北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2016.022

双语影视知识图谱的构建研究

王巍巍　王志刚†　潘亮铭　刘阳　张江涛

清华大学计算机科学与技术系知识工程实验室，北京　100084; †通信作者, E-mail: wangzigo@gmail.com

摘要　提出一种双语影视知识图谱(BMKG)的构建流程。通过半自动化的方法构建了双语影视本体(BMO)，将各个影视数据源对齐到 BMO，以保持异构数据源的语义描述一致性。在知识链接方面，在充分挖掘和利用领域特征的基础上，采用基于 Word2Vec 和 TFIDF 两种向量模型的实体相似度计算方法，使相似度特征增加一倍，大大提升了模型的链接效果。在实体匹配方面，提出基于相似度传播算法的实体匹配算法，并利用影视数据源之间的内在联系，克服了跨语言实体之间计算相似度的语言障碍。实验结果表明，当阈值取到 0.75 以上时，实体匹配的准确率都能达到 90%左右。此外，还建立影视知识图谱共享平台，并提供开放性的数据访问和查询接口。

关键词　影视本体；双语；知识图谱

中图分类号　TP391

Research on the Construction of Bilingual Movie Knowledge Graph

WANG Weiwei, WANG Zhigang†, PAN Liangming, LIU Yang, ZHANG Jiangtao

Knowledge Engineering Group, Department of Computer Science and Technology, Tsinghua University, Beijing 100084;
† Corresponding author, E-mail: wangzigo@gmail.com

Abstract　This paper proposes a method to construct Bilingual Movie Knowledge Graph (BMKG). The authors first builds Bilingual Movie Ontology (BMO) through a semi-automatic way, and aligns each data source with it in order to ensure semantic consistency of heterogeneous data sources. For entity linking, the proposed method makes best use of the field characteristics and calculate entity similarity based on both Word2Vec and TFIDF models, which greatly improve entity linking. For entity matching, a similarity flooding based algorithm is proposed, which utilizes the intrinsic links between the movie data sources, addressing the problem of similarity computation between cross-lingual entities. The experiment results show that the entity matching precision is over 90% when the threshold is above 0.75. In addition, a movie knowledge graph sharing platform is also built to provide open data access and query interface.

Key words　movie ontology, bilingual, knowledge graph

随着互联网和智能设备的普及，影视已经成为人们娱乐生活中不可或缺的一部分，而互联网是人们最重要的影视信息来源之一。人们可以很方便地在优酷土豆、爱奇艺等视频网站上观影，也可以在豆瓣电影、Imdb、百度百科等网站上获取影视以及评论信息。然而，有些用户对影视信息有更深层次的需求，比如制片公司、广告商等往往期望了解影视作品在人员、受众、时间、地域、收视率等不同维度上的统计信息。目前，大部分影视挖掘算法和相关系统的分析效果通常依赖于其背景知识库的质量，因此工业界和研究领域均对高质量影视知识库有着非常迫切的需求。

国家重点基础研究发展计划(2014CB340504)、国家自然科学基金委员会与法国国家科研署双边合作协议(61261130588)、 清华大学自主科研项目(20131089256)、国家科技支撑计划(2014BAK04B00)和 THU-NUS 下一代搜索联合研究中心项目资助

收稿日期: 2015–06–06; 修回日期: 2015–08–17; 网络出版时间: 2015-09-29 15:53:47

国际上，影视本体构建工作进展很快，开放数据云(linked open data, LOD)上已经出现一批如LinkedMdb、Freebase 等著名知识库，但大多以英文知识为主。目前，国内虽然已经出现了比较优秀的中文影视网站，但在影视本体知识库的构建方面相对落后，相对于英文影视知识而言，能够公开获取的中文影视数据源中影视知识的结构化较差，且描述信息较少，缺乏一个统一的语义描述标准。所以，融合优质的中英文影视数据源，构建统一接口、统一语义的双语影视本体知识库，将会为国内的影视信息的挖掘和利用提供重要的基础支撑，同时，对扩大中文影视知识在国际上的影响力具有重要的意义。

