

基于 DBLP 数据的多维异质网络 Graph OLAP 设计与实现

邵连龙, 尹 沐

(北京邮电大学 北京市智能通信软件与多媒体重点实验室, 北京 100876)

摘 要: 为了分析 DBLP 数据中的多种类型的实体信息,挖掘其中特定的知识,首先根据异质网络 Graph OLAP (图联机分析处理)模型,建立相应的数据仓库模型;然后依据实体维的概念,构建多维异质图立方模型;最后针对 Graph OLAP 处理异质网络能力不足的问题,补充了旋转和拉伸操作,并完善了 Graph OLAP 原型系统 Liter Miner。实证表明设计的原型系统可以有效地对 DBLP 数据中的多维异质网络进行分析,挖掘出研究人员需要的知识。

关键词: 图挖掘; 图联机分析处理; 数据仓库; 图立方体

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2014)03-0720-05

doi:10.3969/j.issn.1001-3695.2014.03.019

Heterogeneous multi-dimensional Graph OLAP design and implementation based on DBLP

SHAO Lian-long, YIN Mu

(Beijing Key Laboratory of Intelligent Telecommunications Software & Multimedia, Beijing University of Posts & Telecommunications, Beijing 100876, China)

Abstract: When we want to analyze the entity information and mine the necessary knowledge, according to the heterogeneous multi-dimensional Graph OLAP (graph online analytical processing) framework, firstly this paper modeled the heterogeneous multi-dimensional Graph OLAP data warehouse, then modeled the heterogeneous multi-dimensional graph cube to adapt to entity dimensional analysis. Finally it added rotate and stretch operations to solve the problem of Graph OLAP's poor ability towards heterogeneous network, and implemented heterogeneous multi-dimensional Graph OLAP prototype, Liter Miner. It is proves effective on analyzing heterogeneous networks in DBLP data and mining the knowledge that is needed by researchers.

Key words: graph mining; OLAP; data warehouse; graph cube

0 引言

1993 年关系数据库之父 Codd 提出联机分析处理 (OLAP) 的概念,它成为继联机事务处理 (OLTP) 之后,数据处理及相关数据库领域的热门问题。十多年来从理论研究到实用工具、产品及商业应用取得很大发展,OLAP 已经逐渐成为一项成熟的技术。

随着对于 OLAP 技术的深入研究,可以看到传统的 OLAP 技术无法支持带有图结构的网络分析。直到 2007 年,吴巍^[1]提出了 Link OLAP 的概念,通过将面向实体的分析扩展为面向连接的分析,他发现 Link OLAP 操作在某些特定的分析场景下能提供比传统 OLAP 更优秀的解决方案。他以复杂网络可视化为基础,将 OLAP 技术应用到面向连接的分析,突破了以往传统 OLAP 系统中单调的二维表格表现方式。同年,Chen 等人^[2,3]也提出了基于图结构的 OLAP,即 Graph OLAP 的概念,将 OLAP 技术引入到复杂网络的分析中,解决含有多维度特性的复杂网络的多维分析问题。他提出了 Graph OLAP 特有的两种维度,即信息维、拓扑维,并定义了两个维度上的 OLAP 操

作,为 Graph OLAP 的研究打下了理论基础,开辟了复杂网络多维分析的一个新方向。但是最初的研究并未对 Graph OLAP 的数据仓库模型以及算法设计进行详细考虑,随后一些研究者对这些内容进行了补充。2010 年, Li 等人^[4]提出了一种适合 Graph OLAP 的数据仓库概念模型,即双星模型,并提出了信息维聚集算法 I-OLAPing 和拓扑维聚集算法 T-OLAPing,实现了 Graph OLAP 的原型系统 GraphOLAPer1.0。次年, Li 等人^[5]又在原有基础上提出了基于信息网络数据仓库和信息网络数据立方体的概念,即 InfoNetOLAPer,提出了双星座数据模型,实现了信息维和拓扑维的聚集算法以及上卷下钻的 OLAP 操作,在操作执行方面具有较高的性能。2011 年, Zhao 等人^[6]详细介绍了一个新的数据仓库模型,即基于图的数据立方体 Graph Cube,有效地支持针对大型多维网络进行 OLAP 查询。同时提出了用于 Graph OLAP 的新的查询方式 crossboid,这种查询方式针对多维网络进行交叉查询,并讨论了 Graph Cube 的物化策略。2011 年, Qu 等人^[7]提出了一种信息网络拓扑维的框架,并基于此框架提出了更高效的查询方法以及数据立方体的物化策略,对拓扑维在线分析处理 (T-OLAP) 操作中特定类型度量的优化进行

