DOI:10. 19431/j. cnki. 1673-0062. 2019. 01. 014

中文影视知识图谱构建研究

谢　星,邹银凤,张硕望,欧阳纯萍*

(南华大学 计算机学院,湖南 衡阳 421001)

摘　要:采用半自动化的方法构建中文影视知识图谱。以国内影视领域数据为研究对象,对豆瓣网、百度百科和时光网的数据进行本体对齐,使异构数据源语义描述一致。在知识融合方面,借鉴并优化了 Similarity Flooding 算法的核心思想,实验结果表明,实体匹配的准确率基本保持在 85% 以上。建立了中文影视知识图谱可视化平台,并提供开放性的数据访问和查询接口。

关键词:知识图谱;中文影视本体;知识融合

中图分类号:TP391　**文献标志码**:A　**文章编号**:1673-0062(2019)01-0073-07

Research on Chinese Film and Television Knowledege Graph Construction

XIE Xing, ZOU Yinfeng, ZHANG Shuowang, OUYANG Chunping*

(School of Computer, University of South China, Hengyang, Hunan 421001, China)

Abstract: A Chinese film and television knowledge map was constructed using a semi-automated method. Taking domestic film and television data as the research object, the ontology alignment of Douban, Baidu Encyclopedia and Time Network data is carried out to make the semantic description of heterogeneous data sources consistent. In terms of knowledge fusion, the core idea of Similarity Flooding algorithm is used for reference and optimized. The experimental results show that the accuracy of entity matching is basically above 85%. A Chinese film and television knowledge map visualization platform was established to provide an open data access and query interface.

key words: knowledege graph; Chinese film and television ontology; knowledge fusion

收稿日期:2018-06-30

基金项目:湖南省哲学社会科学基金项目(16YBA323)

作者简介:谢　星(1991-),男,硕士研究生,主要从事自然语言处理方面的研究. E-mail:764453881@qq.com. * 通信作者:欧阳纯萍(1979-),女,副教授,硕士生导师,主要从事语义网技术、本体建模、社交网络分析和自然语言处理等方面的研究. E-mail:16055205@qq.com

0 引 言

在大数据时代下,万维网上的数据剧增,这些数据包含海量的知识。但在互联网上的数据呈现出松散、无序,这将严重影响人们有效获取知识的便捷性与准确性。因此,如何从爆炸式增长的大数据中获取有用的信息,并对其进行计算和分析,已成为学术界研究的热点。一种以动态、清晰、直观、有效的状态来展示知识和知识内部结构及知识之间联系的数据研究方式——知识图谱(knowledge graph)应运而生。知识图谱能够以图的形式将人们难以理解的知识与知识之间的相互关系以可视化的方式展展示出来,使用户能够快速了解其之间的逻辑关系,轻松地从丰富的数据中抓住关键知识点。

目前中文知识图谱的研究主要存在如下问题。第一,由于汉语复杂度问题,在信息抽取、预处理等方面难度大。第二,专业领域的知识图谱的相关研究较少。第三,从多个知识源获得的知识结构化程度不同和准确率不高,如何从多个知识源获取的异构信息融合起来并存入知识图谱需要进行深入研究。

本文以中文影视为例,设计一种关于中文影视本体知识库的构建流程,对其中的半自动化的构建中文影视本体、对象型属性实体链接和基于相似度传播的实体匹配等关键技术进行研究,并构建了影视知识图谱平台。本文的影视知识图谱集成并融合了豆瓣电影,百度百科和时光网的数据源,其中影视实体 89 072 个,人物实体 45 942 个,135 014 个三元组数据,旨在为用户提供影视智能搜索及问答服务。

1 知识图谱概述

知识图谱其本质是为了表示知识,它是一种具有向图结构的知识库。其中图的节点代表实体或者概念,图的边代表实体与概念之间的各种语义关系。三元组是知识图谱的基本组成单位,基本形式主要包括实体 1、关系、实体 2 和概念、属性,属性值等。实体是知识图谱中的最基本的元素,不同的实体之间存在着不同的关系。概念主要指集合、类别、对象类型、事物种类,每一个实体都可以理解成概念的一个外延;属性主要指对象可能具有的属性、特征、特性以及参数;属性值主要指对象指定的属性值。每一个属性-属性值对可用来描绘实体的内在特性,而关系则是连接两个实体,刻画他们之间的关联。

