Journal of Computer Applications ISSN 1001-9081 2016-04-10
计算机应用,2016,36(4):992-996,1001 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2016)04-0992-05 DOI:10.11772/j.issn.1001-9081.2016.04.0992

自动化构建的中文知识图谱系统

鄂世嘉*,林培裕,向 阳
(同济大学 电子与信息工程学院,上海 201804)
(* 通信作者电子邮箱 eshijia1218@vip.qq.com)

摘 要: 为解决当前中文知识图谱构建的准确率低、耗时长且需要大量人工参与的问题,提出一种端到端基于中文百科数据的完整中文知识图谱自动化构建解决方案,并在此基础上开发实现了面向用户的中文知识图谱系统。在此方案中,通过自定义的网络爬虫,原始百科数据的词条属性以及相关的文本信息会不间断地被抓取到本地系统中,并以带扩展属性的三元组形式保存。后端系统则自动通过图数据库 Cayley 以及 MongoDB 数据库系统,对三元组文件数据进行导入,转换为庞大的知识图谱系统,从而在前端为用户提供丰富的基于知识图谱的应用服务。通过与其他知识图谱系统的比较,该方案在构建时间上明显减少,并且知识图谱中的实体及关系数量总规模高于 YAGO、知网(HowNet)和中文概念词典等中文知识图谱系统至少 50%。

关键词: 知识图谱; 网络爬虫; 三元组文件; 知识库; 图数据库
中图分类号: TP311.5 **文献标志码:** A

Automatical construction of Chinese knowledge graph system

E Shijia*, LIN Peiyu, XIANG Yang
(*College of Electronics and Information Engineering, Tongji University, Shanghai* 201804, *China*)

Abstract: To solve the problem that the methods currently used to construct Chinese knowledge graph system are time-consuming, have low accuracy and require a lot of manual intervention, an integrated end-to-end automatically constructed solution based on rich data from Chinese encyclopedia was proposed, and a user-oriented Chinese knowledge graph was implemented. In this solution, some property and related text information of the original encyclopedia data were scraped to local system uninterruptedly by the custom Web crawler, and saved as a triple with extended attributes. Through graph-oriented database Cayley and document-oriented database MongoDB, the data in the archived triple files was imported in the back-end system, and then converted to a huge knowledge graph system in order to provide various services dependent on the Chinese knowledge graph in the front-end system. Compared with other knowledge graph systems, the proposed system significantly reduces the construction time; moreover, the number of entities and relations is at least 50% higher than that of the other knowledge graph systems such as YAGO, HowNet and the Chinese Concept Dictionary.

Key words: knowledge graph; Web crawler; triple file; knowledge base; graph-oriented database

0 引言

一个典型的知识图谱通常包含着一系列概念、实例和关系[1],其为最有效的知识表达形式之一[2]。一些著名的知识图谱包括 Internet Movie Database、YAGO[3-4]、DBpedia[5-6]和 Freebase[7]。近几年来,大量的知识图谱已经被构建起来,并且有关知识图谱的话题也非常丰富,在工业界以及学术界呈现出日益增长的关注态势[8]。这一现象背后的重要原因主要是知识图谱已经逐渐被发现对于各种类型的应用都是至关重要的[9-10]。

尽管大量的知识图谱日益涌现,但当前大多公开的研究工作主要是孤立地强调了知识图谱构建环节的某一方面,诸如知识图谱中的数据表示、存储格式或知识获取方法等问题[11-12];此外,对知识图谱系统的维护及应用问题仍然没有有效解决;另外一个问题是,由于中文的语言特点,不能将语义网络[13]中处理英文的方法直接应用于中文文本处理以及进一步的语义提取。中文的句子结构并不像英文那样具有标准的格式。相反,中文句子结构中会经常省略一些语法结构,因而从非结构化的中文文本中直接自动获取有效的知识或事实是非常困难的。截止到目前,在业界仍然没有对于以上问题的成熟解决方案。