收稿日期: 2013-07-02; 修回日期: 2013-08-29

作者简介: 邵连龙 (1987-), 男, 天津人, 硕士研究生, 主要研究方向为数据挖掘 (along2506@126.com); 尹沐 (1988-), 男, 北京人, 硕士研究生, 主要研究方向为数据挖掘。

了有针对性的深入分析。2012年, Sun等人^[8]提出一种整合元路径选择和用户指导聚类的信息网络分析方法。2013年, Jaka-wat等人^[9]提出一种应用于文献数据的框架, 分析网络节点的动态演化。同年, Han等人^[10]提出一种挖掘半结构化的信息网络模型, 适用于大规模异质信息网络分析。

目前, 已提出的 Graph OLAP 概念模型偏重于分析同质网络, 对于异质网络分析涉及较少, 采用异质网络建模可以更加完整地包含对象之间的关联信息, 蕴涵更加真实的知识。因此本文提出了面向多维异质网络分析的异质网络 Graph OLAP 概念, 它综合了 OLAP 技术和图挖掘技术的优点, 并区别于传统的 Graph OLAP 技术, 尤其适用于以社会网络为代表的复杂异质多维网络的分析。

1 问题描述

现实社会中的网络往往是由各种不同类型的节点和边构成的异质网络。异质信息网络不仅更加符合现实世界, 而且包含更加丰富的信息, 因此异质信息网络分析有潜力发现更加准确的隐含知识。为了更加准确生动地描述异质网络 Graph OLAP 的模型和相关概念, 本文将以文献数据作为背景进行阐述。文献数据除了大量的关系存在之外, 还拥有复杂的维度特征, 包含了如作者合作网络、机构合作网络、文献引用网络等丰富的信息网络, 是异质网络 Graph OLAP 技术的重要应用场景。如图1所示, 文献数据中包含论文、作者、机构等多种实体。通过论文实体相连形成了一个多维的异质网络。网络中不但含有实体, 还有实体间复杂的关系, 如作者合作关系、论文共同引文关系、关键词共现关系等, 各实体本身还具有多种属性, 论文有检索类型属性、被引次数属性, 作者含有姓名、年龄、性别等属性。虽然已有研究者提出多种图联机分析处理模型^[2,7], 但是更多的是对异质网络中提取的同质网络(主要为简单人际关系网络, 如作者合作网络)进行分析, 对于这种多维异质的复杂网络仍无法很好地表达。