2012 年 5 月 17 日,Google 在其官方博客中提出知识图谱概念[1],旨在开发出更加智能化的搜索引擎,提高搜索能力,增强用户搜索体验。不同于传统网页关键字搜索,知识图谱能够从语义上理解用户需求,从而直接展示给用户高效精准的搜索结果。比如,搜索“周杰伦年龄”时,搜索引擎直接将结果“39 岁”呈现给用户,并在搜索结果页面右侧显示有关周杰伦个人简介,身高,国籍等情况,极大优化了搜索智能和体验。如今,随着手机等智能设备快速发展,知识图谱技术在搜索引擎、内容分发、个性化推荐、聊天机器人等领域得到了广泛的使用。

关于知识图谱的构建与研究已取得许多重要的研究成果。例如,在商业界,微软、百度、搜狗等互联网搜索企业推出了微软必应 satori、百度知心、搜狗知立方等商业应用。在学术界,清华大学建成了首个大规模中英文跨语言知识图谱 XLore[2]、中国科学院计算技术研究所以开放知识网络(OpenKN)为基础建立了“人立方、事立方、知立方”的原型系统、中国科学院数学与系统科学研究院陆汝铃院士提出知件(Knowware)的概念、上海交通大学构建并发布了中文知识图谱研究平台 zhishi. me、复旦大学 GDM 实验室开发的中文知识图谱项目等[3]。

此外,在特定的领域内,本体构建方面也有了一些进展,如多语言同义词典 BabelNet、中医药知识图谱[4]和多民族语言本体知识库[5]。但在影视领域内,相关研究较少。

2 中文影视知识图谱的构建流程

中文影视知识图谱的构建流程包括 4 个步骤:

1)影视知识抽取:从不同的数据源中自动地抽取半结构化的信息得到候选知识,并对已抽取的数据进行预处理工作。包括命名实体识别,关系抽取和属性抽取等。

2)影视本体构建:采用半自动化的方式构建中文影视本体,并对已抽取的影视知识进行语义对齐。

3)影视知识融合:在异构数据源之间进行清理、消歧、整理等工作,实现不同数据源的知识融合。并针对知识库中对象型属性值,进行命名实体识别和实体链接。

4)影视知识图谱可视化平台:实现数据的可视

化、影视查询功能和智能问答展示,如图 1 所示。

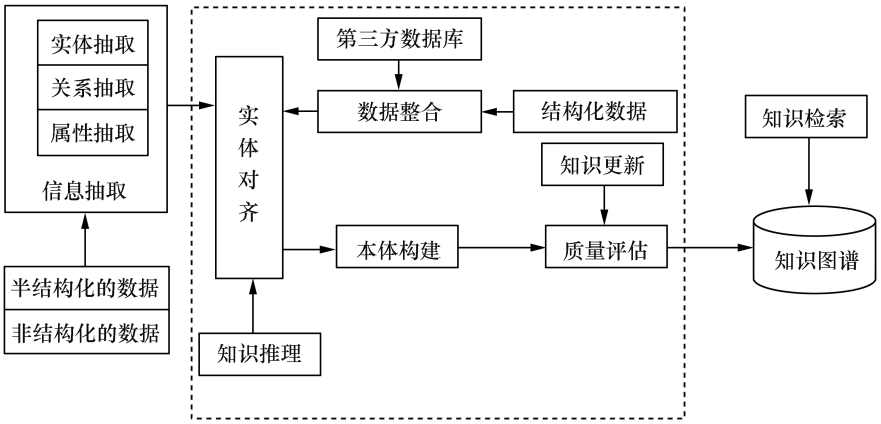

图 1　影视知识图谱构建流程图

Fig. 1　Schme of theconstruction of the film and television knowledege graph

2.1　影视知识图谱数据源

中文影视知识图谱中所包含的影视信息主要是中文影视知识。通过综合考虑影视数据源的规模和质量、数据的获取难易程度以及数据的更新速度,本文选取的数据来源主要如下:

豆瓣电影是国内著名的中文影视评论类网站,它含有丰富的影视信息,并且能够提供结构化的数据、最新的影视评论和影视信息。影视信息更新速度快,链接丰富,语义一致性比较好,并且提供开放性的数据结构;但信息内容相对简单,对影视属性的描述不够全面。

