图数据库在标签系统中的应用研究

王慧孜，范炜
（四川大学公共管理学院信息管理技术系，成都 610064）

摘要：图数据库是NoSQL技术之一，其图数据结构适合用于标签系统的数据存储与检索。本文分析了标签系统的数据存储方式，比较了图数据库与传统关系型数据库，基于“用户-标签-资源”三者关系构建了图数据模型。通过Flickr API采集到的图片数据，以图数据库Neo4j作为应用环境构建了小型的标签系统，使用图数据库语法可以构造出多维度的检索分析，得到直观的可视化网络图。图数据库对于处理呈现出复杂网络结构的标签系统有着明显优于关系型数据库的优势和实用价值。

关键词：图数据库；标签系统；数据管理
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2015.04.004

1 标签系统的数据管理概述

标签（Tagging）是通过用户提供的关键词进行信息组织的方法，它正在改变网络上寻找、使用和分享信息的方式。标签就是用户为资源添加的关键词。标签系统可以被定义成是一系列用户、一系列标签、一些资源对象以及在时间维度上三者之间关系的集合。这个系统中没有层级结构，所有的条目都属于扁平化的空间。

标签系统主要由三个基本要素构成：用户、标签与资源。三者之间的数据关系呈现出网状的图（Graph）结构。根据此共识，已有较多学者与实践者对标签系统的数据结构进行了研究。Kaikai等[1]将大众分类看成是一个复杂的网络结构。他们认为大众分类法使得用户去分享他们个人使用的标签。因为在同一个标签下用户可以分享不同的内容，同一内容可以被赋予不同的标签，这样标签之间是互相联系的，而资源之间亦是如此。这样的特征使得大众分类很容易表示成一个标签和内容的网络。Yin等[2]将Web资源对象的分类问题看成是资源对象与标签的图形优化问题，并且他们也提出了一个有效的迭代算法被提出来解决这样的资源对象与标签的图形优化问题。鲁晓明[3]对社会性标签系统中的复杂网络特性进行了分析，就标签系统中的用户、信息和标签这三个主体之间的联系进行了探讨。

围绕标签系统的分析，底层的数据存储大多数以关系型数据库为支撑，例如SQL Server、Oracle、MySQL等。通常，标签系统的相关查询在关系型数据库中处理为多个二维表的连接。具体而言，当所需的数据涉及到多张表时，就需要通过SELECT语句对FROM子句中的二维表进行选择和投影等操作，同时也需要通过WHERE子句对相应的二维表做条件连接操作。当标签系统中标签、用户与资源三者随着数量增加与关联复杂度提高时，关系模型会因为过多的表连接变得复杂，同时过多的表间连接会降低检索性能，而外键约束也会增加额外的开发和维护的费用，使现有的关系型数据库很难满足不断变化的业务需求。

2 图数据库简介

随着NoSQL技术浪潮的兴起，图数据库（Graph Database）作为NoSQL技术中的一类，以解决社交网络数据管理问题的典型代表。图数据库以图结构作为数

据模型，具有网状数据的原生存储与检索能力。与传统的关系数据库相比，无论是标签系统中过大的数据集，还是标签系统中复杂的网络连接结构，图数据库不仅能与标签系统的底层数据结构良好契合，直观的查询语法支持，以及更富有表现力的数据呈现方式。

图数据库在国内的相关研究主要集中在图数据库中匹配以及查询算法优化、图结构及其数据模型、图数据库技术特性等3方面。沈思、苏新宁[4]分析了面向知识服务的分类表结构，并针对关系数据库的数据机器存储方式在分类表知识更新、删除、添加上存在的弊端，给出了分类表的图数据库存储方式以及具体的检索案例。王余蓝[5]认为在成熟度、安全性等方面虽然图数据库要劣于关系数据库，但在处理复杂数据关联方面远优于关系数据库，适合存储关联关系复杂、关系动态变化等社交性数据。高劲松等[6]针对传统的关系-对象模型实现文献知识元存储的不足，提出以AllegroGraph为图数据库、以RDF作为文献知识元的链接存储方法，并对其进行了实验验证。正如关系型数据库产品众多一样，随着NoSQL阵营的迅速发展，图数据库分支也衍生出一些具体的图数据库产品，下面进行简要分析。

