基于图形数据库的 DBLP 数据存储

王余蓝
(西安交通大学,陕西 西安 710049)

摘 要:DBLP 数据不但数据规模大,而且数据之间存在多种类型的关联关系,包括合作关系、写作关系、引用关系等。采用关系数据存储时不但存在大量的数据冗余,并且难于动态更新,为解决复杂关联关系的高效存储与动态更新问题,提出一种基于图形数据库的 DBLP 数据表示与存储方法,论文、作者、书籍等实体以节点存储,而实体间的各种关系以多类型的边存储。实验表明该方法能够有效支持关联关系的动态增删、多阶查询、深度遍历、广度遍历等操作,有效解决了复杂关联关系的数据存储问题。

关键词:图形数据库;关系数据库;扩展性;NEO4J
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)08-0018-03
doi:10.3969/j.issn.1673-629X.2013.08.005

DBLP Data Storage Based on Graphics Database

WANG Yu-lan
(Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: DBLP data has large data scale and includes complex relationships including co-authorship relationships, writing relationships, citation relationships and so on. Relational database will lead to the data redundancy and is difficult to update. Propose a DBLP method of data presentation and storage based on graphics database to solve the problem of efficient storage and dynamic updating of complex relationship. Papers, authors, books and other entities are stored by node, and the various relationships between the entities are stored by edge. The experimental results show that this method can effectively support the dynamic additions and deletions, multi-step queries, depth traversal, and breadth traversal operation. It is an effective solution for the complex relationship data storage.

Key words: graphical database; relational database; extension; NEO4J

0 引 言

DBLP 全称 Digital Bibliography & Library Project[1],是由德国特里尔大学开发的计算机领域科学文献搜索服务,DBLP 没有采用数据库来存储数据[2],而是使用 XML 文档来存储元数据。DBLP 所有的数据记录都存储在一个名为 dblp.xml 的文档中,虽然 XML 作为数据载体能够有效表示数据关联,但随着文件规模的增大,XML 效率将受到严重制约。

Neo4J 是一种支持亿级节点规模的图形数据库[3],用来专门存储关联性复杂的数据(如社交网络中人物关系)[4~8]。它的内在索引机制与查询优化策略能够有效解决 XML 的大数据问题,同时也克服了传统关系数据库动态更新能力弱、无法有效处理复杂关系的缺点。因此提出一种基于图形数据库的学术合作关系管理方法,对 DBLP 数据库进行解析与存储[9,10],主要从 DBLP 数据格式分析、数据解析、Neo4J 数据库设计、查询应用几个方面展开。

1 DBLP 数据结构分析

dblp.xml 文件包含 DBLP 系统中的所有记录,它使用 dblp.dtd 进行语义约束,dblp.xml 文件一般由如下结构组成:

```
<? xml version="1.0" encording="ISO-8859-1"? >
<! DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
record 1
….
record n
</dblp>
```

收稿日期:2012-11-19 **修回日期**:2013-02-21 **网络出版时间**:2013-04-22
基金项目:国家自然科学基金资助项目(61100166);陕西省教育科技专项(11JK1035)
作者简介:王余蓝(1963-),女,工程师,研究方向为信息系统、实验室管理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130422.1721.024.html

整个 dblp. xml 文件根标签为<dblp></dblp>,其他一级标签分别为 article, inproceedings, proceedings, book, incollection, phdthesis, masterthesis, www。下面仅就关键性标签 article 进行重点分析。Article 是出现在 journal(期刊)中 paper(论文)的记录。一般格式为:

```
<article mdate="2007-04-18" key="journals/csda/PaolettiOM04">
<author>Xavier Paoletti</author>
<author>John O'Quigley</author>
<author>Jean Maccario</author>
<title>Design efficiency in dose finding studies. </title>
<pages>197-214</pages>
<year>2004</year>
<volume>45</volume>
<journal>Computational Statistics &amp;amp; Data Analysis</journal>
<number>2</number>
<ee>http://dx. doi. org/10. 1016/S0167-9473(02)00323-7</ee>
<url>db/journals/csda/csda45. html#PaolettiOM04</url>
</article>
```

重要属性:

key:唯一标识该记录,一般格式:journals/期刊的缩写/字母数字混编(姓名缩写+发表时间),eg: key="journals/csda/PaolettiOM04"。key 值一般与记录内容无关,一旦设定后不会改变。

author:表示作者姓名,基本格式为: first name + lastname。当一篇论文有多个作者时,列出全部作者,每个作者占用一个 author 标签,多个作者排列的先后顺序是有一定意义的。如果某个记录的作者不确定,则无 author 标签,也就是说 author 元素不是必须的。

title:论文题目,该元素必须存在。该标签可能存在以下子标签:字符串, sub, sup, i, tt, ref。

journal:期刊名称。

years:paper 发表的年份,格式为 4 个数字。发表在 journal 上的 paper 的年份一般是确定的。在学术会议中发表的论文年份比较难以确定,因为会议召开时间和会议论文集出版的时间可能不一致。

