中,更为引起人们关注的主题测度是图,即网络拓扑结构,而非传统 OLAP 操作所得到的、经由简单聚集操作得到的单一统计量.技术地来看,信息网络在线分析处理的目的在于使用户能够从多角度、多层次地观察和分析基于图的复杂主题对象.

本文所研究的问题与传统在线事务处理(OLTP)、在线分析处理相对应[6,7],将基于图中心度量的信息网络的多维多层次分析处理称为在线图处理(on-line graph processing,简称 OLGP).

传统的 OLAP 技术以及现有的 Graph OLAP 技术在面向多维的信息网络在线分析处理中都存在较大局限,造成这种局限的原因有两点:其一,在信息网络中,用户关注的主题信息由传统的简单数据值转变为复杂的网络结构,数据间复杂关系的增加,使得传统数据聚集操作和 OLAP 技术不再适用[7,8];其二,现有的 GraphOLAP 技术关注的焦点集中于信息网络的基本操作层面,在处理信息网络的大规模数据方面缺乏良好数据组织、中间结果物化等性能需求的必备基础设施.

如何设计信息网络立方的存储结构,高效、快速地进行信息网络数据立方的物化,是实现信息网络立方的计算和图度量计算融合的关键.本文旨在通过相关基础设施的提出和建模、相关算法的设计,提出在性能和灵活性上都满足现实需要的信息网络数据立方模型和计算方法.本文贡献如下:

(1) 信息网络方体格(InfoNetLattice)的建模和计算

信息网络数据立方(InfoNetCube)面向主题内容进行建模,根据维度组合划分不同的信息网络方体(InfoNetCuboid).如何进行方体格的概念、逻辑和物理结构设计,将主题数据、信息维和拓扑维数据进行清晰、高效的组织,是本文的基本研究内容.

(2) 信息网络数据立方部分物化(partial materialization)策略

InfoNetCube 由于包含多个同构子方体格,且每个方体单元格中均保存一个子图快照,完全物化几无可能.如何进行高效的信息网络数据立方部分物化,采用何种策略、技术路线等是本文的关键内容.

1 信息网络数据立方

为了准确进行多维数据分析,本节首先对信息网络数据立方体中相关概念进行数学定义.

定义 1(信息维). 在无向图 $G(V,E)=G(V,\theta(ID))$中,V 是图中点的集合,E 表示边的集合,函数θ为图 G 的边信息决定函数.设变量 $ID=\{I_1,I_2,\ldots,I_m\}$是 OLGP 中待考察的维度集合.其中,$i=1,2,\ldots,m$.这 m 个信息属性构成的维度集合只能决定图的边集,不能改变图的拓扑结构,称 ID 为信息维集合.对于每个信息维度 I_i,存在一个概念层次集 $H_i=\{h_1,h_2,\ldots,h_m\}$.信息维及其概念分层的配置决定中心度量图的覆盖范围和内容[13,14].

定义 2(拓扑维). 设变量 $TD=\{T_1,T_2,\ldots,T_n\}$是刻画 OLGP 中心度量拓扑结构的一个集合.给定一个无向图,一个图可表示为 $G(V,E)=G(\phi(TD),\delta(TD))$,其中,函数$\phi$为点拓扑决定函数,函数$\delta$为边拓扑决定函数.这 n 个拓扑属性构成的拓扑维决定图的点集合和边集合,从而决定了图的拓扑结构,称 TD 为拓扑维集合.对其中每个拓扑维度 T_i,存在概念层次集 $L_i=\{l_1,l_2,\ldots,l_m\}$.各拓扑维及概念分层的配置决定中心度量图的拓扑形态[13,14].

定义 3(信息网络数据立方体). 信息网络数据立方体是一个四元组,$InfoNetCube=(D,MD,f,aggr)$,其中,

(1) $D=\{ID\cup TD\}$表示维标识集,ID 和 TD 分别表示信息维集合和拓扑维[14]集合;

(2) $MD=\{MID\cup MTD\}$表示指标标识集,其中,$MID_i=\{MID_1,MID_2,\ldots,MID_m\}$表示信息维指标标识集,$MTD_i=$

$\{MTD_1,MTD_2,\ldots,MTD_n\}$表示拓扑维指标标识集;

(3) $f:D\rightarrow MD$ 是 D 到 MD 上的部分映射,称为信息网络数据立方体的基;

(4) $aggr$ 表示 MD 上的聚集函数.

定义 4(维聚集). 在信息网络数据立方体中,

$$Dom=I_1\times\ldots\times I_m\times T_1\times\ldots\times T_n, aggr:2^{|MD|}\rightarrow AGG,$$

其中,AGG 是聚集函数 $aggr$ 的值域.f 关于 $d_i(ID_i$ or $TD_j, 1\leqslant i\leqslant m$ and $1\leqslant j\leqslant n)$的聚集是以下映射[15].

$$f':d_1\times\ldots\times d_{i-1}\times ALL\times d_{i+1}\times\ldots\times d_{m+n}\rightarrow AGG$$

$$\forall(d_1,\ldots,d_{i-1},d_{i+1},\ldots,d_{m+n})\in Dom$$

$$f'((d_1,\ldots,d_{i-1},ALL,d_{i+1},\ldots,d_{m+n}))=aggr(\{k\mid\exists d_i\in Dom, f((d_1,\ldots,d_{i-1},d_i,d_{i+1},\ldots,d_{m+n}))=m\})$$

所谓实体化,是指预先执行某些计算并存储计算结果,使得在数据分析时可以直接使用预先计算好的结果,而不需要从原始数据中计算[15].本文讨论的信息网络方体的实体化都是针对信息维聚集操作和拓扑维的聚集操作进行的.

定义 5(信息网络方体实体化单元集). 信息网络方体的实体化单元集 $M(InfoNetCube)$是满足以下条件的最小集合.

