文章编号：1007-1423(2016)07-0076-08 DOI：10.3969/j.issn.1007-1423.2016.07.018

基于图数据库的电影知识图谱应用研究

陆晓华[1]，张宇[2]，钱进[3]

（1. 四川大学计算机学院，成都 610065；2. 成都航空职业技术学院，成都 610065；
3. 重庆市通信服务产业有限公司移动服务分公司，重庆 404100）

摘要：
知识图谱是一种基于图的数据结构，由节点和边组成，其本质上属于语义网络。近年来，伴随着大数据概念的提出，知识图谱已经成为是当前的研究热点。由于非结构化文本的知识提取和数据可视化这两方面的技术难点，目前知识图谱应用主要局限于搜索引擎和问答系统等方面。着眼于电影知识图谱的设计与实现，通过引入图数据库 Neo4j，为知识图谱的实现提供一种新的思路。

关键词：
知识图谱；图数据库；Neo4j

0 引言

知识图谱，也被称为科学知识图谱、知识域可视化或知识域映射地图，是显示科学知识的发展进程与结构关系的的一系列各种不同的图形。它用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系[1]。

具体来说，知识图谱是把应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，用可视化的图谱形象地展示学科的核心框架、发展历史、前沿领域以及整体知识架构的多学科融合的一种研究方法。它把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律，为学科研究提供切实的、有价值的参考[1]。

近些年，随着大数据时代的到来，知识图谱已经在其他领域有所应用。Google 早在 2012 年就发布了“知识图谱”，利用知识图谱将 Google 的搜索结果进行知识系统化。当用户在搜索某一关键词时，Google 就会在搜索结果的右边给出该关键词相关的信息，极大地方便了人们对信息的搜索。2013 年 2 月，百度也推出了自己的知识图谱。不同于基于关键词搜索的传统搜索引擎，知识图谱可用来更好地查询复杂的关联信息，从语义层面理解用户意图，改进搜索质量。例如在百度的搜索框里输入“马云”的时候，搜索结果页面的右侧还会出现与“马云”相关的人物，如图 1(a)所示；另外，对于包含逻辑关系的搜索语句例如 “马云妻子”，百度能准确返回他的妻子“张瑛”，如图 1(b)所示。这就说明搜索引擎通过知识图谱真正理解了用户的意图。

知识图谱的构建主要包括知识单元的构建、知识单元间关系的构建和知识的可视化三个部分。其中前两个部分是构建知识图谱的最基本任务。以往的知识图谱研究多基于文献来进行研究，将关键词、摘要等结构化信息可以直接作为知识单元进行构建。而大数据概念的兴起，将研究者的目光集中到互联网的海量数据上来。这其中的信息多为非结构化的文本，而且还含有大量杂讯，要对这些信息进行语义分析，提取出能用于可视化知识图谱的知识单元并抽取出知识单元之间的关系就相当复杂了。另外，传统的关系型数据库经历几十年的发展，虽然具备较高的安全性和数据一致性，能够依赖简单的数据结构表达丰富的语义信息，但是对于知识图谱这样连接相对丰富，查询复杂的数据结构，效率上考虑已经不适用了。在本文中，将对 IMDB

(a)　　(b)

图 1　百度搜索知识图谱应用

数据进行抓取并抽取出命名实体和实体关系，并通过 Neo4j 图数据库建立一个电影知识图谱。

1　知识图谱构建流程

知识图谱的构建流程[2]通常包含下面几个重要的环节：构建知识单元、单元关系抽取以及结构化展示。在实现上，流程通常如图 2 所示。通常我们把数据获取和数据清洗归结为数据准备阶段。构建知识单元的操作主要为提取文本中的命名实体信息；单元关系抽取主要是抽取出上一步提取出的命名实体之间的关系；结构化展示即为利用数据可视化技术对提取出的实体和关系进行可视化处理。

在实现上，构建知识图谱通常首先会从维基百科、百度百科等资源中提取所需内容。本文的系统使用的电影及电影人数据来自于 IMBD 网站。利用爬虫技术从互联网空间中抓取的文本包含 HTML 标签等杂讯，需要进行数据清洗。数据准备完成之后，我们通过统计机器学习算法提取文本中的命名实体，继而通过特殊的正则模式匹配找出实体之间的关系，并将其持久化为 csv 文件。最后，我们将所有命名实体及实体关系导入 Neo4j 图数据库，以供数据可视化及知识图谱内部联系的查询。