在 article 和 inproceedings 中可能会含有"crossref"标签,通过该标签可以把 article, inproceedings 与 proceedings 关联起来。crossref 的值与 proceedings 的 key 值对应,即通过 crossref 可以找到收录该 paper 的论文集,crossref 相当于关系数据库中的外键。如:

```
<inproceedings key="conf/naa/Xiang08" …>
<title>Numerical Quadrature …</title>
<year>2008</year>
<booktitle>NAA</booktitle>
<crossref>conf/naa/2008</crossref>
…
</inproceedings>
<proceedings key="conf/naa/2008" …>
<title>NAA 2008, Lozenetz, …</title>
<year>2009</year>
<publisher>Springer</publisher>
…
</proceedings>
```

通过对 dblp. xml 与 dblp. dtd 文件的分析,可以得出如图 1 所示的数据关系,通过该图可以清晰地看出 DBLP 的数据记录以及各数据记录间的关系。

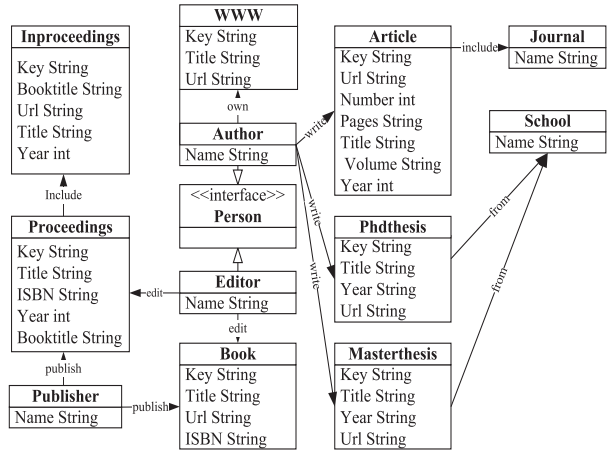

图 1　DBLP 数据中各类关系的 UML 图

2　DBLP 数据解析

目前在处理 XML 文档的问题上,有两套比较流行的机制:SAX(Simple API for XML)和 DOM(Document Object Model)。DOM 机制简单易用,但需要整体加载 XML 文档,导致系统开销过大而内存溢出。针对该问题,SAX 采用事件驱动机制来解析 XML 文档,每当发现文档开始、元素开始、元素结束、文档结束时,就会向外发送事件,通过编写事件监听程序来获取 XML 文档里的信息。SAX 方式占用内存小,速度快。通过综合考虑与分析,最终选用 SAX 机制来处理 dblp. xml 文件,这是由于 dblp. xml 文档比较大,最新的有 997MB。dblp. xml 解析工具的核心代码如下:

```
SAXParserFactory fac = SAXParserFactory. newInstance();//创建 SAX 解析工厂
fac. setValidating(true); //启用 dtd 验证
SAXParser parser = fac. newSAXParser();//得到 SAX 解析器实例
parser. parse(xmlFileName, new MyHandler());//开始解析
…
class MyHandler extends DefaultHandler {…}//处理事件监听器
```

其中 xmlFileName 即为 dblp. xml 的完整路径名。该解析工具的关键部分就是重写 DefaultHandler 类的

一些方法,这些方法是 SAX 机制处理 XML 文档的关键。由于 dblp. xml 文件比较庞大且相当复杂,所以在解析中应注意以下问题:

1)如果启用了 DTD 验证 XML 文档的有效性,则 dblp. dtd 文件必须与 dblp. xml 文件在同一路径下。

2)dblp. xml 文件较大,在解析时需消耗大量内存,需要设置 Java 虚拟机 JVM 与内存相关的参数,增大 Heap Size 值,否则可能会出现 java. lang. OutOfMemoryError 的错误,参考值为-Xms500M - Xmx900M。

3)为了提高解析速度及效率,在编码时应注意,尽量减少不必要对象的创建,减少内存的消耗,及时释放不用的对象。

4)使用 SAX 解析 XML 需要使用 DefaultHandler 的方法 characters()来取元素之间的字符串内容,但是可能出现取值不完整的问题,需要特别注意。出现这种情况的原因可能是 Xerces-J 项目固有的 bug。