FlockDB[7]是用于存储邻接表的分布式图数据库。因为它只是试图解决很少的问题，因此会比其他的图数据库更加简单。它支持横向扩展，是专门为在线、低延迟、高吞吐率这类环境所设计。Twitter就是使用FlockDB来存储社交图和二次索引。

OrientDB[8]是一款拥有文档灵活性的分布式开源图数据库。虽然它是基于文档的数据库，但是却可以像在图数据库中一样用记录之间的直接联系来管理关系。它的存储效率高达150000条/秒，而且用户可以在几毫秒内遍历完部分或全部记录的树和图。

AllegroGraph[9]是一款现代的、高性能的、持久稳固的图数据库，支持语义网的RDFS三元组存储。它支持语义网事实查询语法标准SPARQL，RDFS++和Prolog。它将高效的内存利用与基于磁盘的存储相结合，这使得它在保持良好性能的同时还能够负载数以亿计的RDF。

Trinity[10]是一个以内存云为基础的分布式图系统。内存云是全球范围内可寻址的，内存中的键值对存储在机器集群中。通过分布式的内存存储，Trinity提供对大数据集的快速随机数据存取，这使得Trinity成为了天然的大图处理平台。因为Trinity具有快速图探测和分布式并行计算的能力，对于包含上亿节点的大图它既能支持低延迟的在线查询处理又能支持高吞吐率的离线分析。

Neo4j[11]是一款由Neo Technology 所支持的开源图形数据库。有以下特点：直观性，用图模型来表现数据；可靠性，兼容ACID事务特性；持久和快速性，使用基于磁盘的本地存储引擎；高度扩展性，有着多达数以亿计的节点、关系和属性；可用性，分布在不同的机器上；可表现性，使用有力的人类可读的图形查询语言；快速性，为了实现高速的图形查询提供了有力的遍历框架；通过REST接口或者面向对象的JAVA API进行访问。

通过比较分析，如表1所示，属性图作为最流行的图数据模型能够被以上大多数的图数据库产品支持。除属性图外OrientDB还能支持文档、键值、对象其余3种数据类型，由于它能够支持多元化的数据模型，也就意味着它在拥有完整的本地图功能的同时，还保留着其他数据库的一些特性。AllegroGraph作为建设语义网应用的数据库和应用程序框架，能够将数据和元数据以三元组的形式存储起来，是以上产品中唯一一款支持RDF的图数据库。其中唯一不能支持Java的Trinity只能向用户提供C# API，而剩余的图数据库各自均能支持多种类的程序接口。总之，目前图数据库还在发展初期阶段，各有特点，暂没有具体的标准能够区分出绝对优势产品，还是要根据实际需求来选择最适合的图数据库产品。

表1 图数据库产品基本比较

	FlockDB	OrientDB	AllegroGraph	Trinity	Neo4j
数据模型	属性图	属性图、文档、键值、对象	RDF	属性图、超图	属性图
查询语言	Ruby+构句方法	SQL	SPARQL	TSL	Cypher
API	Thrift API（多种语言）	Java API、Scala API、HTTP API、Gremlin API等	REST协议结构（Java、Python等）	C# API	REST API、Java
许可方式	开源	开源	商业	商业	开源

在图书情报领域，文献资源库以行业传统关系型数据库来驱动。随着图数据库的快速成长，适时将新兴数据管理技术引入到数字资源组织与利用中是与时俱进的表现，也是IT技术服务于资源、服务的体现。基于以上分析，本文以此为出发点探讨图数据库这一新兴数据管理技术在标签系统中的应用。