(1) $f\in M(InfoNetCube)$;

(2) $\forall m\in M(InfoNetCube), d$ 是 m 的任意一个维(信息维或拓扑维),m 关于 d 的维聚集 $m'\in M(InfoNetCube)$;

$M(InfoNetCube)$中的元素称为信息网络数据立方体的实体化单元($InfoNetCuboid$).

引理 1. 对于任意给定的信息网络数据立方体 InfoNetCube,其实体化单元集确定且唯一.

引理 2. 在 $M(InfoNetCube)$上定义如下关系:$\prec_1$:$f\prec_1 g$ 当且仅当 g 是 f 在一个维上的聚集.记$\prec$为$\prec_1$ 的传递闭包,称为 $M(InfoNetCube)$上的聚集关系.

2 信息网络方体格体系结构

信息网络方体格建模的主要目标是对用户关注的主题信息和维度信息进行高效组织,为高效、有序的基于图度量的聚集计算提供支持.

信息网络基方体经信息维上卷和拓扑维上卷,不断进行图聚集,逐次生成高层方体.最终,基方体和衍生方体形成信息网络方体格的体系结构,如图 1 所示.

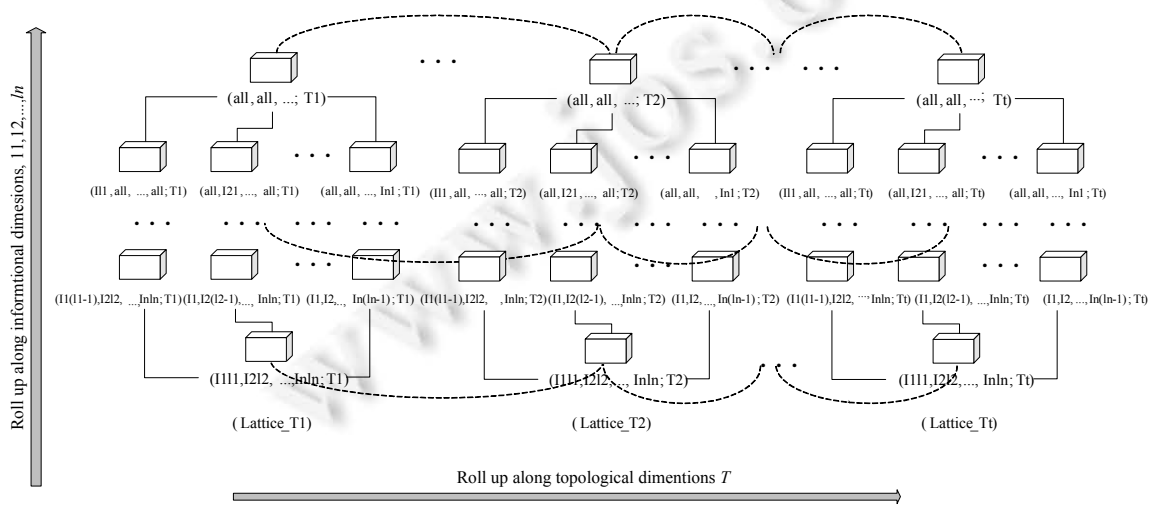

Fig.1 External architecture of the InfoNetLattice design

图 1 信息网络方体格外部体系结构设计

假设当前信息网络的拓扑维 T 包含 t 个概念层次,从特殊到一般,InfoNetLattice 被分为 $Lattice\_T_1$, $Lattice\_T_2$,…,$Lattice\_T_t$ 共 t 个子方体格,其中每个子方体格对应于拓扑维的一个概念层次.沿着特定拓扑维持续上卷,将由 $Lattice\_T_1$ 得到 $Lattice\_T_2$,$Lattice\_T_3$,…,直至最泛化的方体格 $Lattice\_T_t$.

此外,在每个子方体格的内部,方体根据信息维的概念层次自上而下进行组织.位于上方的方体将较位于下方的方体有更泛化的信息维概念层(或其组合).如果我们对 $Lattice\_T_1$ 子格中的基本方体,例如($I1l1$,$I2l2$,…,$Inln$;T_1),沿信息维进行连续的上卷操作,将得到一个从底部方体到顶点方体(all,all,…,all;T_1)的路径.

定义 6(信息网络方体格). 在信息网络方体实体化单元集中,f:$I_1\times\ldots\times I_m\times T_1\times\ldots\times T_{i-1}\times T_{i+1}\times\ldots\times T_n\rightarrow AGG$ 是关于拓扑维 T_i 的维聚集,对 AGG 的所有信息维 I_j 聚集得到实体化单元集 $Lattice\_T_i=\{f(AGG(I_1\times\ldots\times I_{j-1}\times I_{j+1}\times\ldots\times I_m\times T_1\times\ldots\times T_{i-1}\times T_{i+1}\times\ldots\times T_n)),1\leqslant j\leqslant m\}$,称为信息网络的方体格.

引理 3. 信息网络方体格是在拓扑维上对实体化单元集进行的划分,即:

$$\bigcup_{i=1}^{n} Lattice\_T_i = M(InfoNetCube).$$

3 基于基本方体的信息网络数据立方体物化策略

通过上一节对信息网络方体格体系结构分析,在 OLGP 工作的信息网络方体实体化单元集中,每一个实体化单元所包含的信息从传统的数值型聚集值变为子图聚集模型,从而使得 InfoNetLattice 的网络图可以沿着其特定的信息维和拓扑维进行聚集或扩展.同时,由于拓扑维的引入,InfoNetCube 的计算开销将远远大于传统数据立方体,传统的数据立方物化策略已无法应对.本节探讨适用于信息网络数据立方体的不同物化策略.