1.1　数据来源

基于目前的研究和技术，通常的知识图谱具有以下几种类型[3]：(1)领域无关的知识图谱；(2)特定领域的知识图谱；(3)跨语言的知识图谱。其中特定领域的知识图谱，虽然内容不及领域无关知识图谱广泛，但是能够囊括特定领域中的知识内容，更具有针对性，所以在特定领域中具有很好的应用。例如，宜信将知识图谱技术成功应用在互联网金融领域，创立了全球首个基于金融知识图谱的金融云平台，为客户提供个性化的金融服务，取得了很好的效果。

知识图谱类型的多样化导致了知识图谱构建方法的多样化，一般来说，根据知识图谱数据来源划分，又可以将知识图谱构建的方法分为基于网络百科资源的知识图谱构建方法、基于结构化数据的知识图谱构建方法、基于半结构化数据的知识图谱构建方法和基于非结构化数据的知识图谱构建方法。

本文的系统数据采集自 IMDB（互联网电影数据库)电影资料库。IMDB 是一个关于电影、电影演员、电视节目、电视明星、电子游戏和电影制作小组的在线数据库。它是目前全球互联网中最大的一个电影资料库，里面包括了几乎所有的电影，以及 1982 年以后的电视剧集。我们通过 IMDB 的电影及演员介绍页面采集各类实体信息，如图 3 所示分别为 IMDB 电影页面和演员页面。通过对页面标签的正则匹配，我们可以提取出电影中的演职人员名单及其对应的角色；同样，对于演员页面可以提取出其参与拍摄的电影及其饰演的角

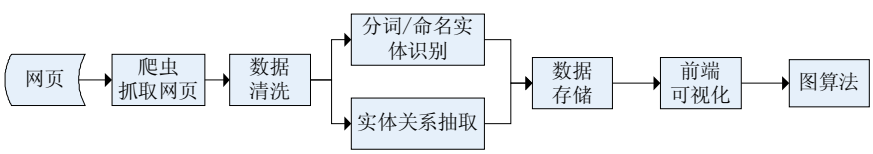


图 2　知识图谱构建流程

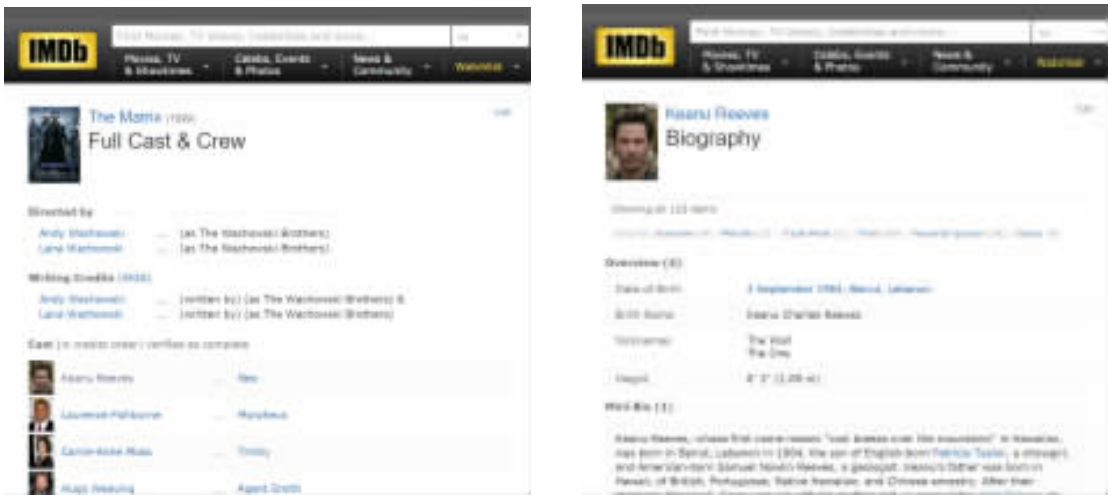


图 3 IMBD 电影及人物页面

色。由此,我们可以得到演员–饰演–角色的关系。另外,对于电影而言,还可以抽取出例如电影分级、发行时间、发行公司、电影类型、电影评分等属性;同样,演员还有身高、生日、性别等人物属性。