3 标签系统的图数据模型

首先分析标签系统三要素的管理模型，以此为基础转换出图数据模型。Smith[12]提出关系数据库中标签系统的数据模型一般分为两种：简单标记模型和协同标记模型。简单标签系统的数据库中通常是包括4张表，如图1-a所示：用户表、资源表、标签表、资源—标签表。在这种标签系统中，每一项资源对用户是唯一的，而一项资源却可以对应很多的标签。协同标签系统的数据库中包括4张表，如图1-b所示：用户表、资源表、标签表、用户—资源—标签表。不同的用户可以对同一项资源添加不同的标签，“用户—标签—资源”表可以唯一确定它们三者之间的关系。

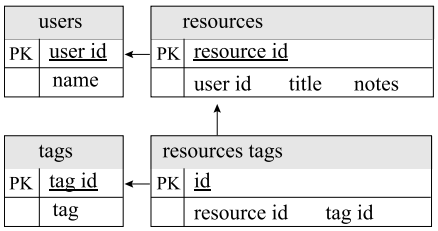

图1-a 简单标记schema

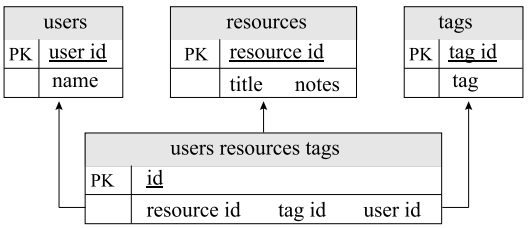

图1-b 协同标记schema

图数据库中“图”是一种数据结构，并不是直接理解字面上的图像、图形。图是顶点和边的一个集合，或者说，是一系列的节点（实体）和连接它们的关系的集合。Robinson[13]提到了三种主要的图数据模型：属性图、RDF三元组、超图。并且主要介绍了属性图模型的特性：它包括节点和关系；节点包含属性；关系也包含属性；关系可以被命名也可以被定向，有一个开始节点和一个结束节点。在一张图中，关系是建立语义情景的关键。

在图数据库的标签系统中每一个用户、每一项资源、每一个标签都可以看做是一个实体，即图中的一个节点。它们之间的关系可以看做是标签“标记”资源、用户“添加”标签、用户“上传”资源的关系。用户可以添加标签或者选择现有的标签对特定的资源进行标记，同时在用户与此资源之间生成一条“用户—上传→资源”的关系。这样就很容易知道哪些用户上传过哪些资源和添加过哪些标签、一个用户用哪些标签标记过哪些资源，如图1-c所示。

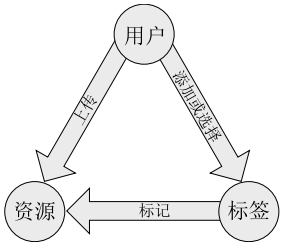

图1-c 标签系统的简单图数据模型

4 图数据库的存储与检索分析

Neo4j是当前最为流行的图数据库之一，本文以Neo4j为代表介绍图数据库的存储与检索方法。Robinson等[13]展示过Neo4j底层的数据结构，由节点存储记录和关系存储记录组成，如图2所示。Neo4j使用Cypher来进行查询。Cypher是一种简洁而且富有表现力的图数据库查询语言，用来描述和查询属性图。向数据库提交的查询时，SQL查询就是从二维表的行列关系中获取所需的相应数据，与SQL相比，Cypher对于查询的描述更为生动形象。当数据集的关系变得越来越复杂时，SQL语句会变得相对来说更加庞大和复杂，而Cypher会保持它的简洁性。

Neo4j不仅提供了适应于图结构的查询语言Cypher，同时对（节点、关系、属性）数据量的支持也达到了以亿为单位来计算。Neo4j数据量主要受到节点、关系、属性和关系类型主键的地址空间限制[14]，如表2所示。

当数据量不断增大时，关系型数据库的查询性能会弱化，但图数据库的性能依旧保持良好不受任何影响。每个查询执行的时间仅仅与遍历的满足查询的图的部分成比例，而不是与整个图的尺寸成比例。图数据库使用无索引邻接，也就是说在数据库中一些相互关