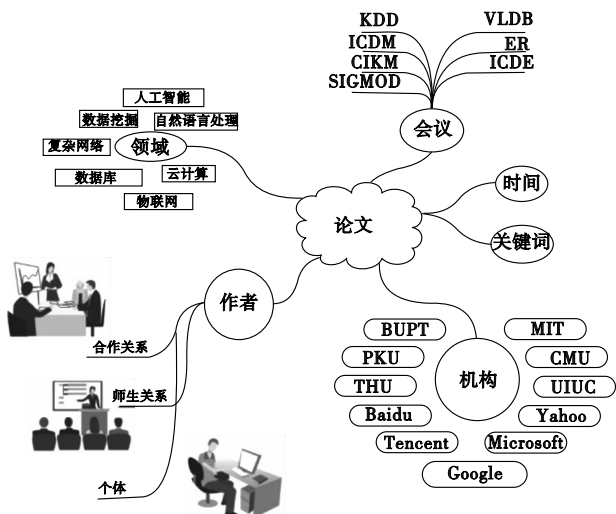


图1 文献数据中的多维网络

定义1 多维网络。一个多维网络 $N = (N, E, V_E, V_R)$ 。其中: $V = \{V_1, V_2, \dots, V_n\}$, V_i 是每一类节点实体的集合, 代表一类实体; $E = \{E_{ij} | E_{ij} = V_i \times V_j, 1 \leq i, j \leq n\}$, 是实体 V_i 与 V_j 关系的集合。当 $n = 1$ 时, N 为多维同质网络, 此时 $i = j$; 当 $n > 1$ 时, N 为多维异质网络, 网络节点代表多种实体, 即相同类实体之间没有关联关系。多维异质网络又分为两种, 若 $\forall E_{ij} \in E, i \neq j$, 则 N 为简单异质网络, 即网络中相同类型的实体节点之

间没有边相连; 否则, 若 $\exists E_{ij} \in E, i = j$, 则 N 成为复合异质网络。复合异质网络都可以拆分成一个简单异质网络和一个同质网络。

$V_E = \{V_{E1}, V_{E2}, \dots, V_{En}\}$, $V_{Ei} = \{V_{Ei1}, V_{Ei2}, \dots, V_{Eik}\}$ 是节点 V_i 集上一系列节点的属性。 V_E 是网络 N 的实体维度信息。 V_{Eik} 代表属于实体 E_i 的节点 k 的各个维度信息。 $V_E = \{V_{Eij}\}$ 是关系集 E 上的维度属性, $\{V_{Eij}\}$ 是实体集合 V_i 与 V_j 之间的关联集合 E_{ij} 的关系维度集合。 V_{Eij} 是节点 i 和 j 的关系维度。

定义2 拓扑维。假设当前分析的图结构为 $G = (V, E, V_E, V_R)$, V 是节点实体的集合, E 为关系集合, V_E 代表实体维度信息, V_R 代表实体关系的维度信息。假设操作 $TD = \{T_1, T_2, \dots, T_n\}$ 是描述图 G 中某一类拓扑结构的集合。设 α 为实体拓扑维度的选择函数, β 为实体关系拓扑维度的选择函数, 则图 G 表示为 $G = (V', E', \alpha(V_E), \beta(V_R))$ 。称集合 TD 为拓扑维集合, 选择函数 α, β 决定了图 G 的拓扑结构, 异质网络 Graph OLAP 可以根据拓扑维集合在中心度量拓扑结构上进行变换。

定义3 信息维。设操作 $ID = \{I_1, I_2, \dots, I_n\}$ 是描述图 G 中某一类信息的集合, 设 γ 为实体信息维度的选择函数, δ 为实体关系信息维度的选择函数, 则图 G 表示为 $G = (V, E', \gamma(V_E), \delta(V_R))$ 。称集合 ID 为信息维集合, 选择函数 γ, δ 决定了图 G 的信息范围, 异质网络 Graph OLAP 可以根据集合 ID 在不同层次信息上进行变换。经过信息维变换得到了图 G 一个网络快照, 即在边的维度属性上限定不同的值得到的网络。

定义4 实体维。设操作 $ED = \{E_1, E_2, \dots, E_n\}$ 是描述图 G 中某一些实体的集合, 设 A 是图 G 的一个聚合子图, $A = \{V', E', V'_E, V'_R\}$, 通过实体关系转换函数 θ , 聚合子图 A 可以表示为 $A = \{\theta(V'), \theta(E'), V'_R, V'_E\}$ 。称集合 ED 为实体维集合, 通过实体维变换可以使异质网络 Graph OLAP 能够在特定实体与关系之间进行转换。根据实体维的特点本文定义了实体维的两种操作:

a) 旋转。将具有可实体化关系的同质网络中的实体和关系相互转换。如图2所示, 作者合作网的本质是一个实体为作者和关系为文献的同质网络, 因此, 可以通过实体维的旋转操作, 使网络变成实体为文献、关系为作者的同质网络。现实中节点和边往往都是由实体构成且可以相互转换。当其属性符合某类具体实体的特征时, 它们就可以转换为此类具体的实体, 如图2所示。