百度百科是百度推出的内容开放、自由的网络百科平台,包含了各种丰富的行业信息,对影视方面的信息数据描述比较丰富和完整,更新速度快。由于人工编辑的原因,其信息结构的一致性比较差,在不同时期呈现的信息质量差异也很大,为知识图谱数据提取抽取工作增加了困难。

时光网具有中文版的 IMDB 之称,拥有权威的影视资源。该网站影视信息描述内容简练、影视内容丰富和质量高,但是影视知识呈现的结构比较松散,这给知识数据的提取带来了一定的挑战。

2.2　影视知识抽取

影视知识抽取是从开放的链接数据中抽取影视知识单元,得到实体、关系以及实体属性等结构化信息。其中的关键问题是如何从异构数据源中自动抽取信息,得到候选知识单元。涉及的关键技术包括:实体抽取、关系抽取和属性抽取。

影视知识抽取一般分为以下几个步骤:

1)网页解析。其步骤主要是对网页模式数据的解析和对网页信息框中信息的初步提取。网页结构解析类似于网站结构分析,需要解析出影视数据中所包含的属性,例如:影视作品包含的导演、时长、制片地区等属性。本文针对不同数据网站制定不同的抽取模板和抽取规则,使用 python 语言制定爬虫程序,实现网页信息的自动爬取,快速并有效地抽取不同数据源网站的影视信息。

2)影视实体抽取。该步骤主要是从网络百科类数据源中抽取出中文影视信息,包括影视本体,属性和属性值。本文选取的数据源网站都包含半结构化的影视数据信息展示模块,这些模块一般由大标题和小标题以及相应的标题值构成。经分析发现大标题一般对应着实体名称,而小标题对应着实体的属性或者是对象类型属性的实体名,标题值则是相应小标题标明的属性值,它可能是一个具体的对象,也可能是一个属性值或者属性值列表。在信息抽取过程中只需要准确识别出这些信息块,便可以有效地从中提取出相应的影视信息。另外,在非结构化的影视简介信息中,使用启发式和定制规则的方法从简介信息中识别和抽取电影作品实体,如电影名称的出现一般会带有书名号“《》”或者引号“”,识别出这些特殊符号之后可以界定电影作品实体的边界,再经算法的进一步处理便可以实现实体的准确的抽取。

3)属性对齐。该步骤主要是对异构数据源中的属性词汇进行统一。由于本文选择的数据源来自三个不同的网站,在部分属性上不同的网站对同一实体的同一属性名使用了不同的描述,而

对同一个数据源网站对于固定属性都使用了固定的描述。因此,通过人为构建同义词映射表可以确保不同数据源属性对齐的正确性。例如对于电影作品别名的描述,百度百科使用了“外文名”和“其他译名”,豆瓣电影使用了“又名”,而时光网使用了“更多片名”,因此可以为这些属性名建立统一的属性同义词映射表(别名、外文名、其他译名、又名、更多片名),使他们的属性值统一对应“别名”这个属性。

4)属性值处理。该步骤的主要任务是对属性值中的长文本信息进行初始化分割,以识别文本中的词汇语义边界(如标点符号、左斜杠、空格、超链接、不同语言单词交界等)。

5)实体识别。公开的影视数据集主要依赖于群体编辑,由于群体的文化水平参差不齐,缺乏统一的数据编辑规范,在数据集上可能会导致上下位关系紊乱,概念语义粒度不一致和产生歧义等问题。实体识别主要是通过基于规则的方法初步确定实体类别。

2.3 影视本体构建

本体构建是对概念本身以及概念与概念之间关系进行形式化描述。通常包含本体需求分析、考察可复用本体,建立领域核心概念,建立概念分层次、定义类和创建属性以及本体评价和进化 6 个阶段[6]。针对不同的领域和不同的实际需求,本体构建方法也有不同。

目前国内外已有许多成熟的影视本体,可以复用这些本体建立概念体系结构,比较权威的影视本体(movie ontology,MO)和 Freebase Film,它们在概念层次结构上都是以影视作品和影视人为核心的扁平化概念层次结构。其中 MO 采用以影视作品为中心的平行概念结构,主要定义了作品、人物、体裁和地区等概念,概念粒度比较细,涵盖面比较小,语义粒度大。Freebase Film 的概念描述体系比较复杂,涉及的概念非常多,具有更精细的粒度。