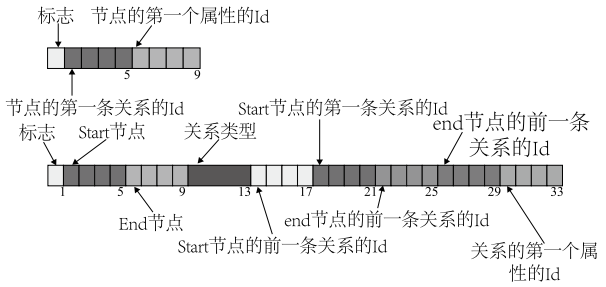

图2 Neo4j关系节点的存储记录底层数据结构

表2 Neo4j数据量支撑情况

节点Nodes	2^{35}(~340 亿)
关系Relationships	2^{35}(~340 亿)
属性Properties	2^{36} to 2^{38}，取决于属性类型 （最大~2740 亿，通常至少~680 亿）
关系类型 Relationship types	2^{15}(~32000)

联的节点物理地相互指向对方，这样可以保证图数据库会把复杂的连接变成快速的图遍历，不管数据集的大小，都能保持毫秒的性能。

现实世界中复杂的数据和真实模型无法一次性得知全貌，由于图模型的灵活性，用户并不需要提前进行详细地建模。图有着良好的可扩展性，这就意味着可以添加任意类型的新的关系、节点和新的子图到一个已经存在的结构，但是丝毫不影响它现有的查询和应用的功能。这一特性预示着将执行更少的迁移，因此也会减少维护的开销和风险。

5 标签系统的图数据库应用实现

基于以上对标签系统与图数据库的综合分析，结合实际应用来探索标签系统的图数据库存储与检索实现方法。以Flickr作为数据来源[15]，从图数据库建模的角度将采集到有关图片（资源）、用户、标签的数据进行整理与转换，并将得到的数据以JSON格式导入到Neo4j中。最后实现了标签系统在图数据库Neo4j中的存储和检索。

5.1 数据源与数据采集

Flickr图片分享网站是最早使用标签的网站之一。用户可以在 Flickr 上传图片、安全分享图片、为图片补充描述信息 (如授权信息、地理位置、人物、标签等)。除此之外，用户也能与家人、朋友、自己或社区中的任何人进行互动，包括留言、添加标签。这些数据不但成为了图片或视频的附加信息，而且更利于今后用户搜索到该资源。

Flickr 对外提供了开放的API，用户可以使用Flickr API从Flickr照片共享服务中检索到各种照片，可以通过API上传图片或视频，当然也通过API获取到所需要的关于图片的任何数据。本文在网页上通过相应API的方法直接从Flickr采集到以“风景”为主题的小部分数据作为图数据库标签系统的数据来源。从以上获取的数据中选取图片的id，依次得到相应id的图片具体信息，主要包括:图片本身信息（如id、标题、原始格式、拍摄时间、图片URL等）、上传者信息（nsid、用户名、真实姓名、用户位置等）、标签信息（标签内容、标记者用户名等），以及其他信息（位置、用户权限、浏览次数、评论次数、安全级别、可编辑性等）。

5.2 数据整理与转换

从图数据库建模的角度出发，以图1-c为参考，为图数据模型的构建，准备数据基础。已经得到的数据里一共包括19个字段及其属性，但是其中大部分数据（如位置、用户权限、浏览次数、评论次数、安全级别、可编辑性等）与标签系统中图片、用户、标签并不直接相关，对下一部分丰富已有的图数据模型没有太大价值。而其中个别字段及其属性与标签系统中图片、用户、标签本身直接相关，见表3，可以用来丰富上一部分中的图数据模型。于是，就得到了包含具体属性的标签系统图数据模型，如图3所示。