3.1 完全物化策略

在大规模数据库系统,尤其是分布式数据库系统中,查询响应时间是衡量数据库性能最重要指标之一.传统数据仓库通过对历史数据进行不同维度、不同概念层次聚集,实现对所有方体格的预计算,从而有效提高查询效率.该计算策略称为完全物化.本文借鉴传统 OLAP 完全物化思想,提出基于底层信息网络方体的完全物化策略.通过对每个 InfoNetLattice 结构分析可知,高维度(或同一维度的高概念层次)的方体由低维度(或同一维度的低概念层次)的方体上卷形成,算法伪代码见算法 1.

算法 1. FULL:信息网络方体完全物化算法.

输入:低概念层次信息网络方体单元集合 IL(*InfoNetCuboid*),待上卷的信息维 I_i,待上卷的拓扑维 T_t;

输出:高概念层次信息网络方体单元集合 IH(*InfoNetCuboid*),TH(*InfoNetCuboid*).

```
(1) foreach cuboid in IL(InfoNetCuboid) do
(2)     infoRes=InfoAggregate(I_i,cuboid);          //沿着特定信息维上卷
(3)     IH(InfoNetCuboid).insert(infoRes);          //将上卷结果存入结果集
(4) endfor
(5) foreach subgraph in IH (InfoNetCuboid) do
(6)     topoRes=TopoAggregate(T_t,subgraph);        //沿着特定拓扑维上卷
(7)     TH(InfoNetCuboid).insert(topoRes);
(8) endfor
```

算法的第 1 行~第 4 行首先根据待聚集维度组合信息,通过维聚集操作将低信息维(或同一信息维的低概念层次)基本方体上卷得到高信息维(或同一信息维的高概念层次)实体化单元集合 IH(*InfoNetCube*);算法的第 5 行~第 8 行沿着拓扑维对上一步得到的信息维实体化单元集执行上卷操作,进而得到高拓扑维概念层次实体化单元集合 TH(*InfoNetCube*).

3.2 基于基方体的部分物化策略

虽然完全物化的计算策略能够有效提高查询效率,但是这种策略可能导致维度组合爆炸.拓扑维的引入,使

得计算的复杂度激增到 $O(2^{m+n})$,其中,m 和 n 分别表示在不考虑概念分层的情况下信息维和拓扑维的数量.同时,由于内存容量、存储空间和计算时间等因素限制,对信息网络方体所有实体化单元集进行计算也是不现实的.

研究表明,用户的 OLGP 需求往往具有局部性[15](包括操作局部性和数据局部性).同时,信息网络立方的数据分布通常呈现出严重的倾斜现象.以合作者网络为例,拓扑维取值“作者”概念层次的子方体格中,权值排名靠前的 Top 5%节点和链接的数目占子方体格总节点和链接数的比率少于 1‰;同一单元格中,该比率则通常小于 1%.因此,进行基于用户一般性需求特点的信息网络立方的剪枝不仅势在必行,而且完全可能做到.

本文首先基于信息网络数据立方体的上卷结果提出了一种基于基方体的部分物化策略,算法伪代码如下.

算法 2. BUPM:基于基本方体的信息网络部分物化算法.

输入:低概念层次信息网络方体单元集合 *ILT*(*InfoNetCuboid*),待上卷的信息维 *I*,待上卷的拓扑维 *T*;

输出:高概念层次部分物化的信息网络方体单元集合 *IH*(*InfoNetCuboid*),*TH*(*InfoNetCuboid*).

```
(1)  foreach cuboid in ILT(InfoNetCuboid) do
(2)    infoTmpResult=InfoAggregate(I,cuboid);          //沿着特定信息维上卷
(3)    IHT(InfoNetCuboid)←infoTmpResult;               //暂存完全物化结果
(4)    infoResult=graphPatternBasedPrun(infoTmpResult); //对完全物化结果进行剪枝
(5)    IH(InfoNetCuboid)←infoResult;                   //将部分物化的最终结果放入结果集
(6)  endfor
(7)  foreach subgraph in IHT(InfoNetCuboid) do
(8)    topoTmpResult=TopoAggregate(T,subgraph);        //沿着特定拓扑维上卷
(9)    topoResult=graphPatternBasedPrun(topoTmpResult);
(10)   TH(InfoNetCuboid)←topoResult;                  //将上卷结果存入结果集
(11) endfor
```

该算法的基本思想与完全物化计算思路类似,算法的第 4 行在执行信息维上卷操作时需要额外空间暂存完全物化结果;算法的第 4 行、第 5 行对信息维上卷操作结果集中实体化单元执行剪枝操作,将满足兴趣度度量约束的子图元素(可能打破图的完整性)存入结果集;算法的第 9 行在执行完拓扑维上卷操作之后可直接对该实体化单元执行剪枝操作.

值得注意的是:传统数据立方体部分物化的目标是从实体化单元集中选择出适当的视图(实体化单元)集合,得以保留的视图是按照某一个或几个维度聚集的所有结果;而在信息网络数据立方体的物化结果中,每一实体化单元所保存的内容变为密集或零散的子图,因此,部分物化的操作目标变成从维聚集后的子图中选择出满足用户兴趣度约束的子图模式.

4 基于透析计算的信息网络方体高效物化策略

基于基方体的信息网络数据立方体物化策略虽然可以实现存储空间和 OLGP 响应时间之间的折中,但是这种计算策略仍然需要在完全物化操作所得到的实体化单元集上执行剪枝操作,因此,算法时间和空间复杂度相对较高,不适合在实际场景中使用.针对此问题,提出了基于透析计算的信息网络数据方体高效物化策略.

4.1 透析原理

透析(dialysis)是通过小分子经半透膜扩散到水(或缓冲液)的原理,将小分子与生物大分子分开的一种分离纯化技术[16].如图 2 所示.

本文借鉴生物学中透析的基本思想,通过设置内含特定用户约束的半透膜,仅允许 InfoNetLattice 中不满足用户兴趣度(图度量模型)约束的子图元素、单元格、方体和子方体格等像小分子一样透过,而满足用户兴趣度约束的子图元素则保持在半透膜一侧备用,从而大幅压缩 InfoNetCube 的空间开销和计算时间.

