



专题:大数据

基于图数据库的 IPTV 内容知识库设计和实现

蔡鑫¹, 阮宜龙², 石屹嵘¹

(1. 中国电信股份有限公司上海研究院, 上海 200122;

2. 中国电信股份有限公司安徽分公司, 安徽 合肥 230001)

摘要: 电信运营商 IPTV 业务平台拥有大量的节目内容信息, 可以建立 IPTV 内容知识库。传统方式下使用关系数据库对知识进行格式转换和存储, 在模型结构复杂性、查询直观性以及效率方面存在局限。利用图数据库技术, 可以基于全新的图模型构建 IPTV 内容知识库。通过研究比较, 提出了基于图数据库的 IPTV 内容知识库存储和查询的思路和实现方法, 并利用某电信公司实际数据进行了场景验证。

关键词: 图数据库; IPTV; 知识库

中图分类号: TP392

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016319

Design and implement of IPTV content repository based on graph database

CAI Xin¹, RUAN Yilong², SHI Yirong¹

1. Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

2. Anhui Branch of China Telecom Co., Ltd., Hefei 230001, China

Abstract: Plenty of program content information in the telecom operators IPTV platform can be used to build knowledge base. The traditional way of using relational database to format conversion and storage of knowledge has limitation in the model structure complexity and query visualization and efficiency. Graph database can be used to construct IPTV content knowledge based on the basis of a new graph model. Through the research and comparison, the idea and method of storing and querying the IPTV content repository based on graph database were proposed, and the actual data of a telecom company was used for the scene verification.

Key words: graph database, IPTV, knowledge base

1 引言

IPTV 业务是运营商转型的重点业务, 在未来 1~2 年将迎来产品生命周期的快速大规模发展阶段。而内容是 IPTV 业务的灵魂, 运营商 IPTV 业务平台中已经沉淀并将继续积累更多的内容。IPTV 平台已经含有一些节目内容标签元数据。利用互联网爬虫技术, 可以在互联网上爬取

更多的内容属性, 与平台本身信息进行融合, 形成更丰富的 IPTV 内容知识库。该知识库能够为 IPTV 产品包设计者以及 IPTV 最终用户提供完备的节目内容信息查询服务, 提升其决策行为的有效性。IPTV 内容知识库, 需要具备支持灵活的元数据扩展性, 满足海量结构化数据的存储和复杂关联查询响应时间的要求。本文尝试采用与传统关系型数据库不同的全新的图数据库技术, 对 IPTV 内容知识库

收稿日期: 2016-11-10; 修回日期: 2016-12-14

进行创新设计和实现,并验证其可行性。

2 图数据库关键技术

图数据库是一种特殊的 NoSQL 数据库,是将数据存储在图(graph)模型中。图模型可看作是从 key-value 模型发展出来的一个分支,不同的是,它的数据之间有着广泛的关联,形成一种图结构,并且支持一些图结构的算法,适合社交网络、推荐系统等^[1]。

2.1 图模型的存储格式

语义网(semantic Web)是知识管理领域的重要分支,旨在建立一个数据标准,使全球的 Web 数据以及数据与数据间的语义关联关系能够被计算机所理解,由 W3C 管理和维护。现已发布了数个语义万维网格式标准,其中重要的一项是基本资源描述框架(resource description framework, RDF)。

RDF 为描述图像、文档和二者之间的相互关系定义了一个简单数据模型^[2]。RDF 的基本单元是由〈主语(subject), 谓语(predicate), 宾语(object)〉构成的三元组(triple)。主谓宾的取值称为资源(resource)。一组 RDF 数据可构成一个 RDF 有向图,其中每一个三元组对应为图结构上的一个“顶点—边—顶点”的子图,陈述了由谓语表示的在主语和宾语所指的事物之间的关系。

基于 RDF 的基本思想,开源图数据产品 Neo4j 采用一种进化的属性图(property graph)模型实现图模型的存储。属性图建模使用 3 种构造单元:节点(node)、关系(relationship)、属性(property),节点即图的顶点,关系即图的边(具有方向和类型),属性可以在节点和关系上面任意多地进行定义。属性图模型如图 1 所示。