表3 图片字段及其个别属性信息

字段	属性
photo	Id、originalformat原始格式
title	_content图片的标题
urls	_content该图片的详细URL
owner	Nsid、username用户名、realname真实姓名、location位置
dates	Posted照片上传日期、taken照片拍摄日期
tags	标签数组tag{ authorid作者id、_content标签内容}

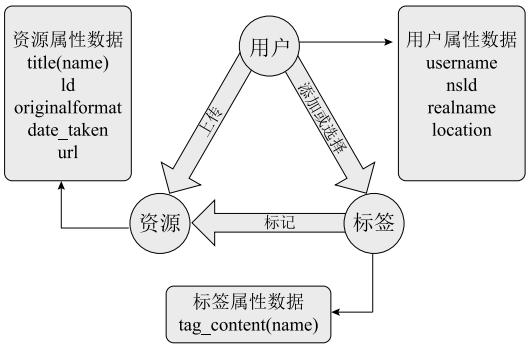

图3 包含具体属性的标签系统图数据模型

5.3 图数据库的存储

从所获得的数据集中随机选择id为“14442547137”的图片具体数据，按照上一部分建模的标准，将其处理成创建语句。

//创建用户节点P，属性包括：name、nsid、realname、location

CREATE (p:P { name: ' YASAX DESIGN ',nsid : '73084566@N04' , realname:'炼刘',location: '上海, 中国' })

//创建资源节点I，属性包括：name、id、originalformat、date_taken、url

CREATE (i:I { name: '卡帕多西亚的日落',id : '14442547137', originalformat :' jpg ',date: '2014-07-12 01:18:06 ' ,url: ' https://www.flickr.com/photos/73084566@N04/14442547137/'})

//创建标签节点T

CREATE (t1:T {name:'旅行'})

CREATE (t2:T {name:'风景'})

//创建用户、资源、标签之间的关系，用户p选择了资源i对其添加了标签t1、t2

CREATE(p) -[: upload]-> (i), (p) -[: add]-> (t1), (p) -[: add]-> (t2), (t1) -[: tag]-> (i), (t2) -[: tag]-> (i)

将以上的创建语句运行在Neo4j就能得到只有一张图片信息的图数据库标签系统，如图4所示。

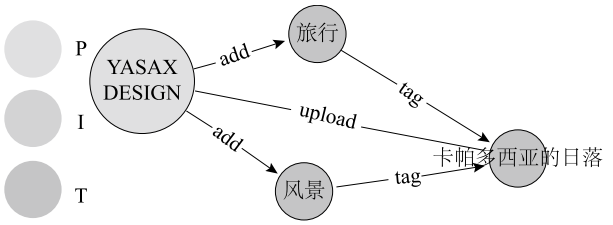

图4 简单建模用例图

上述部分只是包含了从整体数据集中抽取的及小部分数据。笔者将获取到所有的图片数据信息进行整理，导入到Neo4j中，图就会在数据库中生成，即以风景为主题的图数据库标签系统。比如执行所需时间仅仅在2851ms内，一共导入了143个节点，558条关系。与关系模型不同，这些节点和关系不会增加任何额外的复杂性到图中。因此，不必需要担心外键和约束影响了。

5.4 图数据库检索与展示

图数据库使用Cypher语言通过执行“match n return n”语句来进行查询，可以得到整个标签系统中所有的关系和节点，如图6所示。图中所有的节点被分成了3类，其中红色圆形P代表用户节点，一共12个；灰色圆形I代表图片（资源）节点，一共30个，蓝色圆形T代表标签节点，一共101个。通过不同的颜色可以将用户、标签、图片区分开来，这其实就相当于关系型数据库标签系统中的用户表、标签表和图片表。然后再通过各种关系将其连接起来。其中红色箭头代表用户与图片之间的关系“upload”，紫色箭头代表用户与标签之间的关系“add”，绿色箭头代表标签与图片之间的关系“tag”，它们共同表示用户选上传了某项图片并添加某个标签对其进行标记。从图5中可以发现一些有趣的现象，如图的左边大量的标签围绕着同一个用户，说明有些用户非常乐于为图片添加标签；图的右边则是一些稀稀疏疏的标签，表示有些用户总喜欢将相同的标签集合用于同一系列的图片中。