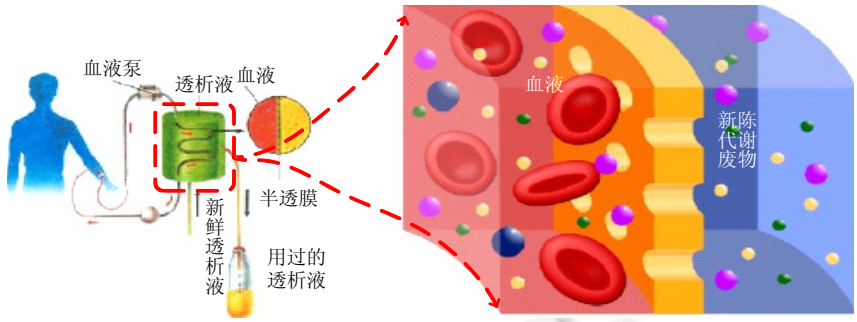

Fig.2 Dialysis principle

图 2 透析原理

4.2 算法设计

图 3 给出了进行 InfoNetLattice 透析计算的技术路线:首先进行基于拓扑维的水平透析,对低拓扑概念层次的方体格尽可能进行整体性剪枝;然后进行垂直透析,大幅削减主题图的规模,实现高效的信息网络数据立方空间优化.

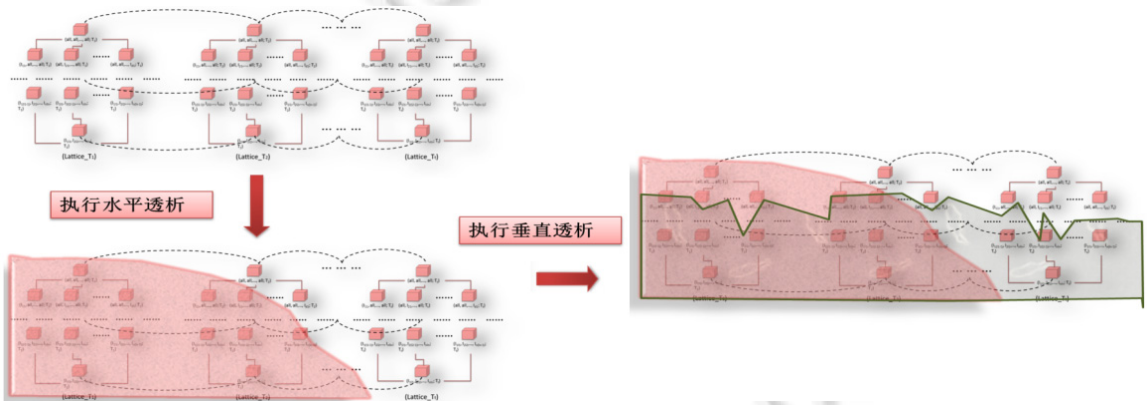

Fig.3 Dialysis computing based partial materialization roadmap of the InfoNetCube

图 3 基于透析计算的信息网络方体部分物化技术路线图

算法 3. DCPM:基于透析计算的信息网络部分物化算法.

输入:原始网络图 $G_1=(V,E)$,信息维待上卷概念层次 I_h,I_l,拓扑维待上卷层次 T_{j-1},T_j;

输出:不同拓扑维概念层次部分物化方体单元集合 *LT*(*InfoNetCuboid*),*HT*(*InfoNetCuboid*).

(1) $G'_j \leftarrow constructHighLevelGraph(G_1,T_j)$;

(2) 根据 *th_Threshold* 对高拓扑维概念层次上卷结果进行剪枝;

(3) 将高拓扑维概念层次提供的剪枝信息映射到低拓扑维对应子图元素;

(4) $G'_{j-1} \leftarrow constructHighLevelTopoGraph(G_1,T_{j-1})$;

(5) 根据 *ih_Threshold* 对高信息维概念层次上卷结果进行剪枝;

(6) 将高信息维概念层次提供的剪枝信息映射到低信息维对应子图元素;

(7) **foreach** *node* **in** G'_{j-1} **do**

(8) **if** (*node* in *lowerInfoPrunedNodes*)

(9) G_j.*remove*(*node*) //从原始数据汇移除对应子图元素

(10) **if** (*node* in *lowerTopoPrunedNodes*)

(11) continue;

(12) 将满足条件的子图元素聚集到网络中;

(13) **endfor**

(14) **foreach** *subGraph* **in** *LTT*(*InfoNetCuboid*) **do**

(15) *reCheck*(*subGraph*) //按照特定子图模式进行二次检查

(16) **endfor**

该算法的基本思想是:通过将高拓扑维概念层次提供的剪枝信息映射到低拓扑维对应子图元素,与高信息维概念层次所提供的剪枝信息相结合,不断对原始数据进行压缩,减少低概念层次实体化单元的初始规模,从而有效降低信息网络方体物化时间开销,提高计算效率.

值得指出的是:图 3 中,阴影部分两次剪枝覆盖的区域未必完全被剪枝掉.事实上,总有一些节点、链接或小规模的强连通子图(团)比较顽强地保留下来.然而,同方体格中的绝大部分图元素都将被透析殆尽.由于透析过程的动态性,且对数据分布具有较紧密依赖关系,透析计算的具体路径和结果边界非常复杂.

5 实验分析

5.1 实验环境

(1) CPU:Intel(R) Corel(TM) i5-3470@3.2GHz;

(2) 内存:16.0GB;

(3) 操作系统:Windows 7 Ultimate 64 位.

5.2 实验数据

本文采用合作者网络作为实验分析对象,实验数据来自 ACM 合作者网络和 Microsoft Academic Graph 真实数据集.为更好地验证算法的性能,本文根据作者合作紧密程度的变化,从 ACM 原始数据集中随机抽取 1 000~12 000 篇文章,从 MAG 原始数据集中随机抽取 5000~40000 篇文章进行实验.数据集各属性数量关系见表 1 和表 2.