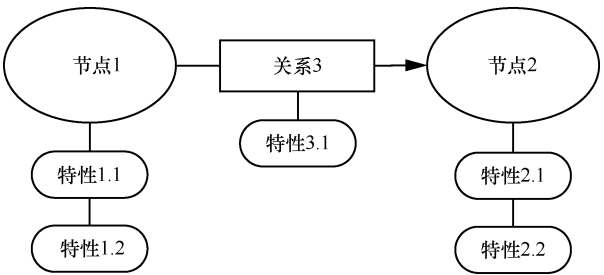


图 1 属性图模型

对比属性图和 RDF,可以看出:

- 属性图的节点融合了 RDF 的主语和宾语;
- 属性图的关系对应 RDF 的谓语;
- 属性图的节点和关系可扩展任意多的属性,具备 NoSQL 弱模型约束特征。

2.2 图数据库的检索方法

图数据库提供强大的图搜索能力。例如 Neo4j 通过专门为操作图数据库设计的 Cypher 语言高效地操作图数据库。图数据库的检索方法包括遍历检索和索引检索。

(1)图的遍历检索

通过遍历算法完成,根据算法从开始节点到与之相连的节点开始查询。图的遍历算法包括以下两种。

- 深度优先(depth first)搜索:就是找到第一个节点,递归地一直往下找,直到找不到合适的节点后,才进行回溯。
- 广度优先(breadth first)搜索:是最简便的图搜索算法之一,这一算法也是很多重要的图算法的原型,属于一种盲目搜寻法,它不考虑结果的可能位置,彻底地搜索整张图,直到找到结果为止。

(2)图的索引检索

通过建立索引,可以更快、更高效地查找某个节点或关系。索引可以分为两类,基于节点的索引和基于关系的索引^[3]。

3 关系模型与图模型 IPTV 内容知识库设计

3.1 传统关系模型设计方式

IPTV 平台与互联网爬虫融合后的内容(content)信息包括:title(内容标题),year(年份),label(电影/电视剧标签),star(主演),area(地区),director(导演),type(节目类型),writer(编剧)。

一个典型的内容数据样例见表 1。

每行代表一个节目内容,其中互联网爬取的属性带有非结构化信息特征,例如:star、director、type、writer 等。

在使用传统关系建模时,需要对这些非结构化信息内可以进一步原子化的数据元素进行解析和概念抽象,形成

表 1 IPTV 内容数据格式

内容标题	年份	标签	主演	地区	导演	节目类型	编剧
集结号	2007 年	电影	张涵予/邓超/袁文康/汤嬿/廖凡/王宝强	中国大陆	冯小刚	传记/战争/历史/动作	刘恒
青春集结号	2015 年	电视剧	焦俊艳/经超/刘小震/石安妮/马程程/王聪	中国大陆	尚敬	剧情/偶像/言情	应良鹏



更具一般性的数据实体。本例中 star、director、writer 中“/”分割的数据,可以抽象为人(person)实体。

接下来内容与人的多对多关系,需要建立内容与人关联关系(content person relation)实体进行抽象描述。可以根据人对于内容的工种(type of work)差别,建立不同类型的联系,即:饰演(act)、导演(direct)、编剧(write),这里还要考虑其他工种(美工、制片人等)信息的可扩展性,需要建立单独的工种实体。

另外,对于节目类型也需要抽象出专门的实体表达,而内容与节目类型之间,也是多对多的关联关系,需要抽象出一个节目内容与类型关联关系(content type relation)实体进行表达。