事实上,难以获取到类似 MO 或者 Freebase Film 那么详尽的影视信息。因此在复用以上两种本体概念体系结构时,要以符合本地数据源的最小粒度来选择概念粒度。以“公司”为例根据 Freebase Film 的分类,它可以进一步分为制片公司和发行公司,但是实际上采用的数据源中仅有百度百科有小部分“公司”相关数据。因此放弃这两个子类别。

在电影数据中,某些属性的描述是一个列表。有时列表中节点的顺序被认为是重要的,比如演员表,通常包含多个演员。一般的演员位置越靠前演员对电影作品的重要性越大,因此本文引入有序节点,区别是添加了一个额外的属性来标记节点的顺序。

目前,本文的中文影视本体,总共创建了 12 个概念和 64 个属性。由于篇幅原因,在此省略影视本体的具体构建详情。

2.4 影视知识融合

影视知识融合是为了将从豆瓣电影,时光网、百度百科三个异构的数据源中抽取的数据进行整合和清理,去除其中冗余、错误的部分,形成结构化的影视知识库。其核心工作包含实体对齐和实体链接两个部分。在实体抽取过程中,我们选择抽取知识的模块主要集中在影视数据源网站的半结构化影视百科知识部分。在具体的处理过程中,实体对齐部分的工作量大且比较复杂,而实体链接部分的问题比较少,并且处理过程比较简单,因此不予以进一步的叙述。下面具体介绍实体对齐过程中的一些具体的处理策略和方法。

实体对齐普遍使用聚类的方法,利用合适的相似度传播算法,而相似度传播算法较为典型的是 SF(similarity flooding)算法。SF 算法的核心思想是将两个元素相似性的部分传播给其在图中各自的邻居,这种传播方式类似于 IP(internet protocol)广播[7]。

本文借鉴了 SF 算法的思想,在实际构图过程中,预先对实体进行剪枝,具体步骤如下:

1)对影视实体进行整理,去除其中类别不同对的实体对;

2)对不同出身年份的影视人实体对和不同上映年份的影视作品进行排除;

3)计算实体对的相似度时,先剔除实体主题词低于一定阈值的实体对,再剔除实体相似度低于另一阈值的实体对。剪枝之后,相似度传播图中的结点得到了明显的下降,大大减少了算法的计算量。

实体对齐技术中不仅需要选择合适的实体匹配算法之外,还需要选择适当地实体特征模型用来计算实体相似度并使计算的实体相似度具有足够大的区分度。实体的标题或者名称在实体的区分度中发挥着重要作用,而实体的属性体现了实体的本质特征。因此在计算实体相似度时主要考虑两个问题:实体主题词相似度和属性相似度[8]。

1)实体主题相似度

实体的主题词是指是的标题或者别名,它是区分实体的标识之一,是表达实体的核心词汇。因此在计算主题词相似度时,我们归并实体的别名、标题和同义词汇构成主题词集,以词集之间的相似度代替标题词相似度。主题词的相似度定义如下:

$$sub_{\text{sim}}(a,b)=\max\left(\frac{\text{lcslen}(x,y)}{\max(\text{len}(x),\text{len}(y))}\right) \tag{1}$$

其中,$x\in N_a, y\in N_b$,其中 N_a 和 N_b 分别为实体 a 和 b 的主题词集,$\text{lcslen}(x,y)$ 为最长公共子序列长度。

2)属性相似度

实体的属性通常包含不同类别,根据不同类别的属性值,分以下几种情况计算属性相似度。

①二值型:

$$P_{\text{sim}}(x,y)=\begin{cases}1, x=y\\ 0, \text{否则}\end{cases} \tag{2}$$

②数值型:

$$P_{\text{sim}}(x,y)=\frac{|x-y|}{\max(x,y)} \tag{3}$$

③字符串型:

$$P_{\text{sim}}(x,y)=\begin{cases}\dfrac{\text{lcslen}(x,y)}{\max(\text{len}(x),\text{len}(y))},\\ x \text{ 或者 } y \text{ 的长度} > 4\\ \text{sim}(x,y), \text{否则}\end{cases} \tag{4}$$