Table 1 Statistic information of ACM co-author network

表 1 ACM 合作者网络数据集特征汇总

论文数量	1 000	2 000	3 000	4 000	5 000	6 000	7 000	8 000	9 000	10 000	11 000	12 000
作者	2 424	4 300	6 028	7 486	8 984	10 250	11 227	12 213	13 724	14 861	15 758	16 430
合作关系	9 024	16 212	24 320	32 084	41 248	47 280	53 914	60 220	68 410	73 224	77 784	81 624

Table 2 Statistic information of microsoft academic graph dataset

表 2 MAG 合作者网络数据集特征汇总

论文数量	5 000	10 000	15 000	20 000	25 000	30 000	35 000	40 000
作者	8 992	14 245	20 212	25 567	29 514	39 656	47 388	56 673
合作关系	43 846	77 824	124 664	166 002	216 472	293 596	364 446	439 110

5.3 实验结果分析

将本文提出的算法与完全物化、基于底层基本方体的部分物化策略进行对比,从算法的总体运行时间、不同方体格的物化效率等方面进行了比较,实验结果如图 4~图 9 所示.其中,FULL 表示算法 1 中的完全物化策略,BUPM 表示算法 2 中基于基本方体的部分物化策略,DCPM 表示算法 3 中基于透析计算的部分物化策略.

5.3.1 方体物化时间

对于合作者网络而言,大多数用户对网络的查询需求集中在作者之间的合作关系,因此,本文首先设计了一组基于通用度量模型,即,满足特定支持度阈值的显著性边(如合作频繁度)的信息网络方体部分物化实验方案.实验结果如图 4、图 5 所示.

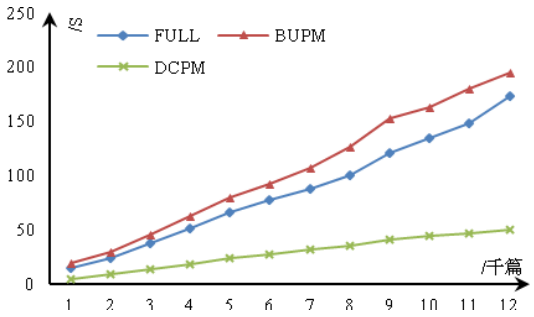

Fig.4 Significant edge based partial materialization time consuming of the InfoNetCuboid (ACM)

图 4 基于显著性边的信息网络单元物化时间比较(ACM)(δ=2)

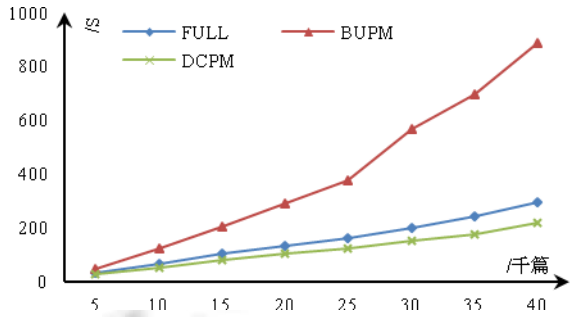

Fig.5 Significant edge based partial materialization time consuming of the InfoNetCuboid (MAG)

图 5 基于显著性边的信息网络单元物化时间比较(MAG)(δ=2)

图 4 和图 5 分别展示了两组数据集上基于显著性边度量(频繁的合作关系)通用度量模型的信息网络方体物化算法时间对比.实验结果表明,完全物化操作所需的计算时间远大于本文提出的基于透析计算的部分物化算法所需要时间.此外,从图中可以看出:随着数据量的增加,本文提出的部分物化策略所需的操作时间增长相对比较缓慢,因此,算法具有较好的扩展性,可以有效应对大规模场景下的应用需求.

考虑到存在部分用户可能对于合作者网络中特定合作模式感兴趣,本文设计了一组基于用户兴趣模型,即中心作者合作关系(如星形结构)的信息网络方体物化实验方案.实验结果如图 6、图 7 所示.

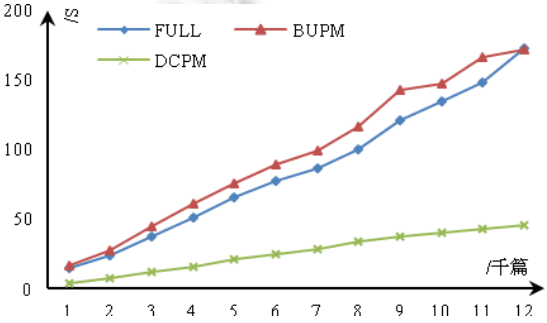

Fig.6 Staring central author based partial materialization time comparision (ACM)

图 6 基于星形中心作者的信息网络方体物化时间比较(ACM)

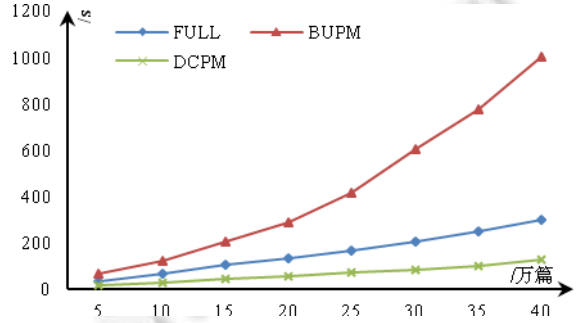

Fig.7 Staring central author based partial materialization time comparision (MAG)

图 7 基于星形中心作者的信息网络方体物化时间比较(MAG)

实验结果表明:在不同的用户兴趣度模型下,基于透析计算的部分物化策略可以有效降低信息网络方体的计算时间开销.分别对图 4~图 7 中的基于基本方体的完全物化策略运行时间和基于透析计算的部分物化策略运行时间的降低百分比取均值,可得出部分物化较基于基本方体的部分物化策略运行效率平均降低 75%,从而表明了本文提出的透析计算策略的有效性.