通过这一个标签系统，我们可以通过检索标签的关键词来寻找相关的图片和用户，通过检索用户的关键词来查看该用户上传过哪些图片和添加过哪些标签。以下是一些标签系统的检索与发现。

（1）一个用户上传的所有图片分别被哪些标签标

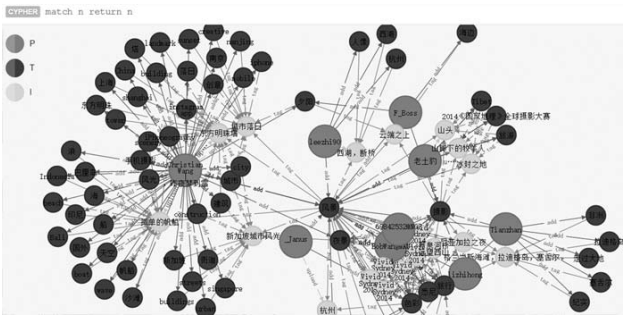

图5 标签系统的全景图

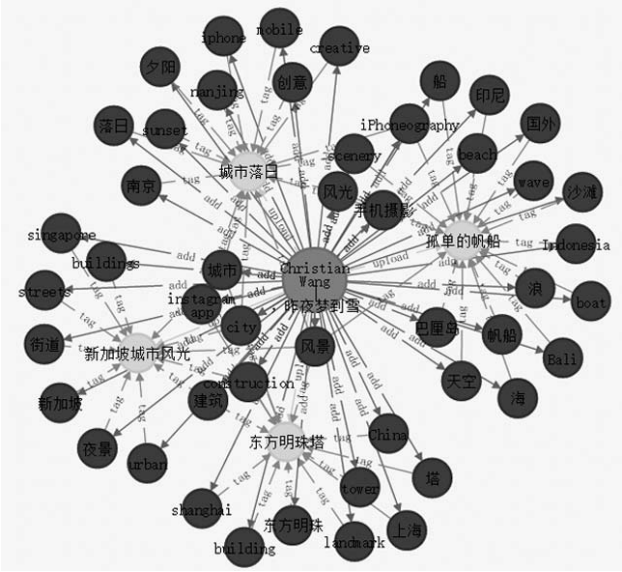

图6 标签系统中用户、标签、图片三者的联系

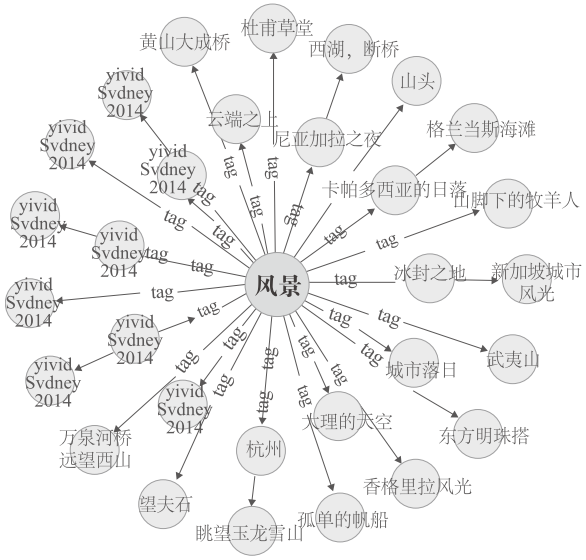

图7 标签系统中标记过的图片

记过。如图6所示，用户“Christian Wang｜，昨夜梦到雪”上传过“新加坡城市风光”“东方明珠塔”“孤单的帆船”“城市落日”等图片，同时对这些图片资源进行了一系列标记，可以直观地发现他使用了“天空”“创意”“夕阳”“手机摄影”等标签对“城市落日”这一图片资源进行了标记，用户、标签、图片三者之间的复杂联系就这样一目了然展现于眼前。