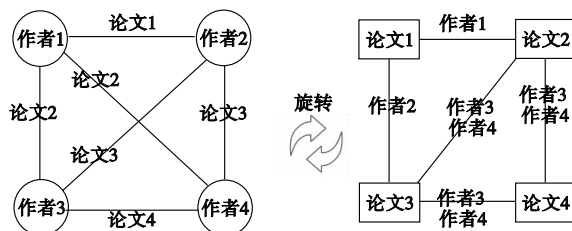


图2 实体维旋转操作

b) 拉伸。将同质网络中可实体化的关系进行实体化操作, 使该同质网络变为由关系和实体组成的异构网络。如图3所示, 拉伸操作将原来表示关系的文献实体化, 变成网络中的新实体。图中的关系随之变成了作者与文献之间的写作关系, 如图3所示。

定义5 关系路径。关系路径是以实体类型为元素组成的一个有序排列。设网络 $N = (V, E)$, 其中 $V = \{V_1, V_2, \dots, V_n\}$, V_i

是每一类节点实体的集合,代表一类实体。 $E = \{E_{ij} | E_{ij} = V_i \times V_j, 1 \leq i, j \leq n\}$, 是实体 V_i 与 V_j 的关系集合。网络中的所有关系都不可实体化。设 $\text{Entity}(V_i)$ 是 V_i 集中实体的类型。 $\text{Entity}(V_i)$ 与 $\text{Entity}(V_j)$ 的关系路径定义为 $\text{Route}(\text{Entity}(V_i), \text{Entity}(V_j)) = \text{Entity}(V_i) \sim \text{Entity}(V_m) \sim \dots \sim \text{Entity}(V_n) \sim \text{Entity}(V_j)$, 其中 $E_i, E_m, \dots, E_n, E_j \in E$ 。关系路径代表了实体间关系的意义,是实体维上的网络标志,以区分不同意义的网络。作者合作网络的关系路径 $\text{Route}(\text{作者}, \text{作者}) = \text{作者} \sim \text{文献} \sim \text{作者}$, 而关键词贡献网络的关系路径 $\text{Route}(\text{关键词}, \text{关键词}) = \text{关键词} \sim \text{文献} \sim \text{关键词}$ 。若是不区分实体维的意义,只有图的拓扑结构,则当两种类型的节点和边完全一样,且节点和边所拥有的维度类型也一样时,就无法区分这两个网络是否为同一个网络。

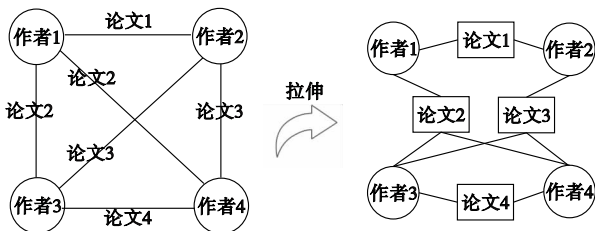


图3 实体维拉伸操作

定义1~5的引入,扩展了图联机分析处理的概念,使异质网络 Graph OLAP 可以更全面地分析多维异构网络信息,对异质网络进行拓扑维、信息维以及实体维的变换;实体维度概念以及旋转和拉伸操作的提出,丰富了原有的 Graph OLAP 操作,增加了异质网络分析的角度;另外,通过关系路径可以更好地理解网络节点和关系的具体意义。