通过对算法 2 的分析可知:基于基本方体的部分物化策略(BUPM)是在完全物化操作基础上执行的剪枝操作,即,算法的运行时间一定比完全物化策略的运行时间长.本文的实验结果也反映了这一事实.

5.3.2 方体格物化时间

根据第 2 节设计的信息网络方体格体系结构,本文通过实验分别计算了不同计算策略下、不同拓扑维层级执行部分物化操作所需的时间开销,实验结果如图 8、图 9 所示.

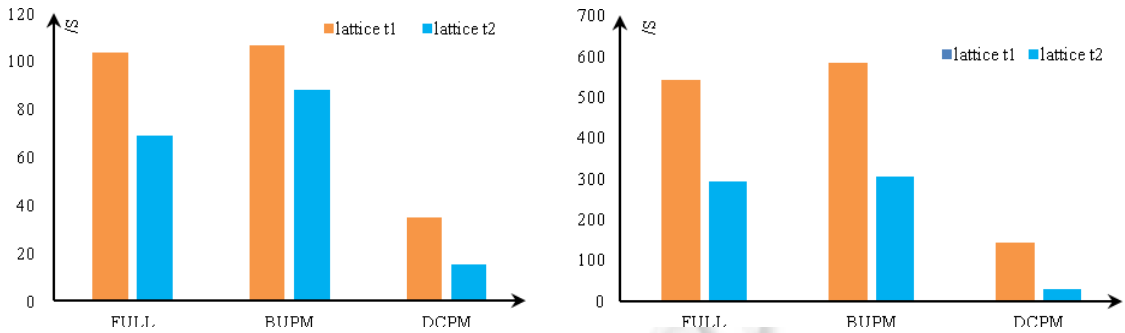

Fig.8 General measure based InfoNetLattice partial materialization time comparision

图 8 基于通用度量模型的信息网络方体格物化时间对比

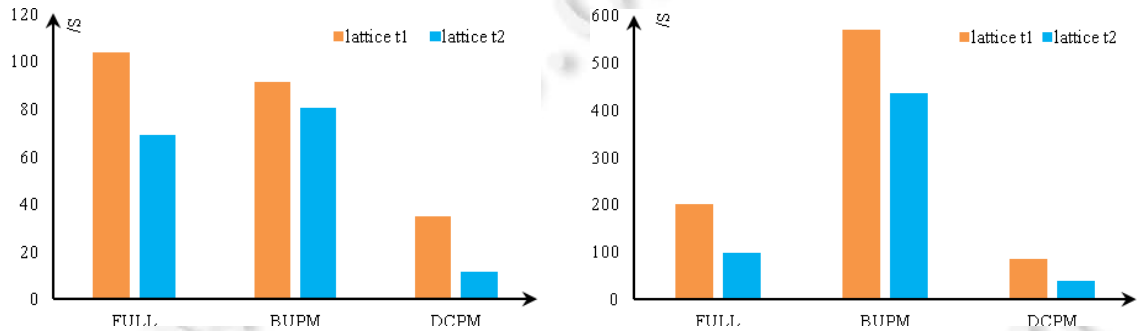

Fig.9 General measure based InfoNetLattice partial materialization time comparision

图 9 基于用户兴趣度模型的信息网络方体格物化时间对比

从图 8、图 9 可以看出:在通用度量模型和用户兴趣度模型中,基于透析计算的部分物化策略不同方体格均具有较高的计算效率,这得益于不同方体格之间和不同信息维层级之间提供的有效剪枝信息.

值得注意的是:在基于用户兴趣度的模型中,基于基本方体的部分物化策略的不同方体格之间的计算时间差距相对较小.这是因为网络中符合该带兴趣度度量的子图模式通常是不平凡的,即,在不同拓扑维层次上的表现出较强的相似性,因此,在不同方体格上的运行时间差距不如通用度量模型下显著.

6 相关工作

Chen 等人[17]首先提出了基于图的在线分析处理,对信息维、拓扑维等基本概念进行了定义,给出了相应的框架设计;该文的扩展版本[18,19]对该框架的技术路线进行详细阐述,但仍然是围绕 I-OLAP 和 T-OLAP 的操作来讨论,并未提出新的观点或相关技术.Qu 等人[20]对拓扑维的在线分析处理做了进一步的讨论,对 T-OLAP 操作中特定类型的度量做了重点分析,但未给出实际的算法与实现;同时,并未对广义的 T-OLAP 指明设计与实现方向.Zhao 等人[21]通过对真实网络进行抽象,提出了 Graph Cube 多维网络模型.但该模型本质上仍与传统数据立方体一致,针对拓扑维的相关操作无法与信息维的相关操作进行有效区分,无法满足信息网络在线分析处理的需求.Li 等人[14]首次提出了以 Graph 数据为中心度量的 OLAP 的概念,文献提出了基于图的数据立方概念和创建过程,但对于信息网络方体格的设计与实现未给出具体的解决方案.Jin 等人[22]提出了一种适用于在线图分析处理的 VisualCube 模型,但该模型只考虑了信息维,无法解决信息网络环境下拓扑维引入所带来的复杂问题.Nie 等人[23]利用维建模的方法对基于图的信息网络数据进行模型设计,实现了多维信息网络仓库模型,为信息网络的在线分析处理提供了必备的基础设施,但关注的焦点仍然集中在原始数据的底层存储结构,该存储结构解决的书物理存储问题,未涉及信息网络的逻辑存储结构进行上层分析处理所需要的技术路线.