整个关系模型的 E-R 如图 2 所示。

3.2 图模型设计方式

关系模型设计中,会使用一个实体(entity)(例如:节目内容)来对一类事物进行抽象,模型描述的是经过抽象的实体以及实体和实体之间的关系,而并非具体的事物实例。

图模型采用一种更为自然的方式来描述事物以及事物和事物之间的关系。在图模型中,属性图的构成要素——节点,则恰恰是用来描述具体的事物实例(例如:电影《集结号》),而对于同一类事物的抽象则并非必须。尽管在图数据库中,也定义了用以表示一类节点集的概念——标签,但标签并不会像关系模型中的实体那样,对节点的结构模式(属性构成)进行强制约束。被标为同一标签的各节点所包含的数据属性有可能不同。

本例中的人可能会有不同的职责分工,并由此通过不同的关系和其他节点关联。例如一个人既可能是演员,可能是导演,也可能是演员兼导演。

在传统关系模型中,可能需要为建立单独的工种实体,并据此在内容与人关联关系实体中区分各类不同的职责分工。而在图模型中,由于关系可以被无限灵活地定义,因此可以通过定义不同的关系对内容和人进行连接,表达所需要的不同业务概念。本例中一个局部的属性图模型如图 3 所示。

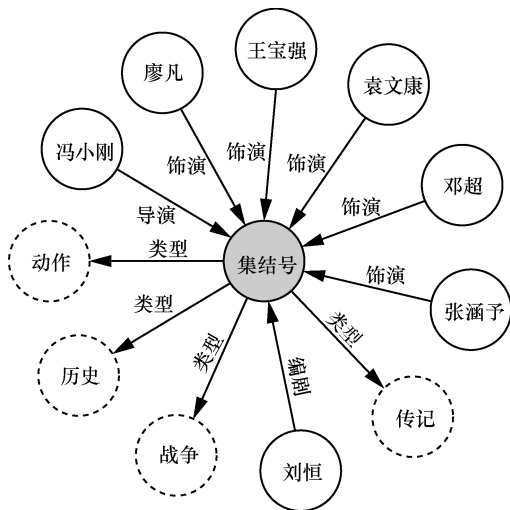


图 3 IPTV 内容知识库图模型局部示例

3.3 图模型的优势

从比较可以看出,传统关系模型在描述实体之间的关系时,需要创建一系列不同的关联表,例如本例中的内容

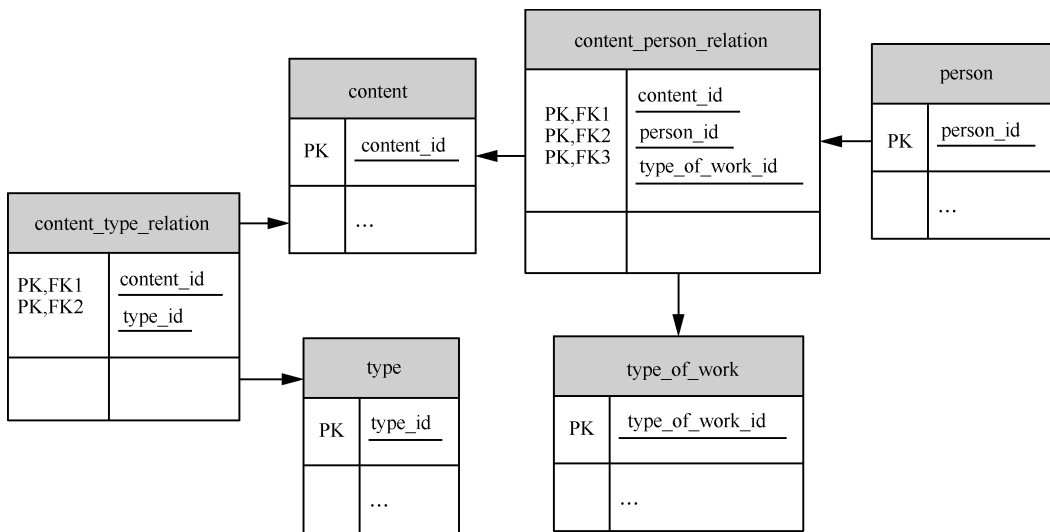


图 2 IPTV 内容知识库关系模型

与人关联关系和节目内容与类型关联关系,以记录这些数据之间的关联关系,而且这些关联表常常不包含除用于连接的外键之外的其他任何有意义的数据,它只是关系模型为满足三范式要求下的策略性的实现方式。