1.2 命名实体识别

知识图谱构建流程中的知识单元构建通常是指提取文本中命名实体的识别。命名实体识别(Named Entity Recognition,简称 NER),又称作“专名识别”,是指识别文本中具有特定意义的实体,主要包括人名、地名、机构名、专有名词等。命名实体识别技术是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。

基于统计机器学习的命名实体提取方法通常包括:隐马尔可夫模型(Hidden Markov Model,HMM)、最大熵 (Maxmium Entropy,ME)、支持向量机(Support Vector Machine,SVM)、条件随机场 (Conditional Random Fields,CRF)[4]。在这 4 种学习方法中,最大熵模型结构紧凑,具有较好的通用性,主要缺点是训练时间复杂性非常高,有时甚至导致训练代价难以承受,另外由于需要明确的归一化计算,导致开销比较大。而条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架,但同时存在收敛速度慢、训练时间长的问题。一般说来,最大熵和支持向量机在正确率上要比隐马尔可夫模型高一些,但是隐马尔可夫模型在训练和识别时的速度要快一些,主要是由于在利用 Viterbi 算法求解命名实体类别序列的效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用,如短文本命名实体识别。

在本文的系统实现中,我们的命名实体提取使用的是 NLTK[10]库中的最大熵算法。命名实体提取示例如图 4(a)所示。NLTK 的命名实体识别使用的是 MaxEnt 分类器,其工作时有两个原则:①总是试图保持均匀分布(即最大化熵);②保持其统计概率与经验数据一致。NLTK 提供了一个持久化的 pickle 文件,即为通过手动标记语料库训练出的 MaxEnt 分类器实例。

1.3 实体关系抽取

在知识图谱构建过程中,单元关系抽取通常在命名实体提取之后进行,就是提取出命名实体之间的联系。基于目前的研究,已经有许多关系实体关系抽取方法被应用在各种实验系统当中。这些方法所遵循的技术方法基本可以归纳为:基于模式匹配的关系抽取、基于词典驱动的关系抽取、基于机器学习的关系抽取、基于 Ontology 的关系抽取以及混合抽取方法[5]。

在关系抽取研究领域,普遍使用基于模式匹配的关系抽取方法。这种抽取方法通过运用语言学知识,在执行抽取任务之前,构造出若干基于语词、基于词性或基于语义的模式集合并存储起来。当进行关系抽取时,将经过预处理的语句片段与模式集合中的模式进行匹配。一旦匹配成功,就可以认为该语句片段具有对应模式的关系属性。

(a)NLTK 命名实体识别　　(b)NLTK 实体关系抽取

图 4

在本文的系统中,一旦我们提取出命名实体,就可以基于模式匹配提取出它们之间的关系。如前所述,我们通常会寻找指定类型的命名实体之间的关系。进行这一任务的方法之一是首先寻找所有 (X,α,Y) 形式的三元组,其中 X 和 Y 是指定类型的命名实体,α 表示 X 和 Y 之间关系的字符串。NLTK 提供了特殊的正则匹配方式,可以方便对词性、命名实体类别等进行模式匹配,提取出我们感兴趣的元组。图 4(b)示例演示了使用 NLTK 抽取组织–地名关系的过程。同样,我们可以使用 NLTK 编写各种模式匹配抽取出人–人、人–电影之间的关系。

2　图数据库设计

在数据存储领域,关系模型曾经是数据存储的主流,近年来逐渐被 NoSQL 数据库取代。NoSQL,泛指非关系型的数据库,通常分为键值(Key–Value)存储数据库、列存储数据库、文档型数据库和图数据库。图 5(a)是来自 db–engines 网站的统计数据,展示了 2013 年以来各类数据库系统的使用情况,其中,图数据库的使用率上涨了 5 倍多。

图形数据库中每个对象是一个节点,而对象之间的关系是一条边。相对于关系数据库来说,图形数据库善于处理大量复杂、互连接、低结构化的数据,这些数据变化迅速,需要频繁的查询——在关系数据库中,由于这些查询会导致大量的表连接,从而导致性能问题,而且在设计使用上也不方便。图形数据库适合用于社交网络,推荐系统等专注于构建关系图谱的系统。图数据库用图来存储数据,是最接近高性能的一种用于存储数据的数据结构方式之一。