本文描述了一个真实的中文知识图谱构建过程,从知识库的组织、知识的获取、知识图谱数据的存储与维护以及知识图谱应用这四个角度重点了介绍了自动化构建中文知识图谱的完整流程;并通过与现有相关知识图谱系统的比较,证明了这种构建方法在构建速度以及系统规模上的优势。

收稿日期:2015-09-06;修回日期:2015-11-12。
基金项目:国家 973 计划项目(2014CB340404);上海市科委科研计划项目(14511108002)。
作者简介:鄂世嘉(1991—),男,辽宁大连人,博士研究生,CCF 会员,主要研究方向:云计算、知识图谱、大数据系统; 林培裕(1993—),男,江苏盐城人,硕士研究生,主要研究方向:知识图谱、大数据系统; 向阳(1962—),男,重庆人,教授,博士,CCF 会员,主要研究方向:管理信息系统、云计算、语义计算、大数据挖掘。

1　知识库的组织

一个知识库的后端通常由一系列概念 $C_1, C_2, \cdots, C_n$，针对每个概念 C_i 的实例 I_i，以及这些概念间的一组关系 $R_1, R_2, \cdots, R_m$ 组成。知识图谱中最核心的一种关系被称为“is-a”关系，其定义了某个概念 A 是概念 B 的一种(例如，“艺术家”是“人”)。“is-a”关系在概念 C_i 之上建立了一种分类系统。这一分类体系是一棵树，其中每一个节点表示了这些概念，每条边表示的是“is-a”关系本身，诸如 $A \rightarrow B$ 这样一条边表示概念 B 是概念 A 的一种。图 1 展示一个微型的中文知识库的组织方式，其详细阐明了上述符号的含义。

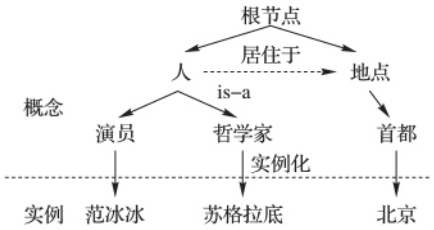

图 1　中文知识库示例

在通常的树型结构中，一个父节点(在分类树中)的一系列实例是其子节点实例的并集，但在本文构建知识图谱情境中，并不强制施加这种限制。因此，节点 A 可能含有并不属于任何 A 的子节点的实例。而且，一般的知识库也都包含许多领域完整性的限制。这些都需要专门的领域人员来进行专门的定义，即人工修正知识库。

2　知识的获取

将现实世界中零散的知识转换为一个结构化的中文知识库并不是一件容易的事情。下面对从网络百科中构建分类树和从百科中抽取实体关系两个关键步骤进行介绍。

2.1　构建分类树

2.1.1　爬取百科原始数据

我们开发并维护着一系列基于 Python Scrapy [14] 框架的爬虫，其能够从百度百科、互动百科等中文百科的页面中抽取需要的部分，爬虫的整体流程框架如图 2 所示。

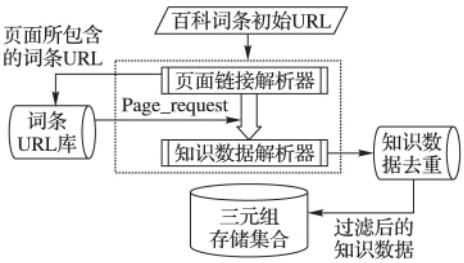

图 2　知识数据获取框架

系统首先对所提供的百科词条初始 URL 进行解析，解析模块分为两部分: 页面链接解析和知识数据解析。前者会基于定义的 HTML 规则将该页面中所包含的其他词条 URL 保存至词条 URL 库中。之后，系统会不断地对解析到的 URL 发出抓取请求，进而将获取到的数据发送至知识数据解析器，由该解析器负责在每个词条页面中获取所需的知识数据。将爬取到的数据存储在本地服务器中，使我们可以基于这些数据在其上构建具体模型，对数据作进一步处理。

