基于图数据库的文献检索方法优化与实现

林启胜[1,2,3],王 磊[1,2],周 喜[1,2],赵 凡[1,2],马 博[1,2]
(1 中国科学院 新疆理化技术研究所,新疆 乌鲁木齐 830011;2 新疆民族语音语言信息处理实验室,新疆 乌鲁木齐 830011;3 中国科学院大学 计算机与控制学院,北京 100049)

摘 要:针对目前文献检索系统对于查询结果仅以文本形式呈现的问题,提出了一个基于图数据库的文献信息检索系统.该采用图数据库Neo4j进行存储,用户通过关键词查询,系统以图节点的可视化形式将信息呈现给用户,可以直观地了解文献的相关信息.实验结果表明,该系统比关系基于数据库的系统更快,执行的时间减少了71%(对于3-节点查询),88%(对于4节点的查询),以及99%的(5-节点查询).

关键词:Neo4j;数据库;数据存储;信息检索;Cypher
中图分类号:TP393 **文献标识码**:A **文章编号**:1000-7180(2017)10-0063-05
DOI:10.19304/j.cnki.issn1000-7180.2017.10.013

Design and Implementation of a Literature Retrieval System Based on Graph Database

LIN Qi-sheng[1,2,3], WANG Lei[1,2], ZHOU Xi[1,2], ZHAO Fan[1,2], MA Bo[1,2]
(1 Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China;
2 Xinjiang Laboratory of Minority Speech & Language Information Processing, Urumqi 830011, China;
3 College of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: For the current literature retrieval system for the results presented in the form of text only problem, a graph database based literature Information Retrieval System. The use graph database Neo4j storage, user query literature by keyword, the system visually in the form of graph nodes of information presented to the user can intuitively understand literature related information. Experimental results show that the system is faster than relational database system, the implementation time is reduced by 71% (for 3-node query), 88% (for 4-node queries), and 99% (for 5-node query).

Key words: Neo4j; graph database; data store; information retrival; cypher

1 引言

对于文献的浏览和检索,传统的基于关键字的文献检索和浏览方式虽然可以返回大量信息,但关注的文献信息形式单一,忽视了文献间的信息及关系复杂,浏览和检索效率不高[1].

目前已有的文献信息存在两个主要的问题:(1)搜索文献返回的最终结果都是以文本的形式呈现;(2)复杂查询中存在一些性能问题.所以,本文主要讨论解决以下研究问题:(1)如何设计一个文献检索系统,直观地呈现检索结果(作者、作者单位、文献来源等);(2)测试执行检索操作所花费的时间.为了解决以上问题,我们实现了基于图数据库的文献检索系统.该系统是一个可扩展、交互和高效检索文献信息系统.为科研人员提供合适的文献信息,有利于研究人员高效快速地查找文献信息.

收稿日期:2016-11-20;修回日期:2016-12-22
基金项目:中科院西部之光——西部博士项目(XBBS201315)

2 图模型和图表数据存储

目前可以用三种方式表示图模型,包括关系数据库、三元组、图形数据库[2-7].关系数据库,如Ora-

cle、MySQL，让用户通过网络数据模型管理图形数据. 这个关系型数据库的类型存储在连接信息节点表和关系表. 尽管关系数据库支持图形数据显示，不过，关系型数据库要花费巨大的代价. 在关系数据库中，如图遍历等操作代价是非常大的，涉及到表关联查询，导致查询花费时间长，效率不高. 三元组是另一种流行的数据存储形式图形数据. 三元组包括主体、谓语、客体，三元组用 RDF(Resource Description Framework)来表示；它是 W3C 的一个标准. 因此，三元组通常被称为一个 RDF 存储. 因为三元组是为图形数据开发的，所以他们比关系数据库更强大. 然而，已知的三元组有可扩展性的问题和性能上的问题. 图数据库是近几年发展起来的数据存储方式. 它是 NoSQL 数据库，是可扩展的，具有数据复制和容错等功能. 相比三元组，图数据库具备这些优势：(1)图形数据库支持的表示无向加权图，而三元组支持非加权图；(2)图形数据库不需要预先定义模式，而三元组需明确定义模式；(3)图形数据库是适合管理大数据. 基于上述优点，图形数据库更适合于具有处理大量真实世界数据的能力.

图数据库已被广泛地用于表示数据构成实体之间的实体和关系类型. 图数据库已在多种在线社交网络服务上得到应用，如 Facebook 和 Twitter 代表人与人的关系图模型，适用于许多复杂关系的领域. 同时也在万维网、社交网络、文献信息和电网等得到应用. 图数据库是理想的表示文献实体，如论文、作者、术语、文献来源和隶属关系，以及引用关系.