总体来说，双语影视知识库的构建工作会面临如下几个挑战。

1) 双语影视本体构建。当前没有成熟可用多语言影视本体，因此，需要根据实际需求，考虑中英文知识平衡性，重新进行构建。

2) 语义信息抽取。从不同的数据源中抽取结构化影视知识，需要进行数据过滤、去噪、清洗、结构化、语义对齐等一系列复杂的预处理过程。

3) 对象型属性实体链接。需要解决关键问题：一是命名实体识别，即如何从属性短文本中，特别是中文文本中进行实体边界的识别；二是领域相似度定义问题，即如何利用影视领域知识，构建具有足够区分度的实体相似度计算公式。

4) 大规模实体匹配以及跨语言实体匹配。需要解决大规模实体匹配的计算可行性问题以及跨语言匹配时，克服实体相似度计算中的语言障碍。

基于上述分析，我们提出一种双语影视本体知识库的构建流程，并对关键技术进行研究，其中包括半自动化的影视本体构建、对象型属性实体链接以及基于相似度传播的实体匹配，为了实现知识共享和可视化，我们还构建了双语影视知识图谱(Bilingual Movie Knowledge Graph, BMKG)应用平台，并开放数据访问和查询接口。

BMKG 集成并融合了豆瓣电影、百度百科、LinkedMdb、DBpedia 等多个中英文影视数据源，包含七十多万个影视实体，一千多万条三元组数据，并建立了 60 万条到多个开放数据源的外部链接。

表 1　综合统计

Table 1　Overall statistics

项目	数量
三元组数量	13616766
LOD 链接数量	614870
影视网站链接数	710841
sameAs 链接数	574971
概念数量	23
属性数量	91
实体数量	728553

表 1 给出知识库的综合统计数据。

1　相关工作

自 20 个世纪 90 年代起，语义网相关技术开始蓬勃发展，本体技术成为研究热点，以 DBpedia、WordNet[1]等为代表的一批优秀的本体知识库开始涌现，标志着语义网技术走向成熟，进入到实际应用阶段。然而，由于本体知识库的构建工作是一项非常复杂、费时费力的系统性工程，其进展相对缓慢，已经成为本体技术发展的瓶颈之一。研究和构建各种本体知识库已经成为当务之急。

国际上，以 DBpedia 为核心的 LOD 开放数据云中本体知识库大多以英文知识为主，尤其是影视领域方面，英文知识库的研究工作一直处于领先地位。Hassanzadeh 等[2]在 2009 年发布影视本体知识库 LinkedMdb，该知识库是以影视知识为中心的链接型知识本体。2010 年，苏黎世大学的 Bouza 等在 LOD 中公布了构建的影视本体 MO①，为大多数的影视数据生产者提供了一个一致的语义规范。大规模知识图谱 Freebase 也含有丰富的影视知识，并建立了一套非常优秀的影视概念体系。

我国的本体构建技术研究还处于起步阶段。在领域本体构建方面，虽然已经有了一些成果，如中文语言本体知识库 HowNet②、医疗领域本体知识库[3]和多民族语言本体知识库[4]，但总体来说，所涉及的领域较少，在规模和质量上都远不能满足现实应用的需求。尤其在有广泛应用前景的影视领域方面，国内还没出现高质量的知识库。

① http://www.movieontology.org/

② http://www.keenage.com/

本体知识库大多都采用半自动化方法构建而成，构建的复杂程度与所用数据源的质量和规模有关。例如，DBpedia 是从维基百科网页数据中抽取多语言的数据[5]，主要侧重于知识的结构化，在进行大规模半结构化数据处理过程中，需要引入大量的人工操作，构建过程十分繁琐复杂。LinkedMdb 的知识规模小，所操作对象数据源基本都是优质的 RDF 数据源，并且主要侧重于建立异构数据源之间的知识链接，构建过程相对简单。