2.1.2　构建百科分类树

在百科中主要有两类页面: 词条文章页面和类别页面，如图 3 所示。

图 3　百科中的两类页面、词条页面和类别页面

一个词条文章页面描述了一个实例，一个分类页面描述了一个概念。特别是在分类页面中，其列出了子类、父类以及相关的孩子节点(具体的词条)。我们解析这些页面来构建一张图，图中的每个节点指的是某个词条或某个类别，而图中的每一条边指的就是百度百科中类别 X 至其某个子类或者从某个类别 X 至一篇 X 类别下的词条文章。理想情况下，词条文章(实例)和类别(概念)应该可以形成一个分类树系统，正如图 1 所展示的那样，但是实际情况并不完全是这样，这个整体的图最终会形成许多环。因此本文利用 Tarjan 算法 [15] 进行剪枝，将该图转换为一个无环图。这个图中的另外一个问题是，从应用的角度来看，其最顶层的类别并不是我们所期望的。诸如“哲学家”和“疾病”这样的分类词条，往往都被埋没在图的深层次中。为了解决这一个问题，本文创建了一系列高层次概念，诸如“历史”和“人物”，之后将其放置在根节点孩子的位置上。本文称这些节点为顶点，如图 4 所示。接下来，将一些想要的百度百科类别放置在相应顶点的前几级子节点中。

图 4　修正的百科分类体系

2.2　抽取实体关系

一个知识库往往拥有着预先定义的关系，诸如“居住”(人，地点)和“写作”(作者，书)。这样一个关系的实例包含了概念的实例，彼此之间存在着一种概念的映射关系 [16]。例如，“居住”(姚明，上海)就是关系“居住”(人，地点)的一个实例。

理论上，我们可以试图去定义这样的一系列关系，之后再去抽取它们的实例 [17]。但是这样做会引起两个问题。首先，百科中包含着上亿条我们可能感兴趣的潜在关系，并且这个

关系集合每天都在发生着变化。因此,想要快速地定义大量有价值的关系是不切合实际的。其次,一个更为严重的问题是,从任何非结构化的纯文本中抽取关系实例显然是非常困难的,并且其所消耗的计算代价也是十分巨大的。

由于这些原因,本文采用了一种务实的方法。本文并不预先定义一系列关系,也不尝试去抽取它们的实例。取而代之的是,直接抓取两个概念在词条页面里所存在的任意关系实例。例如,假设“姚明”这一词条中有一个节标题称作“个人生活”,这其中提到了另一个词条页面——“叶莉”。那么接下来就可以创建一个关系实例——〈姚明,叶莉,个人生活〉,其表示姚明和叶莉之间有一个关系称作“个人生活”。

一般来说,抽取的关系实例具有以下形式〈概念实例 1 的名字,概念实例 2 的名字,表达两个实例间某种关系的文本〉。

2.2.1 从信息盒(InfoBox)中抽取关系

一个 Infobox 往往都处于词条页面的顶部,其对整个页面的内容进行了总结,并且提供了一些重要的统计信息(见图 3)。本文针对这样的页面结构,编写了一系列规则来从这样的 Infobox 中抽取关系实例。例如,图 5 所示的 InfoBox(“计算机”词条中的 InfoBox)。

中文名	计算机	第一台计算机	ENIAC
外文名	computer	别 称	电脑
时 间	1946年	发明者	约翰·冯·诺依曼

图 5 信息盒示例

针对以上所展示的 InfoBox,可以抽取到以下几种关系实例:

1) 〈计算机,计算机,中文名〉;
2) 〈计算机,computer,外文名〉;
3) 〈计算机,1946 年,时间〉;
4) 〈计算机,ENIAC,第一台计算机〉;
5) 〈计算机,电脑,别称〉;
6) 〈计算机,约翰·冯·诺依曼,发明者〉。

2.2.2 从词条文本中抽取关系

使用了一系里基于 HTML 的规则来从词条文本中抽取我们感兴趣的关系实例,如图 6 所示。例如,“牛顿第一运动定律”这个词条页面有一个节标题称作“定律定义”,其中提到了《自然哲学的数学原理》。那么通过这些内容,可以抽取到〈牛顿第一运动定律,自然哲学的数学原理,定律定义〉这样一个关系实例。其表示与《自然哲学的数学原理》的关系是“定律定义”,那么很直观地就可以理解“牛顿第一运动定律”是在《自然哲学的数学原理》中被定义的。