（2）标签标记过的图片。当某一用户对风景类图片较为感兴趣，通过对标签“风景”进行检索，就会得到所有被该标签标记过的图片，如图7所示。如果返回的图片数量越多，就说明该标签被使用的次数越多。通过这种方法就可以找出整个标签系统的热门标签。例如，在此标签系统检索标签“夕阳”“塔”“沙滩”得到的图片数量分别为2、1、1张，而检索标签“风景”一共得到了30张图片，相对于前三者标签“风景”应该为热门标签，在系统进行热门标签推荐时可以被优先考虑。

（3）标记过某一图片的标签。通过标签，用户不仅可以上传和分享资源，还能够更加容易地寻找到图片。如图8所示，图片“The landscape of Shangri-La香格里拉风光”被多个标签所标记，用户通过检索“风光”“自然”“云南”等任何一个标记过该图片的标签都可以找到它。同时很容易得出，一张图片被越多的热门标签标记就越容易被检索到。

（4）两个用户之间添加过的相同的标签，如图9所示。两个互不相关的用户节点之间没有任何一条直接的

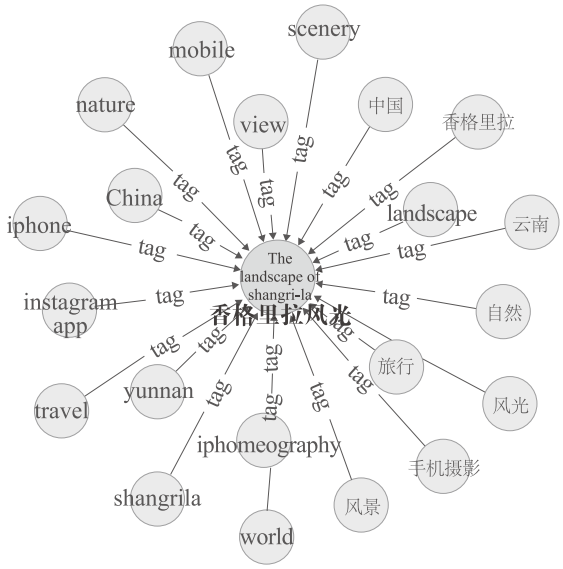

图8 标签系统中标记过某一图片的标签

关系，但是却可以通过一些中间节点（在此处是标签节点）得到它们之间一些隐含的关系。例如用户A添加过“风景”、“旅行”、“夜景”标签，而用户B也添加过“风景”、“旅行”标签、“夜景”，一旦这种共同的标签变多，这里存在着用户A和用户B都偏爱某一类相似的图片的可能性。因为有共同的爱好，标签系统可以分别向他们进行好友推荐。

基于以上对标签系统7个维度的检索，笔者发现以用户为中心，可以获取到用户上传过的图片和添加过的标签以及它们三者之间复杂的关系，同时可以将两个用户添加过的相同标签作为基础数据进行推荐。如果以

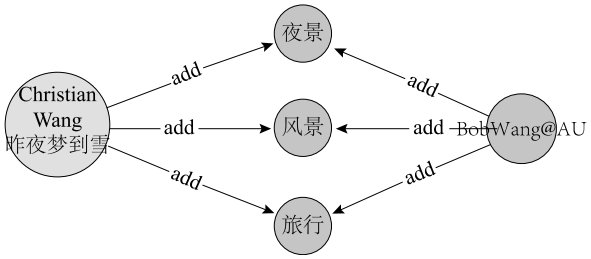

图9 两个用户共同添加过的标签

标签为中心，则可以得到一系列的相关图片；以图片为中心，就能知道标记过该图片的标签。通过图数据库的检索机制，可以直观地看到可视化网络图作为结果返回，而不是通常返回的列表结果。总之，图检索具有多维性、灵活性，其返回结果具有直观性、形象性。