$\text{sim}(x,y)$是用来计算词语相似度的。当属性值为非实体名的字符串时,且长度大于 4 个中文字,默认将 x,y 当做词语处理。词语相似度计算方法,借鉴与文献[8](词语相似度算法的文献)。

④列表型:如编剧表,演员表等属性,通常列表型中包含多个实体对,其计算公式如下:

$$P_{\text{sim}}(x,y)=(x\cap y)/(x\cup y) \tag{5}$$

3)实体相似度

综上所述,定义实体相似度为式(6),其中,N 为属性数量,$P_{\text{sim}_i}(a,b)$ 为相应的属性相似度,w_i 为属性的权重:

$$e_{\text{sim}}(a,b)=\frac{1}{N+1}\left(n_{\text{sim}}(a,b)+\sum_{i=1}^{N}P_{\text{sim}_i}(a,b)w_i\right) \tag{6}$$

由于知识结构的简单一致、信息量丰富,在影视领域中判断两个实体是否相似并实现实体对齐准确率都比较高,且基本保持在 85% 以上。系统每次的数据更新中,对于本文中的三个异构数据源,都需要对其进行两两实体对齐:1)百度百科和豆瓣电影之间的实体对齐;2)时光网和豆瓣电影之间的实体对齐。

3　中文影视知识图谱可视化平台的设计和实现

中文影视知识图谱可视化平台是以动态,清晰,直观,有效的方式展示影视知识和知识内部结构及知识之间的联系而设计的可视化应用平台。它实现了在中文影视领域现有知识的基础之上将知识进行了抽象和精简,并以简单、直观、有效地方式为人们提供影视知识和推荐的服务。

3.1　系统的业务流程设计

影视知识图谱可视化平台是基于 Servlet+JSP+EChart+Mybatis+Mysql+Tomcat 开发的 B/S 系统。主要由电影关系图谱、影片知识图谱、影视知识问答和影视知识库管理四大功能模块组成。其具体的模块设计如图 2 所示。

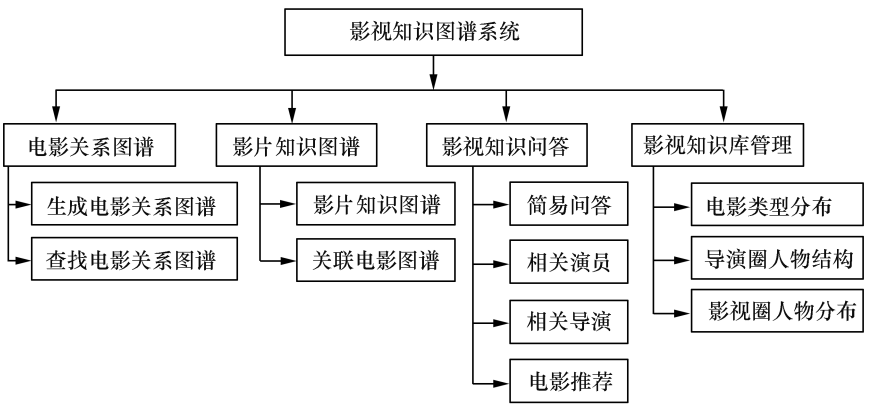

图 2　影视知识图谱可视化平台功能模块框架

Fig. 2　Frame of visulized functional moduels of the film and television knowledege graph

3.2 电影关系图谱

电影关系图谱是以力导向图为知识展现载体的影视知识展现方式之一。其中结点代表电影实体,在图谱中用圆形表示;电影与电影之间的关系用连接两个电影的带标注的弧线表示,标注内容的是两个电影相似的属性之一。弧线的长度表示两个电影之间的相似程度。弧线越短表示两个电影的关系越近,说明两个电影越相关。电影关系图谱功能模块默认显示的是以当前最新电影为核心的电影关系图谱,点击图谱中非核心的结点可以向下扩展一层,整个电影关系图谱可向外扩展三层。电影关系图谱界面如图 3 所示.