3 系统设计

针对传统文献信息检索系统的不足，基于目前 NoSQL 中图数据库的展示数据清晰的优点，开发一个基于图数据库的文献检索系统.

3.1 系统架构

该系统基于 Spring 框架开发的 B/S 系统，使用 Neo4j[3] 图形数据库，用于管理文献信息，并用 D3.js 进行可视化. 图 1 表示了一个概念模式. 可以在概念层面可以更好地了解实体及其关系. 如图 1 所示，概念模式包括五个实体类型和五种关系类型.

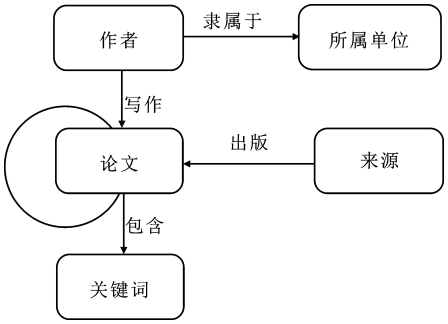

图 1 概念模式

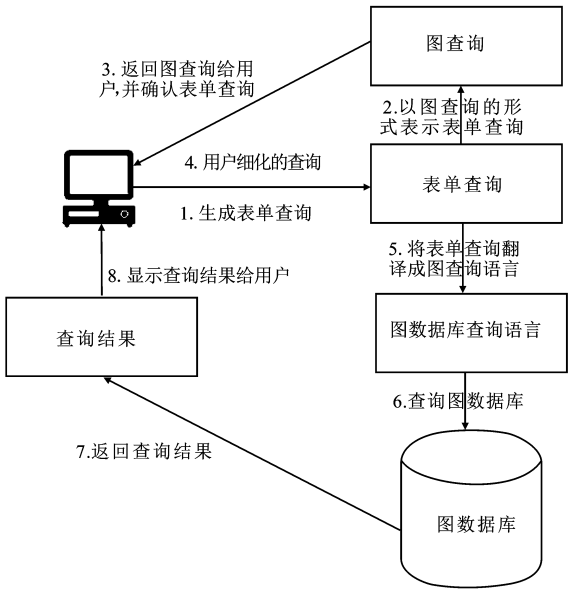

图 2 系统处理流程

该系统由三部分组成：查询生成、查询优化和查询图数据库. 具体流程如图 2 所示.

(1)用户根据他们的信息需求生成表单查询；

(2)以图查询的形式表示表单查询；

(3)生成的图查询发回给用户；

(4)在返回的图查询基础上，用户在必要时细化他们的查询；

(5)细化后(可选)，表单查询被转换成图形数据库查询语言；

(6)系统查询图形数据库；

(7)图形数据库将查询结果发送给系统；

(8)该系统将搜索结果返回给用户.

搜索过程是一个互动和迭代过程，用户可以对他们的查询需求进行确认.

下面对查询生成、查询细化、查询图形数据库进行详细说明.

3.2 查询生成

如图 1 所示，文献信息包含论文作者、文献、关键字、来源和隶属关系. 该系统还可以支持其他类型的文献信息，通过简单的添加，连接它们到现有的实体的模式中. 每个类型都有一个名字的属性. 首先，用户选择一个目标实体. 下一步，用户选择一个或几个属性类型，通过提供每个类型属性的值来限制目标实体. 例如，在搜索论文时，用户可以选择一些类

型的属性类型，如作者和源来限制论文. 通过选择目标类型生成表单查询后，查询被发送到系统，生成一个图查询.

3.3　查询细化

系统根据用户的表单查询生成图查询. 生成图查询的目的是为用户提供一个简单的方法来验证他们的原始查询，并在必要的时刻进行修改. 图查询是使用节点和关系的形式表示表单查询，它直接从表单查询生成并显示表单查询的图形表示. 如图 3 所示，不同类型的文献实体之间的关系是明确的. 目标类型的颜色是黑色，一个或多个用户提供的标签的颜色是红色的. 没有由用户指定的实体，但要连接的目标是蓝色的.