Xu 等人[24]首次提出了面向信息网络的在线图处理模型,设计并实现了相应的基本操作算法,但对于基本方体的处理仍然是采用完全物化计算策略,在维度数量较高的情况下,难以应对计算过程中时间和空间方面剧增的问题.Morfonios 等人[25]通过对社交编著系统的分析,提出了一种针对实体聚集视图的查询搜索框架,并采用完全物化的策略来对所有可能的查询需求进行预计算,显然,这种策略的空间开销过于庞大.

Yin 等人[26]指出,Chen 的模型只适用于同构网络,提出了一种适用于异构信息网络建模的 HMGraphCube 模型,但异构网络的复杂度远高于同构网络,因此,如何对异构信息网络进行高效预计算是不可避免的问题.文献关注的焦点仍然是集中在对新操作的讨论,未提及物化的相关问题.Beheshti 等人[27]指出:当前的信息网络在线分析处理模型过分关注基于图的在线查询和分析处理,并不能很好地支持基于语义驱动的计算,因此提出了一种基于图的拓扑结构进行决策的 GOLAP 模型.但该模型的焦点仍然集中在操作层面,未考虑海量数据场景下的操作效率问题.Wang 等人[28]针对当前 GraphOLAP 模型在分析处理异构信息网络方面存在的效率不足问题,提出了一种面向大规模网络的多维分析处理框架.该框架关注焦点仍然集中在信息网络数据立方的建模,并在此基础上引入了两种新的操作.虽然文献提及了一种基于 2-源路径的部分物化策略,但作者未给出详细的算法描述和实施方案.此外,该方案本质上是为新操作服务的,应用场景相对比较局限.Wang 等人[29]提出了一种适用于大规模信息网络的并行图分析处理框架,虽然该模型可以高效地处理大规模网络图结构,但该模型面向的是基于拓扑维的处理场景,即:更多的关注点集中在网络拓扑结构的变化上,未考虑到信息维的相关信息,本质上仍属于 OLGP 较具体的一组应用场景.

Sabine 等人[30]通过对 OLAP 以及著作数据的调研,对信息网络在线分析处理的应用场景进行了分析和归纳,虽然指出了几个可能的研究方向,但是未对信息网络立方的物化工作方向进行预测.同时,近年来也鲜有相关工作涉及具体的部分物化技术和处理思路.

7 总结与展望

本文借鉴医学的透析原理,首次提出了基于透析计算的信息网络数据立方 InfoNetCube 剪枝策略和部分物化原理.本文贡献可总结为:

(1) 提出了信息网络方体格的概念,提出了 InfoNetLattice 外部体系结构和内部体系结构.提出了信息网络方体及信息网络方体单元在信息网络立方体中的结构特点和联系;

(2) 提出了基于透析计算的 InfoNetCube 部分物化策略,设计和实现了相应的算法.实验结果表明,该部分物化策略可以有效降低信息网络方体物化过程的计算时间和空间开销.

本文在基于用户兴趣度度量模型的实验中,算法按照传统的 DFS 策略执行时间相对较长.如何充分利用网络提供的信息来提高算法的子图模式匹配效率,是本文后续工作中需要解决的问题.正如相关工作中提到的,本文的研究场景仍然是基于同构网络,但是本文所设计的信息网络方体格体系结构具有很好地适用性,后续的研究工作将考虑将算法应用场景扩展到异构信息网络上.

References:

[1] Han JW, Yan XF, Yu PS. Scalable OLAP and mining of information networks. In: Proc. of the 12th Int'l Conf. on Extending Database Technology: Advances in Database Technology. ACM Press, 2009. [doi: 10.1145/1516360.1516505]

[2] Han JW, Sun Y, Yan X, Yu PS. Mining knowledge from databases: An information network analysis approach. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2010. [doi: 10.1145/1807167.1807333]

[3] Han JW. Mining heterogeneous information networks by exploring the power of links. In: Proc. of the Int'l Conf. on Discovery Science. Berlin, Heidelberg: Springer-Verlag, 2009. [doi: 10.1007/978-3-642-04747-3_2]

[4] Aggarwal CC, Wang HX, eds. Managing and Mining Graph Data. Vol.40. New York: Springer-Verlag, 2010.

[5] Newman MEJ. Networks: An Introduction. Oxford University Press, 2010.

[6] Gray J, Chaudhuri S, Bosworth a, Layman A, Reichart D, Venkatrao M, Pirahesh H, Pellow F. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery, 1997,1(1):29−53. [doi: 10.1023/A:1009726021843]

[7] Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997,26(1):65−74.

[8] Sarawagi S, Agrawal R, Megiddo N. Discovery-Driven exploration of OLAP data cubes. In: Proc. of the Int'l Conf. on Extending Database Technology. Berlin, Heidelberg: Springer-Verlag, 1998. 168−182. [doi: 10.1007/BFb0100984]

[9] Sun YZ, Wu TY, Yin ZJ, Cheng H, Han JW, Yin XX, Zhao PX. BibNetMiner: Mining bibliographic information networks. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2008. [doi: 10.1145/1376616.1376770]

[10] Burdick D, Doan A, Ramakrishnan R, Vaithyanathan S. OLAP over imprecise data with domain constraints. In: Proc. of the VLDB. 2007. 39−50.

[11] Morfonios K, Konakas S, Ioannidis Y, Kotsis N. ROLAP implementations of the data cube. ACM Computing Surveys (CSUR), 2007,39(4):12. [doi: 10.1145/1287620.1287623]

[12] Zhang N, Tian Y, Patel JM. Discovery-Driven graph summarization. In: Proc. of the ICDE. 2010. 880−891. [doi: 10.1109/ICDE.2010.5447830]

[13] LI C, Yu PS, Zhao L, Xie Y, Lin WQ. InfoNetOLAPer: Integrating InfoNetWarehouse and InfoNetCube with InfoNetOLAP. In: Proc. of the VLDB 2011. Demo, 2011.