6 结语

以关系为中心的社会化网络呈现的是网状结构，在一定程度上图数据模型可以很好地与社会化网络相契合，但是现有的标签系统绝大多数都是依靠关系型数据库来实现。本文尝试将图数据库应用到标签系统的数据管理与利用上，分析图数据库关键技术的基础上，提出标签系统的图数据基本模型。从Flickr上采集到的数据应用到该模型上，在Neo4j环境中构建了一个小型的标签系统。图数据库检索具有多维性和灵活性，而且其返回的可视化网络图更加直观形象，所展示的数据更少地被扭曲或失真。

相对于关系型数据库的行业成熟根基，图数据库技术的发展还在初期，但图数据库技术在大数据环境下有着灵活的数据存储与多维检索分析的适配潜力，可以预见未来有更多由图数据库驱动的数据应用出现。

参考文献

[1] Shen K, Wu L. Folksonomy as a Complex Network[J/OL]. [2015-3-15]. http://arxiv.org/pdf/cs/0509072.pdf.

[2] Yin Z, Li R, Mei Q, et al. Exploring Social Tagging Graph for Web Object Classification[C].Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 957-966.

[3] 鲁晓明. Web2.0中社会性标签系统的复杂网络特性研究[J]. 现代情报,2007,27(12):64-66.

[4] 沈思, 苏新宁. 知识服务环境下分类表的知识组织探究[J]. 图书情报工作,2014,58(7):113-118.

[5] 王余蓝. 图形数据库NEO4J与关系据库的比较研究[J]. 现代电子技术,2012,35(20):77-79.

[6] 高劲松, 马倩倩, 周习曼, 等. 文献知识元语义链接的图式存储研究[J]. 情报科学,2015,33(1):126-131.

[7] FlockDB [EB/OL].[2015-3-15].https://github.com/twitter/flockdb.

[8] OrientDB [EB/OL].[2015-3-15].http://www.orientechnologies.com/orientdb/.

[9] AllegroGraph [EB/OL].[2015-3-15].http://franz.com/agraph/allegrograph/.

[10] Trinity [EB/OL] .[2015-3-15].http://research.microsoft.com/en-us/projects/trinity/.

[11] Neo4j [EB/OL].[2015-3-15].http://neo4j.com/.

[12] Smith G. Tagging: People-powered Metadata for the Social Web[M]. CA :New Riders, 2007:138-143.

[13] Robinson I, Webber J, Eifrem E. Graph Databases[M]. Massachusetts :O'Reilly Media, Inc.2013:4.

[14] The Neo4j Manual [EB/OL].[2015-3-15].http://neo4j.com/docs/2.1.5/capabilities-capacity.html.

[15] Flickr [EB/OL].[2015-3-15].http://www.flickr.com.

作者简介

王慧孜，女，1994年生，2012级信息管理与信息系统学生，E-mail：wanghuizi1994@foxmail.com。
范炜，男，1981年生，博士，副教授，研究方向：信息组织与检索，通讯作者，E-mail：fanw@scu.edu.cn。

Application of Graph Database in Tagging System

WANG HuiZi, FAN Wei
（Department of Information Management & Technology, School of Public Administration, Sichuan University, Chengdu 610064, China）

Abstract: Graph database is one of NoSQL technologies. Graph data structure is suitable for storage and retrieval of tagging system. This paper analyzes the implementations of data storage of tagging system, and compares graph database with traditional relational database. It proposes a graph data model based on “user-tag-resource” triadic relation. Then, It collects the image data from Flickr API, and imports these tagging data into Neo4j. The multi-dimensional retrieval analysis with the graph database grammar and intuitive network diagrams are demonstrated. Finally, it indicates that graph database has obvious advantages and practical value than relational database on the processing of the tagging system which presenting a complex network structure.

Keywords: Graph Database; Tagging System; Data Management

（收稿日期：2015-03-26；编辑：刘伟）