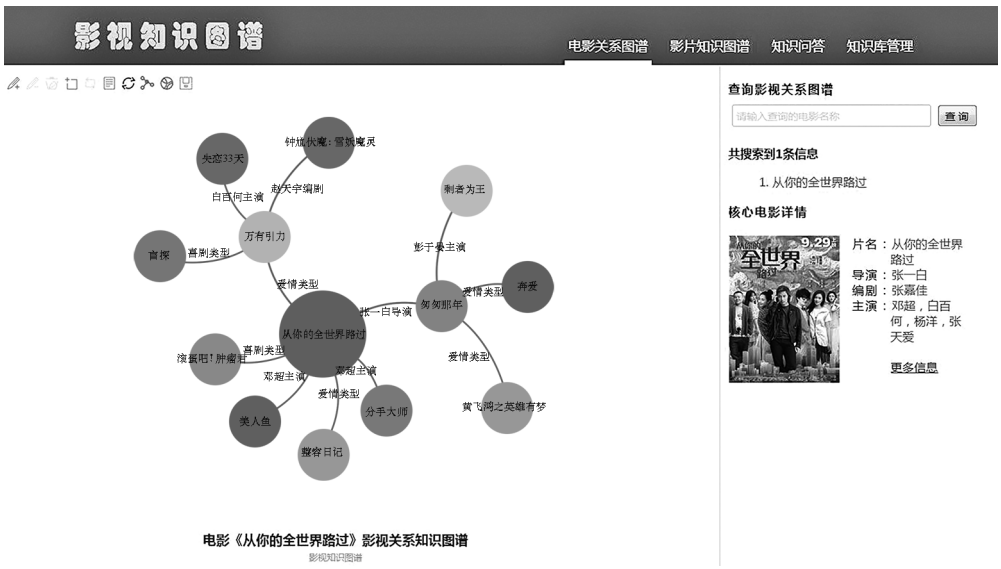

图 3 电影关系图谱界面

Fig. 3 Interface of the film relation graph

查找电影关系图谱需要在输入框中输入电影名称的全名或者部分,如果知识库中包含查找的电影,系统将返回所有的查找结果列表,如果没有则返回默认的电影关系图谱界面。默认的查找结果显示的是排序最靠前的电影关系图谱和知识卡片,点击任意查找结果可以实现图谱间的切换。

生成电影关系图谱模块是将不同的电影,通过相似的属性值为连接关系而形成一个完整的影视知识图谱。其中以电影实体为结点,电影间相似的属性和属性值对为连接线将不同的电影实体连接起来,并以力导向图展现不同电影间的结构图谱。生成电影关系图谱功能模块,每次只能生成指定结点下的一个直接子结点集。若需要生成一个大型的多层电影关系图谱,则需要遍历每一层的父结点,逐层生成子结点直至结束。影片知识图谱展示的是单部电影的具体信息。除了使用传统的知识卡片的方式展示电影的具体信息外,还引入了电影信息图谱和指定相关方式的电影关系图谱。

电影信息图谱主要展示的内容是电影的属性和属性值对,是单部影片知识的一种图形化展示方式。如图 4 电影《摆渡人》的信息图谱。通过比较电影信息图谱中各个结点与中心结点距离的远近,可以直观地反映出电影实体中各部分属性信息的重要程度。指定相关方式的电影关系图谱中用户可以根据需要指定电影间的相关方式(即连接关系),它展示的是与当前电影具有相同的指定属性值的电影及他们的相关程度。其中结点越大,距离核心电影越近的影片,与核心电影相同的属性值越多或者拥有的相同的属性越重要,他们的相关程度越大。

3.3 影视知识问答

影视知识问答模块是一个基于影视知识图谱的简易问答系统,它能在一定程度上理解用户的查询意图,直接返回问题的答案,同时能够根据具体的查询信息给出相应的推荐。不同于传统的搜索引擎或者问答系统,基于影视知识图谱的简易问答系统依赖于影视知识图谱中存储的知识,虽只能理解影视领域的相关问题,却能针对问题做出精确回答。例如,搜索问题“从你的全世界路过的导演是谁?”系统将返回“《从你的全世界路过》的导演是:张一白”。