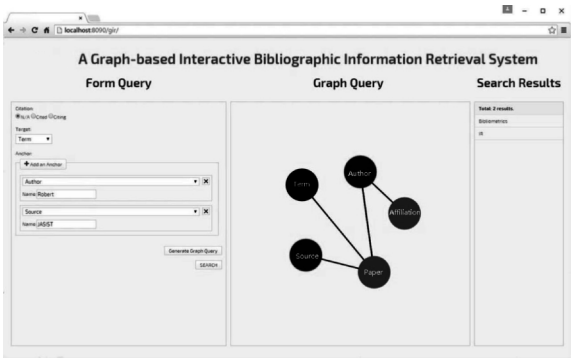

图 3　各类文献实体之间关系

基于图的查询，用户可以捕获和验证所生成的表单查询的意义. 这个过程是互动的，用户可以在必要时，反复完善他们的原始查询. 图查询组件采用了 d3.js. 使用表单来表示复杂的搜索上下文可能不是很直观，而提供交互式界面，通过显示图形查询提供交互式界面可以消除在表单查询的制定过程中可能出现的任何混乱. 生成图形查询的步骤是可选的，这意味着用户可以让系统直接生成基于表单查询的图形查询语言.

3.4　查询图数据库.

一旦用户确认表单查询时，系统转换表单查询为图查询语言. 该系统采用的 Neo4 j 作为数据库[4]，使用 Cypher 作为查询语言. Cypher 已测试是处理图数据库最高效的查询语言. Cypher 语言具有高可读性和可维护性，能够被高效地开发使用.

系统基于图 1 文献实体中的概念模式的关系，将表单查询翻译为 Cypher 语言. 当将表单查询转换成图查询语言时，对在表单查询中没有直接显示的文献实体之间的连接进行探索和补充，以形成图形查询语言的完整路径. 转换成 Cypher 查询语言后，所生成的 Cypher 查询语言直接发送到 Neo4 j 图数据库中. 最后把查询结果返回给用户. 图 4 所示为系统生成的 Cypher 查询语言.

```
MATCH (cited_o:Affiliation)←(cited_a:Author),
      (cited_a:Author)→(cited_p:Paper),
      (cited_p:Paper)←(cited_s:Source),
      (cited_p:Paper)→(cited_t:Term),
      (citing_o:Affiliation)←(citing_a:Author),
      (citing_a:Author)→(citing_p:Paper),
      (citing_p:Paper)←(citing_s:Source),
      (citing_p:Paper)→(citing_t:Term),
      (cited_p:Paper)←(citing_p:Paper),
WHERE cited_o.name= 'Happy University'AND
      citing_t.name= 'NoSQL'AND
   citing_s.name= 'VLDB'
RETURN DISTINCT cited_a.name
```

图 4　系统生成的 Cypher 查询语言

4　实验与结果

为了测试所提出的系统的性能，尤其是使用 Neo4 j 图数据库[8]，我们构建了另一个系统，基于关系数据库作为底层数据存储，该系统具有相同的接口和功能，用这个系统去进行相同的查询任务，对其性能进行测试. 由于关系型数据库采用关系型数据模型，需要预先定义概念模型. 图 5 显示了关系的概念模型(E-R 关系图). 概念模型存储在 MySQL 数据库.

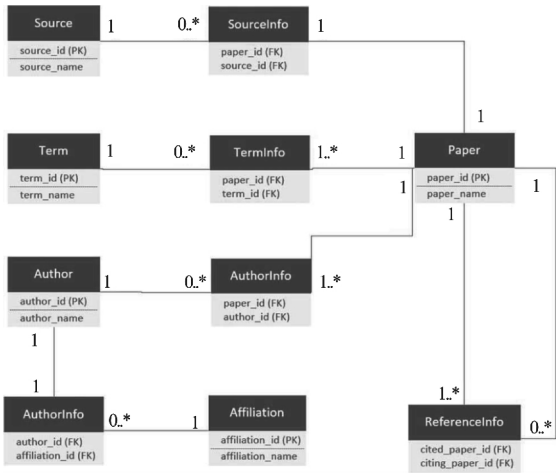

图 5　关系的概念模型

实验使用 ACM SIGKDD 会议上的公开数据集，表 1 表示数据集中的文献实体和关系的数量.

我们构建四组查询，从两个节点到五个节点的查询.

比较查询执行时间是一个好的方案，我们可以排除人的因素(例如，用户花费的时间生成查询)，并专注于系统内在要素，给定这两个系统具有相同的

接口. 测试环境是在 Windows7,64 位操作系统, Inter i5 处理器,500 GB 硬盘,16 GB 内存的电脑上. 如图 6 所示,关系数据库(MySQL)在执行查询有两个节点表现较好. 关系数据库执行的所有查询都小于 0.4 s,好于图数据库的 1 s.