2 异质网络 Graph OLAP 数据仓库模型

区别于传统 OLAP 中的分布式度量值,异质网络 Graph OLAP 侧重于研究不同实体间的关联意义,且异质网络 Graph OLAP 中包含了信息维、拓扑维以及实体维的多种操作,因此本文所提出的数据仓库模型更注重体现实体和关联的异质性。基于多维网络定义提出的数据仓库概念模型如图4所示。

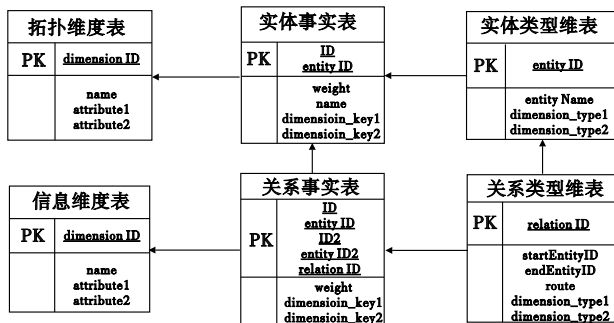


图4 异质网络Graph OLAP数据仓库模型

图4中有两种事实表,分别为实体事实表和关系事实表(即节点事实表和边事实表)。维度表有两种,第一种是实体、关系的维度表,包含信息维度表和拓扑维度表;第二种是实体、关系的类型维表,包含实体类型维表和关系类型维表。

实体事实表包含了所有的实体类型,主键为实体 ID、实体类型 ID (entityID)、属性包含节点名称 (name)、节点权重 (weight)、节点的维度值 (dimension_key)。实体类型 ID 作为本表外键连接了实体类型维表。

关系事实表包含了全部的关系,即网络中的边信息。它的

主键是边两端的实体 ID 和实体类型 ID,以及关系类型 ID (relationID)。关系类型 ID 同时作为外键关联了关系类型维表。属性包含关系权重 (weight)、关系的维度值 (dimension_key)。

拓扑维度表用来记录实体的各种维度属性,主键为维度 ID (dimension_ID),属性为当前拓扑维度的各种属性值。信息维表是记录关系的各种维度属性,主键为维度 ID (dimension_ID),属性为当前信息维度的各种属性值。

实体类型维表包含了所有的实体类型,主键为实体类型 ID (entityID),属性为实体的类型名称 (entityName) 和实体对应的维度类型 (dimension_type)。

关系类型维表包含了所有的关系类型,主键为关系类型 ID (relationID),属性包含关系的起始实体类型 ID (startEntityID)、终止实体类型 ID (endEntityID)、关系路径 (route)、关系对应的维度类型 (dimension_type)。关系类型维表可通过起始和终止 ID 关联到实体类型维表。实体类型维表和关系类型维表合起来组成了实体维表。

3 多维异质图立方

给定一个多维网络 $N = (V, E, V_A, V_B)$, 一个多维异质图立方包含了 N 在实体节点维度 DV 上的所有聚合网络。维度总数 $|DV| = \sum_{i=1}^n |DV_i|$ 。每个聚合网络 DV' 可以叫做一个子立方体 (cuboid)。一个多维网络 N 有 $2^{|DV|}$ 个子立方体,每个子立方体的维度 $\dim(DV')$ 是 DV 中维度的个数。另外, DV' 中属于不同实体类别的相同维度名称算做不同的维度。若子立方体的 $\dim(DV') = DV$, 则称这个子立方体为基本子立方体 (base cuboid); 若子立方体的 $\dim(DV') = \{\}$, 则称这个子立方体为顶点子立方体 (apex cuboid)。

多维异质图立方如图5所示。节点类型有文献和关键词,文献维度有领域和检索类型,关键词维度有领域。不同实体类型的维度名即使相同,也应算做不同的维度。因此,这个网络的拓扑维度图立方共有三个节点维度,分别为文献领域 (article_field)、文献索引类型 (article_index) 和关键词领域 (key_field), 共 $2^3 = 8$ 个子立方体。图立方上的每个子立方体都代表了一个或一类聚合网络,同时也代表了一个层次的网络视图。在各个子立方体之间的变换,改变了图的拓扑结构,也改变了图中节点和边的意义。本文基于多维异质网络图立方,重新定义了拓扑维 OLAP 操作,用于进行维度层次的转换,从而审视不同维度层次上的网络视图。