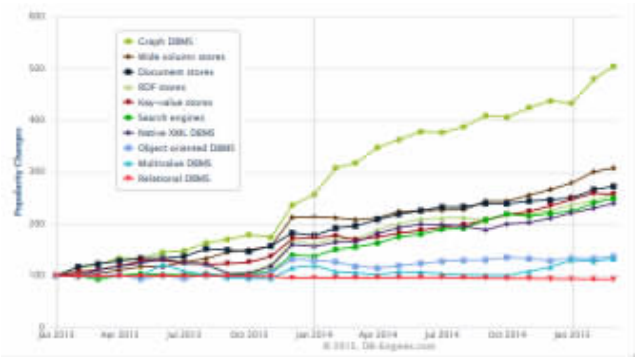
知识图谱是基于图的数据结构,它的存储方式主要有两种形式:RDF 存储格式和图数据库[6]。图数据库的代表有 Neo4J、Titan、OrientDB、DEX、AllegroGraph、GraphDB 等。图 5(b)展示了目前比较流行的基于图存储的数据库性能情况。基于 S Jouili 的研究,Neo4j[7]在存储查询等方面性能均优于其他图数据库,在工业上具有广泛的应用[8]。

通常,现实生活中的实体和关系普遍都比较复杂当然,而且常常查询涉及到 1 度以上的关联查询,如果使用关系型数据库存储知识图谱会形成性能瓶颈。对于复杂的关系网络,基于图数据库存储优势非常明显。首先,在关联查询的效率上会比传统的存储方式有显著的提高。当涉及到 2~3 度的关联查询时,基于知识图谱的查询效率会比关系型数据库高出几千倍甚至几百万倍。其次,基于图的存储在设计上会非常灵活,一般只需要局部的改动即可。例如我们有一个新的数据源,我们只需要在已有的图谱上插入就可以。与此相反,关系型存储方式灵活性方面就比较差,它所有的 Schema 都是提前定义好的,如果后续要改变,它的代价就非常高。最后,把实体和关系存储在图数据结构是一种符合整个故事逻辑的最好的方式。

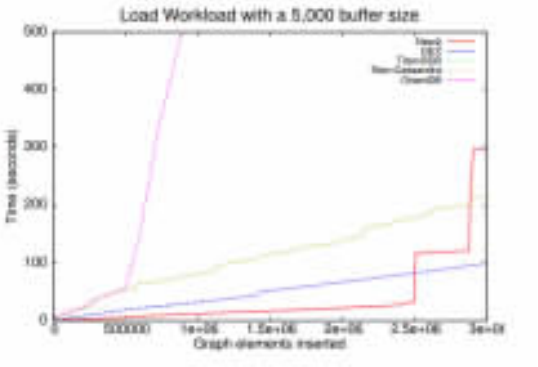
在本文的系统中,我们设计的节点主要有两种类型,分为 Movie 和 Person 类型,而关系类型主要有 4 种,分别为:ACTED_IN、DIRECTED、PRODUCED、WROTE。具体的,Movie 还有电影分级、发行时间、发行公司、电影类型、电影评分等属性;而 Person 也有身高、生日、性别等属性。此外,所有的关系都是有向边,例如 ACTED_IN 就是一条有 Person 指向 Movie 的边,其属性是演员在该电影中饰演的角色名。

3　系统实现及实验结果

本文的系统主要使用 Python 实现,主要分为 3 个



(a)图数据库应用发展



(b)各类图数据库性能比较

图 5

模块:数据采集模块、数据提取模块以及 Neo4j 图数据库导入模块。数据采集模块主要是通过 Python 的 urllib2 库爬取相关的 IMDB 网页,然后使用 BeautifulSoup 库清洗掉网页中的 HTML 标签,这样得到的纯文本数据以供后面的数据提取工作。命名实体的识别和实体关系的提取主要是通过 Python 的 NLTK 库实现。我们将识别出的命名实体及关系分别存储为 csv 文件,以方便后续导入 Neo4j 图数据库。在最新版本的 Neo4j 系统中,提供了一个大规模并行的可伸缩 csv 导入工具,该工具为 Neo4j 目录./bin/neo4j-import。在使用 neo4j-import 时,需要将待导入的 csv 文件表头定制为指定格式——显示地节点指定 ID 和 LABEL 以及边的 START_ID 和 END_ID 等。