表 1 数据集中的文献实体和关系的数量

实体	数量	关系	数量
Paper	628 725	Paper-Paper	631 531
Author	584 665	Paper-Author	1 308 561
Source	11 539	Paper-Source	520 534
Term	280 405	Paper-Term	5 386 834
Affiliation	1 000	Author-Affiliation	595 442

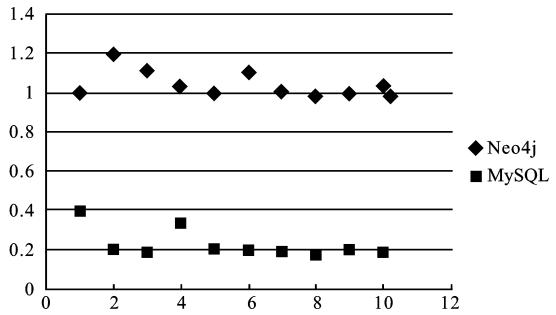

图 6 对 2 个节点查询时间对比

然而,如果查询变得复杂,对于三个节点查询,关系数据库的执行时间少于 4 s,而 Neo4j 图数据库则是 1s 左右(如图 7 所示).

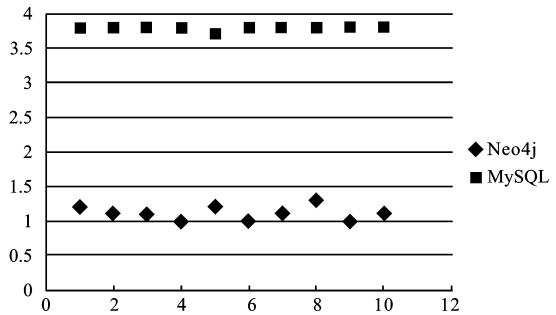

图 7 对 3 个节点查询时间对比

对于四个节点的查询,关系数据库的执行时间在 10 s 左右,而 Neo4j 图数据库则少于 2 s(如图 8 所示).

对于五个节点的查询,关系数据库的执行时间要花费半个小时,而 Neo4j 图数据库则依然少于 2 秒(如图 9 所示). 可见,对于复杂查询,Neo4j 图数据库性能更好.

影响执行效率的原因是在关系数据库中是将两个或多个表的记录连接起来,这严重影响了查询的

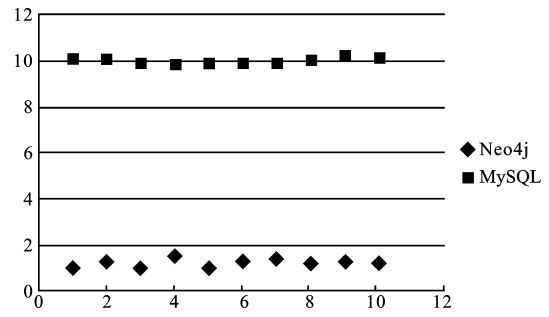

图 8 对 4 个节点查询时间对比

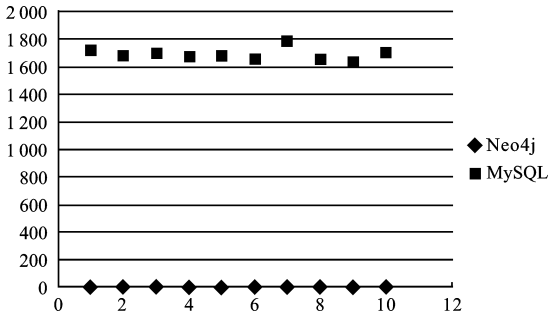

图 9 对 5 个节点查询时间对比

效率. 由于查询涉及更多的节点,连接操作是必要的. 在处理查询具有五个节点而言,关系数据库处理五个连接的操作. 因为每个表含有多条记录,处理这些操作是非常耗时. 另一个原因是索引所花费的时间. 关系数据库使用索引是为了方便快速检索. 虽然索引找到有用的数据很快,不过我们还需要更多的时间来遍历索引数据发现他们的关系. 这是因为关系记录两者之间是不明确的,并且需要遍历找出现有的连接. 在图形数据库,连接和遍历导致巨大的性能开销. 而图形数据库不是查找表; 图遍历执行的速度是常数,跟图的规模大小无关. 图形数据库不需要费时连接操作. 执行不同的查询组的图形数据库时间并不受节点的查询数量的影响. 基于图数据库的系统对不同查询组的结果,在执行时间上大体是相同的(如图 10 所示),相比于关系数据库,执行的时间分别减少了 71%(对于 3-节点查询),88%(对于 4 节点的查询),以及 99%的(5-节点查询). 实验证明,对于多节点的复杂查询,基于图形数据库设计的系统性能更好. 实验结果表明,基于关系数据库系统的系统对于五个节点的查询处理花费了 30 多分钟. 在现实中,这样的一个系统性能是不可忍受的. 目前已有的系统是采用分治思想将查询分成几个短的查询,合并子结果结果来解决复杂查询的问题.