这些关系模型中存在的大量关联表,对于文件树这样的递归结构和社交图谱这样的网络结构等复杂查询带来严重的性能问题^[4]。图形关系上的每次操作都会导致关系型数据库的一次代价昂贵的联接(join)操作,这种操作以两个表的主键集合间的集合操作来实现,性能影响会随着表中的元组数量增加和查询的连接深度增加而愈加显现。

与之相对的,图模型原生支持关系,且图模型中的关系可以进行任意地定义,对所有不同的节点之间不同的连接含义有无限的包容性,甚至任何两个关系都可以不一样。因而,不再需要像关系模型那样,为了连接不同类型实体而需要定义不同种类的关联,关系的维护变得异常简单。

图模型的设计原理,决定了图遍历执行的速度是一个常数,跟图的规模大小无关。不像在关系数据库的联接操作那样,不涉及降低性能的集合操作。对于大规模深度遍历而言,极大地提高了性能。

4 图数据库 IPTV 内容知识库实证

4.1 IPTV 内容知识库图模型构建

实际验证中,以某电信公司 IPTV 节目内容作为测试数据样本,以本文提及的开源图数据库产品 Neo4j 作为测试系统平台,开展相关测试。

由于图模型的节点对应的就是具体事物实例,因此大规模图模型的构建过程,就是实例数据的导入过程。主要步骤如下。

(1) CSV 文件准备

对表 1 格式的源数据中 star、director、type、writer 等非结构化字段进行解析,转换为多个分别描述节点和关系的 CSV 文本文件,包括:节点数据集(content.csv、person.csv、type.csv)、关系数据集(content_actor.csv、content_director.csv、content_writer.csv、content_type.csv)。

(2) 节点数据导入

以内容标签的节点为例,需要执行以下的语句完成数据导入。

```
USING PERIODIC COMMIT 10000
```

```
loadcsv with headers from "file:///D:/Neo4j/ content.csv"
```

```
AS row
```

```
create(:content {content_id:row.content_id,title:row.title,
year:row.year,label:row.label,area:row.area});
```

(3) 关系数据导入

以内容和演员(actor)之间的关系为例,需要执行以下的语句完成数据导入。

```
USING PERIODIC COMMIT 10000
```

```
loadcsv with headers from "file:///D:/Neo4j/ content_
actor.csv" AS row
```

```
match (l:person {name:row.actor})
```

```
match (r:content {content_id:row.content_id})
```

```
merge(l)-[rel:ACT]->(r) on create set rel.c_name=row.
c_name;
```

4.2 IPTV 内容知识库图检索实际验证

场景验证将以电影《集结号》为起点,展开与之相关联的信息的查询。实验将逐步增加连接深度,测试对检索性能的影响。

第 1 个检索将返回《集结号》的所有演职人员(演员、导演、编剧)。语句如下。

```
match (l1)-[rel1]->(m) where m.title = '集结号'
```

```
return l1;
```

第 2 个检索深入一层,将返回《集结号》的所有演职人员参与的所有其他电影。

```
match (l2)<-[rel2]-(l1)-[rel1]->(m) where m.title = '集
结号' and l2.label = '电影'
```

```
return l2,l1;
```

第 3 个检索再深入一层,将返回《集结号》的所有演职人员参与的所有其他电影的所有演职人员。

```
match(l3)-[rel3]->(l2)<-[rel2]-(l1)-[rel1]->(m) where
m.title = '集结号' and l2.label = '电影'
```

```
return l3,l2,l1;
```

按照前面的逻辑继续,第 4 个检索再深入一层。

```
match(l4)<-[rel4]-(l3)-[rel3]->(l2)<-[rel2]-(l1)-[rel1]
->(m) where m.title = '集结号' and l2.label = '电影'
```

```
return l4,l3,l2,l1;
```