此外,本文还定义了很多解析 HTML 页面的规则来从列表、表格等页面结构中抽取关系,其原理与从 InfoBox 和从词条文本中抽取关系类似,这里就不赘述了。

图 6 从词条文本抽取关系

2.3 关系的优先级

通过使用一组相对并不庞大的规则集,可以从现有的中文百科平台中抽取到上亿级别的关系实例。将这些关系实例组织成了一个关系图,其中的节点表示概念实例,彼此之间的边表示这些点之间的关系实例。在使用这一关系图时,发现对一些特定的应用需求,一个相对较小的图是更有价值的。因此,有时候可能会需要对该关系图进行剪枝,以去除某些特定的关系。为了达到此目的,使用了一系列规则来对这些关系的实例赋予一定的优先级,那么在必要时就可以对一些优先级较低的边进行剪枝。

定义的规则首先会将 Infobox 中抽取的关系赋予较高的优先级,接下来为抽取的关系赋予的优先级依次降低:从词条文本中抽取的诸如像上文中提到的〈牛顿第一运动定律,自然哲学的数学原理,定律定义〉这样的关系、相互映射的关系实例、非相互映射的关系实例。

一个相互映射的关系实例是指的是,对于某种关系的源概念实例和目的概念实例,在我们的关系图中同样有一条边从该关系的目的节点指向源节点。直观地来说,我们认为这种关系的紧密性要强于非相互映射关系。例如,姚明和叶莉之间具有相互映射关系,因为在各自的词条页面中,都有提到对方的词条。但另一方面,姚明的页面中提到了威廉王子,但是在威廉王子的页面中,却并没有提到姚明。因此,姚明和威廉王子之间的关系是非相互映射关系,进而我们认为这种关系是不紧密的。

3 知识图谱数据的存储与维护

3.1 带有扩展属性的三元组存储

从网络上抓取的事实数据通过具有扩展属性的三元组文本存储在本地服务器中,文本中的每一条记录以〈Subject, Predicate, Object, Label〉表示。与传统普通三元组文件相比,具有扩展属性的三元组能够在简洁有效地表达关系的基础上,通过 Label 标签,丰富关系的元数据信息。根据不同的应用需求,Label 标签可以表达多种含义以及拥有不同的数据结构:其可通过文本语义理解的形式,精准表示三元组关系的时间属性。目前我们所构建的知识图谱系统以网络爬虫抓取该关系的时间作为 Label 的值;此外,Label 标签中还可保存关系的可信度,从而可以对关系进行可行性排序。

3.2 知识图谱数据的维护

为了让该数据能够更加灵活地被使用,在知识图谱构建过程中,很重要的一个环节就是对已获取的知识数据进行维护。其整体流程如图 7 所示。

从数据结构的角度考虑,知识图谱代表了一张巨大的关系图,而三元组文本形式的事实数据则对应关系图中的边。因此,本文采用图数据库 Cayley 与文档型数据库 MongoDB 对三元组数据进行持久化的存储,并不断根据抓取到的三元组文件对数据进行更新。MongoDB 作为 Cayley 的后台数据库,对于三元组文本形式数据的 CRUD 操作都由 Cayley 向 MongoDB 发出请求来完成。

对于每一条三元组形式数据,MongoDB 均以带有扩展属性的三元组形式的文档记录存储在由 Cayley 自动创建的集合“quads”中。通过第 2 章所述的时间属性,保留事实数据的时序性。

根据 Predicate 形式的不同,三元组数据主要被分为以下三类:1) Predicate 值为“is-a”或“instance-of”,该类三元组数据

构成了整个知识库的分类体系; 2) Predicate 值的前缀为“attribute:”,该类三元组数据表示某一概念实例 Subject 的属性 Predicate 的值为 Object; 3) Predicate 值为除前两种之外的普通字符串,该类三元组数据在知识库中的数据量最大。