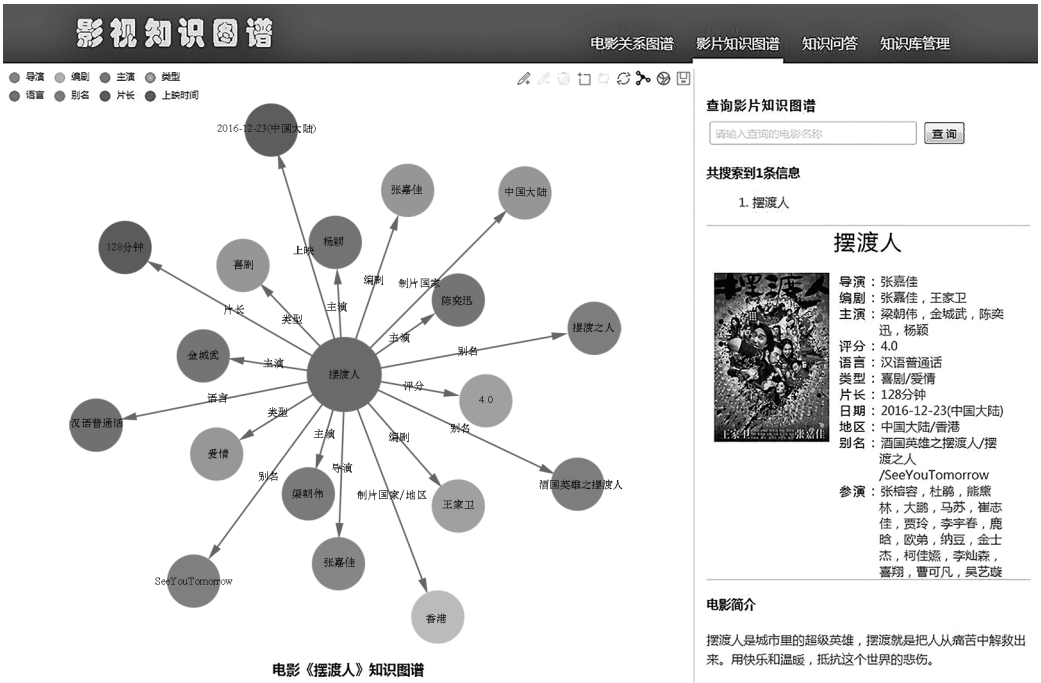

图 4　电影《摆渡人》的信息图谱

Fig. 4　Graph of the information of the film The Ferryman

影视知识问答模块的知识展示,除了直接给出问题答案之外,系统会根据问题的答案和关键字给出相近的或者相关的电影推荐和相关的演员等信息。当用户没有提问或者用户提出的问题系统无法解析时,系统默认推荐的是当前系统中最新且评分较高的电影以及参演作品比较多的明星知识卡片。

4　结　论

本文融合优质的中文影视数据资源构建了丰富的中文影视知识库,为影视领域下步的知识化和智能化建设提供了可借鉴思路。同时,设计了影视知识的可视化的展示方式,并实现了基于影视知识图谱的部分智能化应用。如影视知识问答,指定相关方式的电影查询及展示,可以为用户提供更好的中文影视知识服务,尤其是在影视的推荐服务中,用户可以通过指定相关方式的电影关系图谱,针对个人的实际需求自定义个性化推荐服务。在中文影视领域具有一定的应用和推广价值。

参考文献:

[1] AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google,2012.

[2] 王巍巍,王志刚,潘亮铭,等. 双语影视知识图谱的构建研究[J]. 北京大学学报(自然科学版),2016,52(1):25-34.

[3] 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.

[4] 阮彤,孙程琳,王昊奋,等. 中医药知识图谱构建与应用[J]. 医学信息学杂志,2016,37(4):8-13.

[5] 赵小兵,邱莉榕,赵铁军. 多民族语言本体知识库构建技术[J]. 中文信息学报,2011,25(4):71-74.

[6] 张文秀,朱庆华. 领域本体的构建方法研究. [J]图书与情报,2011(1):16-20.

[7] MELNIK S, GARCIAMOLINA H, RAHM E. Similarity flooding: a versatile graph matching algorithm and its application to schema matching[C]//International Conference on Data Engineering,2002. Proceedings. San Jose, US: IEEE,2002:117-128.

[8] ZOU Y, OUYANG C, LIU Y, et al. A similarity algorithm based on the generality and individuality of word[M]. Natural Language Understanding and Intelligent Applications: Springer International Publishing,2016.

(责任编辑:周泉)