3 使用 Neo4J 图数据库存储与检索 DBLP 数据集

在解析 DBLP 数据集的基础之上,使用 Neo4J 作为持久化引擎,开发出计算机科学学术论文检索系统。该系统为 B/S 架构,遵循 MVC 开发模式。DBLP 数据集在 Neo4J 中的数据模型如图 2 所示。

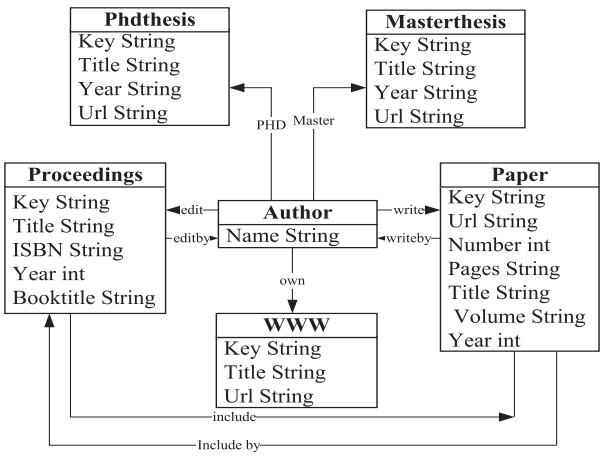

图 2 DBLP 在 Neo4J 图数据库中的存储模型

为了提高查询效率,在常用的查询字段作者名字、论文标题、论文集标题等上分别建立全文索引,支持三种查询模式:按标题、按作者、按论文集。图 3 是按作者姓名搜索时的输入界面。

图 3 搜索页面

图 4 是输入姓名"E. F. Codd"时的查询结果,利用模糊匹配技术找到含有"E. F. Codd"的作者姓名。图 5 是用户点击作者姓名"E. F. Codd"后最终的查询结果。在检索论文的同时,系统还检出与当前作者合作过的其他作者并加以显示(图 5 底部),实验表明 Neo4J 具有良好的整体查询效率。

图 4 搜索结果页面

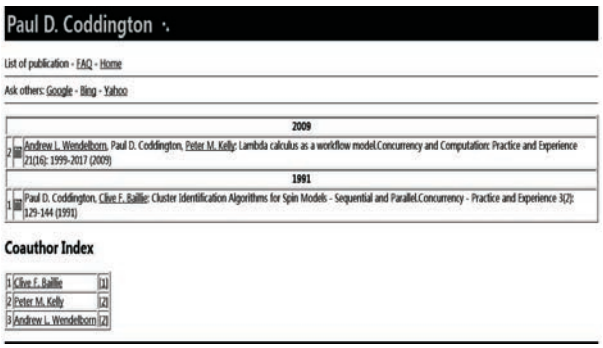

图 5 搜索结果

4 结束语

图形数据库善于处理大量复杂、互连接、低结构化的数据,当使用图形数据库 Neo4J 对 DBLP 这种复杂的图状结构数据进行存储时,可以高效地支持研究热点分析、中心作者发现等数据分析服务。系统实现了 DBLP 格式分析、XML 数据解析、数据迁移、信息检索等功能,但在解析 DBLP 数据时,遇到主键并不唯一、人名解析乱码等问题,尚需在下一步工作中完善。

参考文献:

[1] Elmacioglu E, Lee D. On Six Degrees of Separation in DBLP-DB and More[J]. ACMSIGMOD Record, 2005, 34(2): 33-40.

[2] 张 辉,赵郁亮,徐 江,等. 基于 Oracle 数据库海量数据的查询优化研究[J]. 计算机技术与发展,2012,22(2):165-167.

[3] Neo4j Technology, The Neo4j Manual v1.7[DB/OL]. 2012. http://neotechnology.com/.

[4] Angles R, Gutierrez C. Survey of graph database models[J]. ACM Computing Surveys (CSUR), 2008, 40(1): 1-6.

[5] 陈 锐. 网络结构与数据库模式在陕西科技信息网中的应用[J]. 情报杂志,2004,23(10):57-58.

[6] 乔秀全,杨 春,李晓峰,等. 社交网络服务中一种基于用户上下文的信任度计算方法[J]. 计算机学报,2011,34(12):2403-2412.

[7] 付晓燕. 社交网络服务对使用者社会资本的影响-社会资本视角下的 SNS 使用行为分析[J]. 情报杂志,2010(4):20-22.