本文所构建的电影知识图谱示意如图 6 所示,其中,(a)图为全量数据可视化之后的局部截图,(b)为随机查询的 25 条边视图以及它们之间的联系;(c) 为随机查询的 25 个节点视图以及它们之间的联系。从我们构建的电影知识图谱,可以非常容易地分析电影节点及电影人节点,以及它们之间的关系,推理出演员之间是否认识或者间接认识,从而推断出是否存在合作的可能等。

Neo4j 系统提供了名为 Cypher 的查询语言。Cypher 是一种可以对图形数据库进行查询和更新的图形查询语言,它类似于关系数据库的 SQL 语言。Cypher 的语法并不复杂,但是它的功能却非常强大,它可以实现 SQL 难以实现的功能。在本文的知识图谱中,我们可以通过编写 Cypher 查询语句,探索数据内部的关系。例如,六度分割理论中曾指出任何两个人之间所间隔的人不会超过六个。只要数据足够完整,采用 Cypher 可以很容易地找到任何两个人之间是通过哪些人联系起来的,而这一点是 SQL 很难实现的。

程序 1 所示的 Cypher 语句,可以查询 Kevin Bacon 和 Meg Ryan 之间到最短路径,如图 8(a)所示:Kevin Bacon 和 Tom Cruise 合作出演过电影 A Few Good Men;而 Tom Cruise 和 Meg Ryan 通过 Top Gun 结识。

程序 1 Cypher 查询最短路径

```
MATCH p=shortestPath(
  (bacon:Person {name:"Kevin Bacon"})-[*]-(meg:Person
{name:"Meg Ryan"})
)
RETURN p
```

相似地程序 2 所示的查询语句,可以查询到 Tom Hanks 和 Tom Cruise 两位演员之间相距 1 跳的联系。查询结果如图 8(b)所示。

程序2 Cypher 查询 2 度联系

```
MATCH (
tom:Person {name:"Tom Hanks"})-[:ACTED_IN]->(m)<-
        [:ACTED_IN]-(coActors),
(coActors)-[:ACTED_IN]->(m2)<-[:ACTED_IN]-(cruise:
        Person {name:"Tom Cruise"})
    RETURN tom, m, coActors, m2, cruise
```

(a)

(b)

(c)

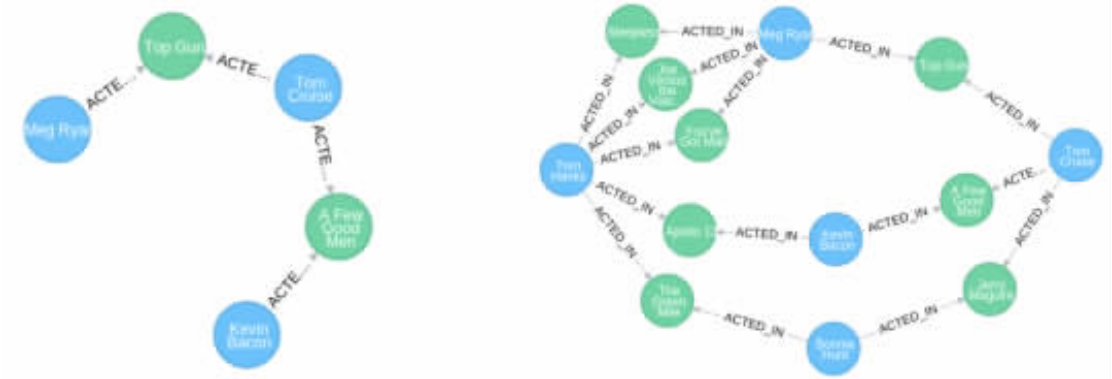
图 6 电影知识图谱示例

知识图谱的另一个应用是可以用于推荐系统。这其中，最著名的就是 Taher H .Haveliwala 设计的 PersonalRank 算法[9]：在计算所有节点相对于用户 u 的相关度时，PersonalRank 算法从用户 u 对应的节点开始游走，每到一个节点都以 1-d 的概率停止游走并从 u 重新开始，或者以 d 的概率继续游走，从当前节点指向的节点中按照均匀分布随机选择一个节点往下游走。这样经过很多轮游走之后，每个顶点被访问到的概率也会收敛趋于稳定，这个稳定的概率就可用进行排名。在本文的系统中，我们可以编写简单的 Cypher 语句给 Tom Hanks 推荐好友，推荐结果如表 1 所示。