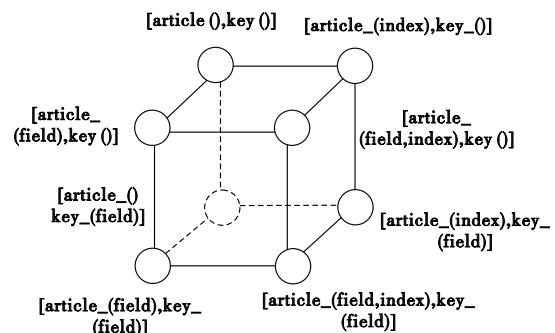


图5 多维异质图立方

4 优化

构建异质网络 Graph OLAP 多维异质图立方能够为各维度提供高效的查询操作,主要关注操作的响应时间。本文针对这个

中还提供了重名处理、文献去重等数据清洗功能,在可视化展示方面集成了多种布点算法以及网络参数计算功能,进一步丰富了原型系统的分析手段。本文的实验环境是:a) CPU 为 Intel Pentium 4 CPU 3.00 GHz;b) 内存为 2 GB;c) 操作系统为 Microsoft Windows XP Professional;d) 编程语言为 Eclipse Java。

表 3 DBLP 文献数据集

数据集	节点	边	数据集	节点	边
1	44	177	5	1 988	52 684
2	209	1 631	6	4 056	97 813
3	540	7 875	7	4 999	119 667
4	1 049	34 422			

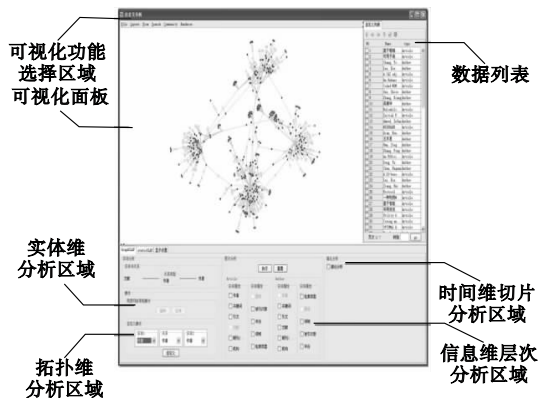


图8 异质网络Graph OLAP 原型系统

5.3 实验分析

本文采用的物化策略为顺便物化策略。在操作过程中,将已生成的子立方体和所需的中间子立方体保存。以拓扑维操作为例,使用拓扑维下钻从顶层网络开始,逐层向下下钻,共钻取四层。这样的操作能够适应用户最复杂的查询需求。

图 9 是两种策略的时间占用情况,其中横坐标为节点数,纵坐标为操作时间(单位 ms)。可以看到拓扑维操作的时间耗费大致呈线性增长,而不物化的策略增长要比有物化策略的时间增长速度快。图 10 是两种策略的内存耗费情况,其中横坐标为节点数,纵坐标为占用内存空间(单位 MB)。两种情况的内存占用在节点值为 500 ~ 1000 时的耗费速度增加较快,因为内存占用与图的大小成正相关,本例中 1 049 个节点有 34 422 条边,比其他图的平均边数要大,因此内存占用也较多,而部分物化的内存占用情况和增长幅度均比不物化时要大。