如图 7 所示,网络爬虫不间断地抓取百科页面词条的 InfoBox 或相关的文本信息,以三元组文本形式将知识数据在本地存储起来。数据导入程序采用多线程的方式不间断扫描三元组文件,并在数据导入层进行了去重处理,保证导入到 MongoDB 中数据的唯一性。基于此套流程,知识图谱系统的后端就可以在无人工操作监督的情况下快速构建,使得数据量不断累积,从而使得知识图谱覆盖的知识领域更加丰富。

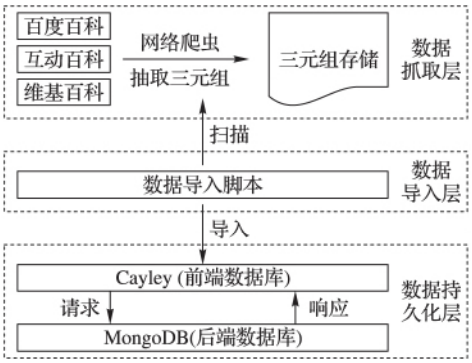

图 7　知识图谱数据的维护框架

4　知识图谱的应用

在从百科中构建了最初的中文知识图谱后,只相当于对互联网中的数据进行了基本的格式化处理,形成了一份归档,但对于普通用户来说,仍然是无法很好使用的一份资源。因此,当前很多的知识图谱系统仅局限于为专业人员进行二次开发而使用,没有充分体现出知识图谱对于普通用户的巨大实用价值。基于本文介绍的知识图谱构建方法,我们为普通用户在 Web 前端实现了面向内容理解的中文知识图谱(Content Understanding-Knowledge Graph, CU-KG)系统,其拥有丰富的应用界面,从而让用户透明地使用及完善中文知识图谱。

4.1　关系搜索

在 CU-KG 中,目前用户可以针对知识图谱所包含的海量关系展开搜索。通过在网页中输入两个实体的名称,即可获得两个实体间的多种关系,如图 8 所示。

图 8(b) 展示了针对用户输入的两个实体“周克华”与“重庆市”之间的关系。可以看到针对这两个直接实体,通过本文构建的知识图谱系统,不仅可以找出实体间的直接关系,还可以找出与直接实体相关联的中间实体(图中的圆点实体),以及他们之间的关系,如“周克华杀害了朱彦超”这样的关系实例。

4.2　知识库的多维搜索

与传统知识图谱不同,CU-KG 为用户提供了多维的知识库搜索功能,用户通过搜索实体名字,可以从多种属性的角度来浏览实体的相关信息; 并且在知识库中融合了多种数据类型,除了基础的文本信息,还有图片、视频数据。这些多源异构的数据通过在三元组中 Predicate 的 attribute 前缀进行标识。例如〈周克华, attribute: 图片, URL〉在系统中将实体“周克华”的图片通过统一资源定位符(Uniform Resource Locator, URL)找到其在本地存储的位置,进而在前端进行统一的展示,视频数据的表达也是如此。用户可以选择按实体、关系以及属性浏览等多种方式全角度理解实体的各种特征,如图 9 所示。

(a) 用户输入接口

(b) 针对用户输入的搜索结果

图 8　关系搜索中的用户输入和搜索结果

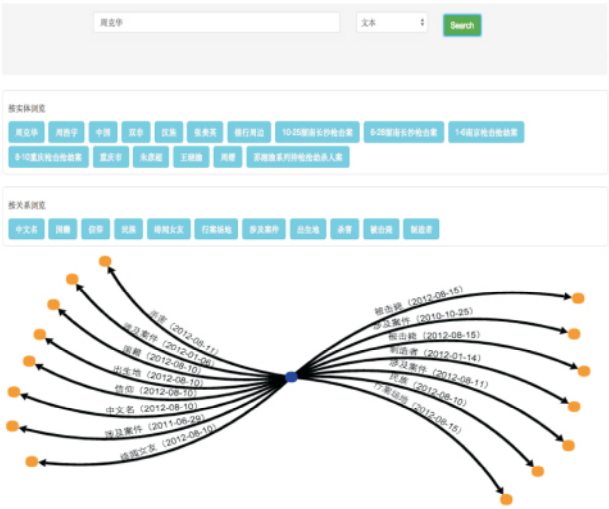

图 9　知识库的多维搜索示例

4.3　带时序的实体关系演化

通过采用扩展的三元组对知识数据进行组织,其保留了关系的时间维特征。用户可在 CU-KG 中了解某个时间范围内实体之间的关系随着时间的推移而发生的变化,以实体“周克华”在“2012-08-10”至“2012-08-13”之间的关系演化为例,“2012-08-10”的初始关系如图 10 所示。