[14] Li C, Zhao L, Tang CJ, Chen Y, Li J, Zhao XM, Liu XL. Modeling, design and implementation of graph OLAPing. Ruan Jian Xue Bao/Journal of Software, 2011,22(2):258−268 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3771.htm [doi: 10.3724/SP.J.1001.2011.03771]

[15] Pei J, Chai W, Zhao C, Tang SW, Yang DQ. An algebra for online analytical processing data cube. Ruan Jian Xue Bao/Journal of Software, 1999,10(6):561−568 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/10/561.htm

[16] Yang Y. The principles and developments of the hemodialysis system. Chinese Journal of Medical Instruction, 2001,25(5):288−296 (in Chinese with English abstract).

[17] Chen C, Yan XF, Zhu FD, Han JW, Yu PS. Graph OLAP: Towards online analytical processing on graphs. In: Proc. of the Int'l Conf. on Data Mining (ICDM 2008). 2008. [doi: 10.1109/ICDM.2008.30]

[18] Chen C, Zhu F, Yan XF, Han JW, Yu P, Ramakrishnan R. InfoNetOLAP: OLAP and mining of information networks. In: Yu PS, Faloutsos C, Han JW, eds. Proc. of the Link Mining: Models, Algorithms and Applications. Springer-Verlag. [doi: 10.1007/978-1-4419-6515-8_16]

[19] Chen C, Yan XF, Zhu FD, Han JW. Graph OLAP: A multi-dimensional framework for graph data analysis. Knowledge and Information Systems (KAIS), 2009,21(1):41−63. [doi: 10.1007/s10115-009-0228-9]

[20] Qu Q, Zhu FD, Yan XF, Han JW, Yu PS, Li HY. Efficient topological OLAP on information networks. In: Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2011. [doi: 10.1007/978-3-642-20149-3_29]

[21] Zhao PX, Aggarwal C, Wang M. gSketch: On query estimation in graph streams. In: Proc. of the 38th Int'l Conf. on Very Large Data Bases (VLDB 2012). Istanbul, 2012.

[22] Jin X, Han JW, Cao LL, Luo JB, Ding BL, Lin CX. Visual cube and on-line analytical processing of images. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2010. [doi: 10.1145/1871437.1871546]

[23] Nie ZY, Li C, Tang CJ, Xu HY, Zhang YH, Yang N. Design of multi-dimensional information network data warehouse model for online graph processing. Journal of Frontiers of Computer Science and Technology, 2014,8(1):51−60 (in Chinese with English abstract).

[24] Xu HY, Li C, Tang CJ, Li YT, Dai SC, Yang N. On-Line graphic processing: Information network oriented on-line analytical processing. Journal of Frontiers of Computer Science and Technology, 2012,6(9):797−809 (in Chinese with English abstract).

[25] Morfonios K, Koutrika G. OLAP cubes for social searches: Standing on the shoulders of giants? In: Proc. of the WebDB. 2008.

[26] Yin M, Wu B, Zeng ZF. HMGraph OLAP: A novel framework for multi-dimensional heterogeneous network analysis. In: Proc. of the 15th Int'l Workshop on Data Warehousing and OLAP. ACM Press, 2012. [doi: 10.1145/2390045.2390067]

[27] Beheshti SMR, Benatallah B, Motahari-Nezhad HR, Allahbakhsh M. A framework and a language for on-line analytical processing on graphs. In: Proc. of the Int'l Conf. on Web Information Systems Engineering. Berlin, Heidelberg: Springer-Verlag, 2012. [doi: 10.1007/978-3-642-35063-4_16]

[28] Wang PS, Wu B, Wang B. TSMH graph cube: A novel framework for large scale multi-dimensional network analysis. In: Proc. of the IEEE Int'l Conf. on Data Science and Advanced Analytics (DSAA 2015). Vol.36678. IEEE, 2015. [doi: 10.1109/DSAA.2015.7344826]

[29] Wang ZK, Fan Q, Wang HJ, Tan KL, Agrawal D, Abbadi AE. Pagrol: Parallel graph olap over large-scale attributed graphs. In: Proc. of the 2014 IEEE 30th Int'l Conf. on Data Engineering. IEEE, 2014. [doi: 10.1109/ICDE.2014.6816676]

[30] Loudcher S, Jakawat W, Morales EPS, Favre C. Combining OLAP and information networks for bibliographic data analysis: A survey. Scientometrics, 2015,103(2):471–487. [doi: 10.1007/s11192-015-1539-0]

附中文参考文献:

[14] 李川,赵磊,唐常杰,陈瑜,李靓,赵小明,刘小玲.Graph OLAPing 的建模、设计与实现.软件学报,2011,22(2):258−268. http://www.jos.org.cn/1000-9825/3771.htm [doi: 10.3724/SP.J.1001.2011.03771]

[15] 裴健,柴玮,赵畅,唐世渭,杨冬青.联机分析处理数据立方体代数.软件学报,1999,10(6):561−569. http://www.jos.org.cn/1000-9825/10/561.htm

[16] 杨焱.血液透析系统的基本原理及发展.中国医疗器械杂志,2001,25(5):288−291.

[23] 聂章艳,等.面向 OLGP 的多维信息网络数据仓库模型设计.计算机科学与探索,2014,8(1):51−60.

[24] 徐洪宇,李川,唐常杰,徐洪宇,张永辉,杨宁.在线图处理:面向信息网络的在线分析处理.计算机科学与探索,2012,6(9):797−809.

刘光明(1989−),男,山东临沂人,硕士生,主要研究领域为数据库,数据挖掘,信息网络.

任艳(1983−),女,工程师,主要研究领域为电子产品数据资源建设与应用.

李川(1977−),男,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘,信息网络数据分析,社会网络分析,生物信息学.

杨宁(1974−),男,博士,讲师,CCF 专业会员,主要研究领域为时空数据挖掘,时态序列挖掘,异构信息网络挖掘,网络上的信息处理.

唐常杰(1946−),男,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库系统,数据挖掘和数据仓库,知识工程,计算机安全.