BMKG 涉及到两种语言的数据源，中文选用半结构化网页数据源，英文选用优质的 RDF 数据源。因此，可以借鉴上述两种知识库的构建方法，分别构建中英文影视知识库。

在构建知识库的过程中，为了实现知识融合，需要对各个异构的知识库进行大规模的实体匹配。随着 OAEI 这项实体匹配方面的国际性竞赛的不断举行，越来越多的实体匹配算法开始涌现。PARIS[6]、SIGMA[7]和 RiMOM[8]是比较有代表性的算法，都采用基于图的相似度传播(Similarity Flooding[9])思想，能够充分利用数据的结构化进行实体匹配。在跨语言实体匹配方面，基于通用算法，克服了实体相似度计算中的语言障碍。文献[10]通过中文维基页面，建立英文维基与百度百科之间联系，并提出基于因子图的知识链接方法，取得非常不错的效果。

2　双语影视知识图谱的构建流程

BMKG 构建的基本流程包括 5 个步骤，如图 1 所示。

1) 本体构建：通过复用现有的知识本体，半自动化构建双语影视本体；

2) 语义信息抽取：从数据源中抽取结构化影视知识，并在语义上对齐到双语影视本体；

3) 对象型属性实体链接：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工作；

4) 实体匹配：在异构数据源之间进行实体匹配，实现不同数据源的知识融合；

5) 双语知识图谱共享平台：双语影视知识库的可视化应用平台，实现数据可视化和查询功能。

2.1　数据源

BMKG 选择数据源的标准为：影视数据源的规模和质量；数据的获取难度；数据源是否保持更新。所以，目前我们主要从如下数据源抽取影视知识：

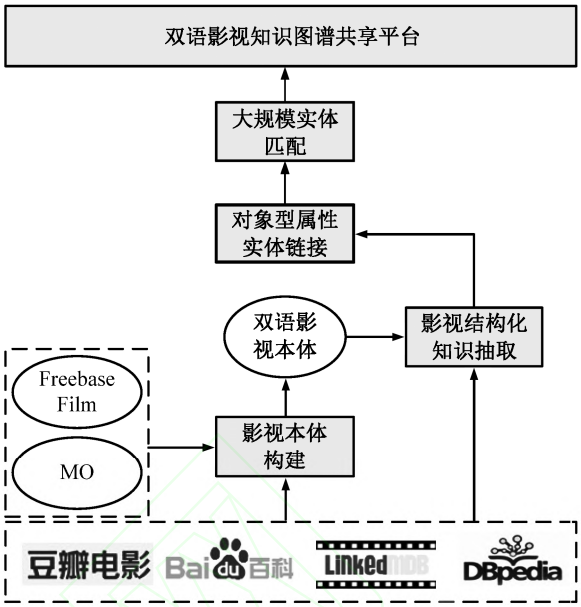

图 1　双语影视知识图谱构建流程图
Fig. 1　Pipeline of the BMKG

1) 豆瓣电影是当前最著名的中文影视评论网站之一，提供最新的影视介绍以及评论信息，并且提供开放性的数据访问接口。其数据具有结构化、链接丰富、语义一致性好的优点。目前我们获取了 127406 个影视作品，70534 个影视人，但信息内容相对简单，同时也缺乏丰富的影视属性描述。

2) 百度百科是当前最大的中文百科全书。近几年来，随着百度公司的不断努力，其数据，尤其是在影视信息方面，无论是在规模上还是质量上都有显著的改进，影视信息较为丰富，可以作为豆瓣影视数据有效补充。目前我们抽取了 69861 个影视实例，42012 个影视人。然而，由于是基于人工编辑的半结构化文本，且不同时期编辑的网页数据质量差异很大，所以，对语义信息抽取工作带来一定的挑战。

3) LinkedMdb 是一个开放性、高质量的英文影视知识库，它从 Imdb，Freebase，DBpedia 等数据源抽取知识，包含 85620 部影视作品、107768 位影视人、6148121 个三元组、162199 个内部链接以及 541810 个外部网页链接。遗憾的是，该知识库自 2010 年 2 月后就不再更新。