基于 Neo4j 图数据构建的电影知识图谱系统，具

表 1 Tom Hanks 推荐结果

Recommended	Tom Cruise	Zach Grenier	Helen Hunt	Cuba Gooding Jr.	Keanu Reeves
Strength	5	5	4	4	4
Recommended	Tom Skerritt	Carrie-Anne Moss	Val Kilmer	Bruno Kirby	Philip Seymour Hoffman
Strength	3	3	3	3	3



(a) (b)

图 8 Cypher 查询示例

有界面友好,查询方便,应用广泛的特点,在工业上具有广阔的前景。本文虽然提出了基于图数据库的电影知识图谱研究实现,然而系统中还存在着许多不足和可以提高的地方。首先,我们只是提供了有限类型的边和点节点;此外,仅仅提供了基于 Cypher 的查询功能,还未集成相关的图算法。后续工作中我们将结合 Spark GraphX[11],运用图算法进行大规模的知识图谱分析,进而可以方便的实现社区发现、用户影响力、人群划分等功能。

参考文献:

[1]秦长江,侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报,2009(1):30-37,96.

[2]金贵阳,吕福在, 项占琴. 基于知识图谱和语义网技术的企业信息集成方法[J]. 东南大学学报:自然科学版, 2014(02):250-255.

[3]梁秀娟. 科学知识图谱研究综述[J]. 图书馆杂志, 2009(6):58-62.

[4]孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(6):42-47.

[5]徐健,张智雄,吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术, 2008(8):18-23.

[6]Abreu D D, Flores A, Palma G, et al. Choosing Between Graph Databases and RDF Engines for Consuming and Mining Linked Data[J]. Cold, 2013.

[7]Webber J. A Programmatic Introduction to Neo4j[J]. Addison Wesley Pub Co Inc, 2012:217-218.

[8]Jouili S, Vansteenberghe V. An Empirical Comparison of Graph Databases[C]. 2013 International Conference on Social Computing. IEEE Computer Society, 2013:708-715.

[9]Haveliwala T H. Topic-Sensitive PageRank: a Context-Sensitive Ranking Algorithm for Web Search[J]. Knowledge & Data Engineering IEEE Transactions on, 2003, 15(4):784-796.

[10]Loper E, Bird S. NLTK: The Natural Language Toolkit[C]. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1. Association for Computational Linguistics, 2002:63-70.

[11]Xin R S, Gonzalez J E, Franklin M J, et al. GraphX: a Resilient Distributed Graph System on Spark[C]. First International Workshop on Graph Data Management Experiences & Systems. ACM, 2013:1-6.

作者简介:

陆晓华(1988-),男,江苏苏州人,硕士,研究方向为机器学习、计算机视觉

张宇(1962-),教授,研究方向为模式识别,

钱进,工程师,研究方向为通信传输

收稿日期:2015-01-12 修稿日期:2016-02-25

Implementation of Movie Knowledge Graph Based on Graph Database

LU Xiao-hua[1], ZHANG Yu[2], QIAN Jin[3]

(1.College of Computer Science Sichuan University, Chengdu 610065
2.Chengdu Aeronautic Polytechnic,Chengdu 610065; 3. Chongqing Communcation Industry Services Co. Ltd.,Chongqing 404100)

Abstract:
Knowledge graph is a graph-based data structure, consisting of nodes and edges, and it is essentially a semantic network. In recent years, along with the proposed concept of big data, knowledge graph has become the current research focus. As technical difficulties of knowledge extraction of unstructured text and data visualization, the current applications of knowledge graph mainly limited in the aspects of search engine and Q/A system. Focuses on the design and implementation of movie knowledge graph, by the introduction of the Neo4j graph database, provides a new way of thinking for the realization of knowledge graph.

Keywords:
Knowledge Graph; Graph Database; Neo4j

(上接第 75 页)

An Improved Wavelet Threshold Denoising Method

LI Peng, YU Liang

(College of Computer Science, Sichuan University, Chengdu 610065)

Abstract:
In the process of acquisition, transmission and storage, digital image inevitably will be polluted by a variety of noise, this not only influence the visual effect of the image, but also will bring trouble in further processing and analysis of the image.Therefore, it is very important to remove the noise of the image.For soft and hard threshold wavelet denoising method exist some shortcomings, proposes an improved wavelet threshold denoising method, the denoising effect of this method is superior to the hard and soft threshold method and some other improved algorithms.

Keywords:
Image Noise; Soft and Hard Threshold; Wavelet Transform; Threshold Denoising