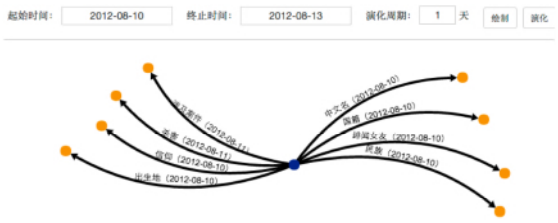

图 10　“周克华”实体“2012-08-10”的关系图

若选择演化周期为 1 d,通过不断点击演化按钮,关系图会因时间的推移,而有新的关系加入进来,图 11 为演化至

“2012-08-13”时的关系图，若某一天有新的关系产生，则会以不同类型的连线在图中进行标识。

图 11 “周克华”实体关系演化后的关系图

可以看到，新产生的关系为“2012-08-11”生成的。

4.4 知识图谱系统的应用前景

知识图谱系统对现实世界中零散的知识数据进行了组织。基于中文知识图谱系统，除了上文介绍的部分知识图谱的示例应用，其在问题理解、深度信息搜索等领域都可以有广泛的应用。例如，在智能问答领域，对于问题“周克华的女友出生于中国的哪个省份”，经依存语义分析，并结合知识图谱中现有的〈周克华，绯闻女友，张贵英〉、〈张贵英，出生地，四川省〉、〈中国，省份，四川省〉等关系实例，可得到该问题的答案为“四川省”。这种基于知识图谱的问答方法可以对原有的问题系统解决方案提供良好的补充，提升问答系统的准确率。

5 讨论与评价

构建通用的知识图谱是一项系统而又庞大的工作。现有的中文知识图谱中，更常见的是面向领域的知识图谱，由于其所覆盖的知识数据相对通用知识图谱来说较少，因而更容易构建。而截止到目前，在中文知识图谱领域还没有公开的标准通用知识图谱系统。通过我们实践发现，领域知识图谱的构建可以在通用知识图谱构建的基础上来完成。因为当通用知识图谱具有一定规模后，通过对其中的数据分类，形成多种类别的子知识图谱，进而再加上一定专业领域人工修正，即可形成较为准确有效的领域知识图谱。

对于现有知识图谱系统的评价，可以将系统中的关系与开放的诸如 Freebase、WordNet 等百科知识库数据进行对比来进行评价。但对于中文知识图谱，现在仍然缺少相对标准的事实关系库，并且现有的评价方法主要是对抽取的关系数据进行人工的标注与验证。本文的目的是构建真实高可用的中文知识图谱系统，CU-KG 知识图谱系统的所有事实数据都是直接基于百科的半结构化信息所获取的，百科类数据是目前中文知识图谱构建领域中最高可信度的数据源，因此其准确率可以得到充分保证，且通过对词条页面的超链接分析，对同义词列表进行了区分，避免了歧义性的问题。

从非结构化数据中直接抽取可能的三元组事实是当前研究的方向之一，主要是依赖中文语法分析、语义依存技术进行内容分析。对于此类关系需要与百科中获取的标准关系进行比较，对关系的准确率和召回率进行评价。但所获得知识的准确率将难以满足可用性的要求，因此目前还不能大规模应用于中文知识图谱的自动化构建过程中。

CU-KG 知识图谱系统的当前规模及与相关公开的知识系统的对比如表 1 所示。

表 1 CU-KG 与其他带有中文的知识图谱系统对比

知识图谱名称	实体数量	事实数量
WordNet	155 287	479 887
YAGO	1 054 638	5 000 000
HowNet	191 924	462 433
北大中文概念词典	66 590	—
CU-KG(本文的知识图谱)	3 399 615	5 651 614