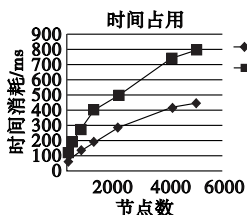


图9 拓扑维逐层下钻操作时间效率

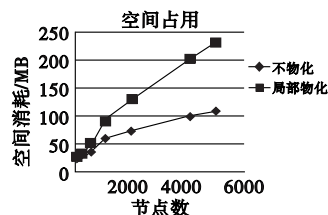


图10 拓扑维逐层下钻操作空间效率

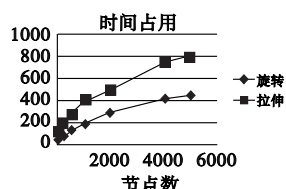


图11 旋转拉伸操作时间效率

图 11 是旋转和拉伸两种实体维操作的时间占用情况对

比,其中横坐标为节点数,纵坐标为操作时间(单位 ms)。可以看出,随着节点数的增多,旋转和拉伸操作的时间消耗近似线性增长,而且拉伸操作要比旋转操作消耗更多的时间。

6 结束语

图联机分析处理是近几年刚刚兴起的研究领域,与传统的 OLAP 技术不同,它侧重于对图数据进行多角度多粒度的分析。本文基于文献异质网络提出了异质网络 Graph OLAP 概念,设计了异质网络数据仓库和数据立方体模型,首次提出了实体维概念以及实体维的旋转、拉伸操作,提出了异质网络 Graph OLAP 的位图索引和物化策略,并将上述功能集成在原型系统 LiterMiner 中。未来,笔者的工作重点是:a) 改进异质网络 Graph OLAP 算法,进行优化设计,提升原型系统性能;b) 异质网络 Graph OLAP 操作研究,目前,异质网络 Graph OLAP 的分析方法较为单一,将丰富的图挖掘算法与异质网络 Graph OLAP 技术相结合是解决这一问题的有效方法;d) 异质网络 Graph OLAP 粒度研究,主要包括粒度选择与粒度存储,当粒度划分较为合理时,查询算法以及存储效率会大大提高。

参考文献:

- [1] 吴巍. 复杂网络可视化与 Link OLAP[D]. 北京: 北京邮电大学, 2007.
- [2] CHEN Chen, YAN X, ZHU Fei-da, *et al.* Graph OLAP: towards on-line analytical processing on graphs[C]//Proc of the 8th IEEE International Conference on Data Mining. [S. l.]: IEEE Press, 2008: 103-112.
- [3] CHEN Chen, YAN X, ZHU Fei-da, *et al.* Graph OLAP: a multi-dimensional framework for graph data analysis[J]. *Knowledge and Information Systems*, 2009, 21(1): 41-63.
- [4] LI Chuan, ZHAO Lei, TANG Chang-jie, *et al.* Modeling, design and implementation of graph olaping[J]. *Journal of Software*, 2011, 22(2): 258-268.
- [5] LI Chuan, YU P S, ZHAO Lei, *et al.* InfoNetOLAPer: integrating InfoNetWarehouse and InfoNetCube with InfoNetOLAP[J]. *Proceedings of the VLDB Endowment*, 2011, 4(12): 1422-1425.
- [6] ZHAO Pei-xiang, LI Xiao-lei, XIN Dong, *et al.* Graph cube: on warehousing and OLAP multidimensional networks[C]//Proc of ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM Press, 2011: 853-864.
- [7] QU Qiang, ZHU Fei-da, YAN Xi-feng, *et al.* Efficient topological OLAP on information networks[C]//Proc of the 16th International Conference on Database Systems for Advanced Applications. Berlin: Springer-Verlag, 2011: 389-403.
- [8] SUN Yi-zhou, NORICK B, HAN Jia-wei, *et al.* Integrating meta-path selection with user-guided object clustering in heterogeneous information networks[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: ACM Press, 2012: 1348-1356.
- [9] JAKAWAT W, FAVRE C, LOUDCHER S. OLAP on information networks: a new framework for dealing with bibliographic data[C]//Proc of East-European Conference on Advances in Databases and Information Systems. 2013: 361-370.
- [10] HAN Jia-wei, ZHAI Cheng-xing, ROTH D, *et al.* Mining heterogeneous information networks[C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013.