(下转第 26 页)

```
R3(config)#router bgp 200
R3(config-router)# address-family ipv6
R3(config-router-af)#neighbor 200::1 activate
R3(config-router-af)#network 300::/64
R3(config-router-af)#network 3::1/64
R3(config-router-af)#redistribute ospf 1
R3(config-router-af)#no synchronization
R3(config)#ipv6 router ospf 1
R3(config-rtr)#redistribute bgp 200
```

(3)BGP4+运行状态。

```
R2#show bgp ipv6 unicast neighbors
BGP neighbor is 200::2,  remote AS 200, external link
  BGP version 4, remote router ID3.3.3.3
  BGP state = Established, up for 02:18:54
Last read 00:00:53, last write 00:00:53, hold time is 180, keepalive
interval is 60 seconds
```

信息显示,BGP 状态是 Established,说明 R2 和 R3 之间建立了邻居关系,路由信息交换成功。

(4)IPv6 路由表。

```
R1#show ipv6 route rip
IPv6 Routing Table - 9 entries
R   2::1/128 [120/2]
      via FE80::CE0C:9FF:FE6C:0, FastEthernet0/0
R   3::1/128 [120/2]
      via FE80::CE0C:9FF:FE6C:0, FastEthernet0/0
R   4::1/128 [120/2]
      via FE80::CE0C:9FF:FE6C:0, FastEthernet0/0
R   300::/64 [120/2]
      via FE80::CE0C:9FF:FE6C:0, FastEthernet0/0
R2#show ipv6 route bgp
IPv6 Routing Table - 11 entries
B   3::1/128 [20/0]
      via 200::2
B   4::1/128 [20/1]
      via 200::2
B   300::/64 [20/0]
      via 200::2
R3#show ipv6 route bgp
IPv6 Routing Table - 11 entries
B   1::1/128 [20/2]
      via 200::1
B   2::1/128 [20/0]
      via 200::1
B   100::/64 [20/0]
      via 200::1
R4#show ipv6 route ospf
IPv6 Routing Table - 9 entries
OE2  1::1/128 [110/2]
      via FE80::CE0D:9FF:FE6C:1, FastEthernet0/0
OE2  2::1/128 [110/1]
      via FE80::CE0D:9FF:FE6C:1, FastEthernet0/0
O   3::1/128 [110/1]
      via FE80::CE0D:9FF:FE6C:1, FastEthernet0/0
OE2  100::/64 [110/1]
      via FE80::CE0D:9FF:FE6C:1, FastEthernet0/0
```

信息显示,R1、R2、R3 和 R4 都同步到了本地域内路由和远端的域内路由。

3 结束语

文中讨论了常用 IPv6 单播路由协议的新特性,开展仿真实验模拟路由协议在不同网络场景中的组网过程,分析了路由协议的运行状态,验证了路由交换信息。下一步,将继续研究 IPv6 单播路由协议的传输容量、服务质量以及安全问题。

参考文献:

[1] 杨惠仁,吕 波,谢晓尧. IPv6 驻地网部署方案研究[J]. 计算机技术与发展,2007,17(11):60-62.

[2] Malkin G,Minnear R. RIPng for IPv6[S]. RFC 2080,1997.

[3] Coltun R,Ferguson D,Moy J. OSPF for IPv6[S]. RFC 2740,1999.

[4] Gupta M,Melam N. Authentication/Confidentiality for OSPFv3[S]. RFC 4552,2006.

[5] Hopps C. Routing IPv6 with IS-IS[S]. RFC 5308,2008.

[6] Bates T,Rekhter Y,Chandra R,et al. Multiprotocol Extensions for BGP-4[S]. RFC 2858,2000.

[7] 杨 闽. 用于 IPv6 的 RIPng 的研究[D]. 天津:天津大学,2005.

[8] 王宇杰,王 锋. OSPFv3 及其在高校下一代校园网建设中的应用[J]. 计算机应用,2002(5):72-73.

[9] 吴许俊,朱长水,王 巍. IPv6 网络 OSPFv3 路由协议的研究与仿真[J]. 电子设计工程,2012,20(13):71-75.

[10] 赵玉兰,张弘宇,冀 超,等. IS-IS 路由协议互操作性测试的研究[J]. 计算机科学,2012(s1):146-150.

(上接第 20 页)

[8] 于婷婷,窦光华. 社交网络服务兴起的社会学意义[J]. 当代传播,2011(6):55-57.

[9] Tong H,Gallagher B,Faloutsos C,et al. Fast Best-effort Pattern Matching in Large Attributed Graphs[C]//Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,NY,USA:ACM,2007:737-746.

[10] Rahm E,Thor A. Citation analysis of database publications[J]. SIGMOD Record,2005,34(4):48-53.

基于图形数据库的DBLP数据存储

作者： 王余蓝， WANG Yu-lan
作者单位： 西安交通大学,陕西 西安,710049
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013(8)