图形数据库的一个好处是易于构建查询. 在图形数据库使用的查询语言是直观和容易构造的,因

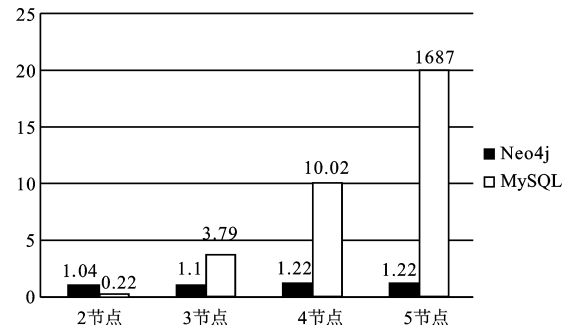

图 10　查询时间对比

为这些语言都是基于天然图形的特征. 表 2 是使用 Cypher 和 SQL 对五个节点的查询，可看出，使用 Cypher 语言查询更简洁. 所以系统响应时间也更快.

表 2　使用 Cypher 和 SQL 对五个节点的查询语句

Neo4j(Cypher)	MySQL(SQL)
MATCH (cited_o: Affiliation) <-(cited_a: Author)WHERE ->(cited_p: Paper) <-(citing_p: Paper) -> (citing_t: Term) WHERE citing_t = 'Database' RETURN cited_o. name	SELECT affiliation. affiliation_name FROM affiliation, affiliation_info, author_info, reference_info, term_info, term affiliation. affiliation_id = affiliation_info. affiliation_id AND affiliation_info. author_id = author_info. author_id AND author_info. paper_id = reference_info. cited_paper AND reference_info. citing_paper = term_info. paper_id AND term_info. term_id = term. term_id AND term. term_name = 'Database'

5　结束语

本文讨论了基于图数据库的文献检索系统的设计与实现. 首先，对目前存在的文献检索系统中存在的不足进行了简要介绍，针对已有的不足，提出了基于图数据库的设计，并对图数据库、系统架构、系统使用进行了介绍. 最后，实验对比并验证该系统与传统的基于关系型数据库相比，有着简洁和互动界面、良好的用户体验，能处理复杂的查询，具有更好的查询速度，为基于图数据库开发的应用提供了借鉴作用.

参考文献：

[1] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.

[2] 路莹，罗荣庆，王青春，等. 基于图形数据库 Neo4J 的合著网络研究与实践[J]. 中华医学图书情报杂志，2016,26 (4)：13-16.

[3] Jordan G. Spring data Neo4j，practical Neo4j[M]. Springer,2014：261-318.

[4] Holzschuher F，Peinl R. Performance of graph query languages：comparison of cypher，gremlin and native access in Neo4j[J]. Joint Edbt/icdt Workshop Graphq，2013，1(1)：195-204.

[5] Kaliyar R K. Graph databases：A survey[C]// Computing，Communication & Automation (ICCCA)，2015 International Conference on. IEEE，2015：785-790.

[6] Singh M，Kaur K. SQL2Neo：Moving health-care data from relational to graph databases[C]// Advance Computing Conference (IACC)，2015 IEEE International. IEEE,2015：721-725.

[7] Zhu Y，Yan E，Song I Y. The use of a graph - based system to improve bibliographic information retrieval：System design，implementation，and evaluation[J]. Journal of the Association for Information Science and Technology，2016.

[8] Miller J J. Graph database applications and concepts with Neo4j[C]// Proceedings of the Southern Association for Information Systems Conference. Atlanta，GA，USA，2013.

作者简介：

林启胜　男，(1992-)，硕士研究生. 研究方向为大数据存储应用、大数据分析. Email：a2950240@163. com.

王　磊　男，(1974-)，博士，研究员. 研究方向为计算机及应用、大数据分析.

周　喜　男，(1978-)，博士，研究员. 研究方向为物联网应用技术、大数据分析.

赵　凡　男，(1980-)，博士研究生，副研究员. 研究方向为信息安全、大数据分析.

马　博　男，(1984-)，博士研究生，副研究员. 研究方向为计算机应用技术.