从表 1 可看出，CU-KG 相对于其他中文知识图谱在实体数量和事实数量这两个基本指标上优势明显，系统总规模平均约比同类高出至少 50%。伴随着新的实体以及更多的事实数据不断地被添加到知识图谱系统中，CU-KG 的系统规模优势将越来越大。

在构建时间方面，根据能获得的数据，当前业界在知识图谱的初始构建时间达 12.5 h[18]，而通过本文所介绍的知识图谱自动化构建流程，能够几乎随时地将互联网知识数据转换至知识图谱系统中，无需进行长时的等待。

除此之外，CU-KG 充分利用了多源异构的大数据资源，在应用过程中，将文本、图像和视频有机地融合在一起，从多个角度向用户展示了知识，弥补了传统知识图谱的不足。

6 结语

本文详细地阐明了一种自动化构建中文知识图谱系统的方法，该方法基于可信度高的数据源，使用定制的知识图谱数据维护框架，持续地从网络数据中抓取知识数据并以带扩展属性的三元组的数据格式进行中间存储，将三元组数据通过图数据库技术对知识数据进行持久化存储，并对所实现的 CU-KG 知识图谱系统的应用进行了充分地说明。实践结果表明，该构建方法能够快速构建大规模的中文知识图谱系统，并在实体和关系的数量上有着较为明显的优势。

在当前的 CU-KG 系统中，由于采用了 NoSQL 技术，现有的数据存储方式对于传统的图路径搜索支持并不是特别理想，对于复杂的关系路径查询所消耗的时间依然较高，后期仍然需要研究相关算法，来优化关系搜索所要消耗的时间。并且在未来可以扩展 CU-KG 所依赖的知识数据源，从而充分丰富现有知识图谱的实体及关系数量。

参考文献：

[1] LENAT D B. CYC: A large-scale investment in knowledge infrastructure [J]. Communications of the ACM, 1995, 38(11): 33-38.

[2] SINGHAL A. Introducing the knowledge graph: things, not strings [EB/OL]. [2014-10-10]. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html#!/2012/05/introducing-knowledge-graph-things-not.html.

[3] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge [C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 697-706.

[4] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a large ontology from Wikipedia and WordNet [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217.

[5] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a Nucleus for a Web of Open Data [M]. Berlin: Springer, 2007: 722-735.

[6] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia — a crystallization point for the Web of data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165.

(下转第 1001 页)

起来,实现算法的 MapReduce 化; 另一方面,可以将该算法的思想拓展到最大、闭频繁项集的挖掘领域。

致谢　非常感谢邓志宏教授提供的 PrePost 算法的代码。

参考文献:

[1] AGRAWAL R, IMIEILNSKI T, SWAMI A. Mining association rules between sets of items in large databases [C] // Proceedings of 1993 ACM SIGMOD Conference on Management Data. New York: ACM, 1993: 207 - 216.

[2] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules [C] // VLDB 1994: Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 1994: 487 - 499.

[3] LIN K C, LIAO I E, CHEN Z S. An improved frequent pattern growth method for mining association rules [J]. Expert Systems with Applications, 2011, 38(5): 5154 - 5161.

[4] GUPTA R, SATSANGI C S. An efficient range partitioning method for finding frequent patterns from huge database [J]. International Journal of Advanced Computer Research, 2012, 2(2): 62 - 69.

[5] 李也白, 唐辉, 贺玉明. 基于改进的 FP-tree 的频繁模式挖掘算法 [J]. 计算机应用, 2011, 31(1): 101 - 103. (LI Y B, TANG H, HE Y M. Frequent pattern mining algorithm based on improved FP-tree [J]. Journal of Computer Applications, 2011, 31(1): 101 - 103.)

[6] SUCAHYO Y G, GOPALAN R P. CT-PRO: a bottom-up non recursive frequent itemset mining algorithm using compressed FP-tree data structure [C] // FIMI 2004: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations. Piscataway, NJ: IEEE, 2004: 212 - 223.