第 5 个检索更加深入一层。

```
match (l5)-[rel5]->(l4)<-[rel4]-(l3)-[rel3]->(l2)<-[rel2]
-(l1)-[rel1]->(m) where m.title = '集结号' and l2.label =
'电影'
```

```
return l5,l4,l3,l2,l1;
```

图检索性能验证见表 2。



表 2 图检索性能验证

性能	检索连接深度				
	1	2	3	4	5
访问节点规模/个	8	62	346	4 197	22 234
检索执行时间/ms	71	97	191	521	2 841
执行速度/(节点数 $\cdot$ ms $^{-1}$)	0.113	0.639	1.812	8.055	7.826

由实际验证结果可知,体现图检索性能的执行速度(访问节点规模/检索执行时间),即单位时间访问的节点数,并未因为检索连接深度以及所访问节点数量的增加而出现显著降低。与之相反,总体趋势呈现不断提升并逐渐收敛于一个稳定值,符合(当图模型规模足够大时)图遍历执行速度是一个常数的预期。

5 结束语

随着知识管理领域相关理论研究的日益深入,基于图模型的图数据库技术逐渐从实验室走向生产应用。得益于设计理念上对于连接的原生支持能力,使得图数据库技术在欺诈检测、基于图的搜索、主数据管理、推荐引擎、社交网络等具有复杂关系特征的场景中大有用武之地。但也应该看到,当前图数据库产品市场仍处于早期阶段,有几款相对成熟的商业型产品,如 infinite graph、trinity 等,但未得到大规模应用。而 Neo4j 作为当前比较成熟且较高性能的开源图形数据库,尚不支持分布式存储,难以应对大数据环境下更加海量的图模型存储和检索的需求。相信在未来的几年,图数据库的产品成熟度将会进入快速通道,会成为在数据平台方面的重要选项,有必要持续跟踪和关注。

参考文献:

- [1] 项灵辉. 基于图数据库的海量 RDF 数据分布式存储 [D]. 武汉: 武汉科技大学, 2013.
XIANG L H. Mass RDF data distributed storage based on graph database[D]. Wuhan: Wuhan Science and Technology University, 2013.
- [2] 信息组织 [EB/OL]. [2016-09-10]. <http://www.docin.com/p-440989969.html>.
Information organization [EB/OL]. [2016-09-10]. <http://www.docin.com/p-440989969.html>.

[docin.com/p-440989969.html](http://www.docin.com/p-440989969.html).

- [3] 康杰华, 罗章璇. 基于图形数据库 Neo4j 的 RDF 数据存储研究[J]. 信息技术, 2015(6): 115-117.
KANG J H, LUO Z X. RDF data storage research based on graph database[J]. Information Technology, 2015(6): 115-117.
- [4] 图形数据库、NoSQL 和 Neo4j [EB/OL]. [2016-09-10]. <http://www.infoq.com/cn/articles/graph-nosql-neo4j>.
Graph database, NoSQL & Neo4j[EB/OL]. [2016-09-10]. <http://www.infoq.com/cn/articles/graph-nosql-neo4j>.

[作者简介]



蔡鑫(1975-),男,中国电信股份有限公司上海研究院高级工程师,主要研究方向为数据规划、信息管理、数据标准。

阮宜龙(1977-),男,中国电信股份有限公司安徽分公司高级工程师,主要研究方向为数据规划、大数据平台运营管理。

石屹嵘(1970-),男,中国电信股份有限公司企业信息化验证实验室主任、高级工程师;中国电信股份有限公司上海研究院大数据专业资深专家,负责完成了中国电信股份有限公司集团和省公司几十项科研项目,包括企业信息化架构设计和验证、信息系统性能评测和优化、云计算和大数据关键技术验证等;中国电信股份有限公司集团级优秀人才。目前的研究方向主要为大数据平台技术、大数据分析和模型设计、信息系统架构设计、企业架构设计、安全架构设计等。获得国家多项专利授权,发表论文 10 多篇,并获得 CISSP 和 CISA 等多项国际认证。