[7] ZAKI M J, GOUDA K. Fast vertical mining using diffsets [C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York: ACM, 2003: 326 - 335.

[8] LI Z F, LIU X F, CAO X. A study on improved Eclat data mining algorithm [J]. Advanced Materials Research, 2011, 328/329/330: 1896 - 1899.

[9] DENG Z H, WANG Z H, JIANG J J. A new algorithm for fast mining frequent itemsets using N-lists [J]. Science China Information Sciences, 2012, 55(9): 2008 - 2030.

[10] LIN K C, LIAO I E, CHANG T P. A frequent itemset mining algorithm based on the principle of inclusion-exclusion and transaction mapping [J]. Information Sciences, 2014, 276: 278 - 289.

[11] VO B, LE T, COENEN F. Mining frequent itemsets using the n-list and subsume concepts [C] // Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway, NJ: IEEE, 2014: 1 - 13.

[12] GOETHALS B, ZAKI M. Frequent itemset mining implementations repository [EB/OL]. [2015-02-20] http://fimi.ua.ac.be/data/.

Background

This work is supported by the National Natural Science Foundation of China (61272029).

XU Yongxiu, born in 1991, M. S. candidate. Her research interests include data mining.

LIU Xumin, born in 1956, Ph. D., professor. Her research interests include computer aided geometric design, graphics and image processing, data mining.

XU Weixiang, born in 1956, Ph. D., professor. His research interests include data mining, analysis and integration for transport systems, cloud computing.

(上接第 996 页)

[7] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C] // Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 1247 - 1250.

[8] BUTLER D. Science searches shift up a gear as Google starts Scholar engine [J]. Nature, 2004, 432(7016): 423 - 423.

[9] FERRUCCI D, BROWN E, CHU-CARROLL J, et al. Building Watson: an overview of the DeepQA project [J]. AI Magazine, 2010, 31(3): 59 - 79.

[10] PAVLIDIS Y, MATHIHALLI M, CHAKRAVARTY I, et al. Anatomy of a gift recommendation engine powered by social media [C] // Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2012: 757 - 764.

[11] DEROSE P, SHEN W, CHEN F, et al. Building structured Web community portals: a top-down, compositional, and incremental approach [C] // VLDB 2007: Proceedings of the 33rd International Conference on Very Large Data Bases. New York: ACM, 2007: 399 - 410.

[12] NIU F, ZHANG C, RÉ C, et al. DeepDive: Web-scale knowledge-base construction using statistical learning and inference [EB/OL]. [2014-10-10]. http://www.cs.stanford.edu/people/chrismre/papers/deepdive_vlds.pdf.

[13] Scrapy 1.0 documentation [EB/OL]. [2015-07-11]. http://doc.scrapy.org/en/latest/index.html.

[14] TARJAN R E. Finding optimum branchings [J]. Networks, 1977, 7(1): 25 - 35.

[15] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web [J]. Scientific American, 2001, 284(5): 28 - 37.

[16] PANKRATIUS W J. Building an organized knowledge base: concept mapping and achievement in secondary school physics [J]. Journal of Research in Science Teaching, 1990, 27(4): 315 - 333.

[17] ZHU J, NIE Z, LIU X, et al. StatSnowball: a statistical approach to extracting entity relationships [C] // Proceedings of the 18th International Conference on World Wide Web. New York: ACM, 2009: 101 - 110.

[18] DESHPANDE O, LAMBA D S, TOURN M, et al. Building, maintaining, and using knowledge bases: a report from the trenches [C] // Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013: 1209 - 1220.

Background

This work is partially supported by the National Basic Research Program (973 Program) of China (2014CB340404), the Shanghai Municipal Science and Technology Research Project (14511108002).

E Shijia, born in 1991, Ph. D. candidate. His research interests include cloud computing, knowledge graph, big-data system.

LIN Peiyu, born in 1993, M. S. candidate. His research interests include knowledge graph, big-data system.

XIANG Yang, born in 1962, Ph. D., professor. His research interests include management information system, cloud computing, semantic computing, big-data mining.