4) DBpedia (Wikipedia) Movies 是结构化的维基百科 RDF 数据，包括 10 多万部影视作品、10 多万影视人以及大量链接信息，其数据质量类似于百

度百科，是 LinkedMdb 的有效补充。

5) Freebase 是共享的全球性知识图谱，其中 Film/TV 等影视类数据是其重要的组成部分。截止到 2015 年 5 月，有超过 40 万的影视作品以及数百万影视相关实体信息。相比较其他知识库而言，Freebase 提供了更为详细的影视数据，其概念和属性也颇为丰富。但是，自 2014 年之后，Freebase 不再提供完整的 RDF 数据集下载。

2.2 双语影视本体构建

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，一般包含本体需求分析、考察可复用本体、建立领域核心概念、建立概念分类层次、定义类和创建属性以及本体评价和进化 6 个步骤[11]。针对不同的领域和不同实际需求，本体构建方法也有所不同，我们研究了当前多语言影视领域本体实际情况，给出双语影视本体的构建思路。

2.2.1 复用已有本体，建立概念结构体系

当前已有许多成熟的影视本体，如国际上比较权威的 MO 和 Freebase Film。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，其中以体裁和地区最为详细，具有 3~4 层的分类层次，但概念的涵盖面较小，语义粒度较大。Freebase Film 的概念描述体系较为复杂，涵盖影视信息的各个方面，涉及概念非常多，语义粒度也较细，但实际中，我们很难获取到如此详尽的影视信息。

在概念层次结构上，上述本体都是以影视作品和影视人为核心的扁平化概念层次结构。我们复用这种概念体系结构，但在概念粒度的选取上，采用契合本地数据源的最小粒度方案。以“公司”为例，根据 Freebase Film 的分类可以进一步分为制片公司、发行公司两个类，但实际上所采用的数据源中仅百度百科有部分公司相关数据，且信息量较少，无法支持更细粒度的概念分类，因而，放弃使用这两个子分类。当然，如果数据能够有效支持上述两个分类，我们会尽量在更细的概念粒度上进行描述。

在核心词汇的选取上，我们尽量使用标准影视词汇集：英文词汇方面，主要从上述本体中进行抽取；中文词汇方面，我们根据考查词汇在当前大型影视网站的流行度，选取流行度最高的词汇集。最后手工对齐中英文的影视词汇，构建双语核心影视词汇集。

2.2.2 建立多元影视属性描述结构

在影视数据中，一些属性有多元信息的描述需求，比如，演员表属性要分别描述演员名、演员 id、角色等多种信息，通常的三元组无法同时进行描述，因而，本文引入中间节点(匿名节点)来承接这些多元信息。

有些属性描述是一个列表，但有时节点在列表中的顺序被认为是重要的，如演员表通常有多个演员，但主演应该排在更前面的位置。因此本文引入有序节点，它是匿名节点的一种，区别是添加了一个额外的属性来标记节点的顺序。表 2 是用有序列表来描述演员表属性的示例。

根据观察，绝大部分影视数据(例如演员属性)的内容文本的编辑顺序基本表现了实体的重要性，因此本文节点的顺序主要依据字符串或表格中实体出现的先后顺序进行确定。

现阶段的双语影视本体，共建立了 23 个概念和 91 个属性，由于篇幅原因，所构建的双语影视本体将在影视共享网站平台上给出。

2.3 影视结构化知识抽取

影视结构化知识抽取是从各异构数据源抽取影视知识，并对各种格式的数据进行分析，统一语义、统一结构的过程，大体包括如下 5 个模块。

1) 网页解析。该模块主要是网页模式的分析以及网页中表格信息的抽取。其中，采用基于树编辑距离的自适应学习方法[12]，可以有效提升表格抽取的效率，有效抽取大部分模式的表格数据。

2) 影视信息抽取。主要任务是从百科类数据

表 2 匿名节点实例

Table 2 Blanknode example

主语	谓语	宾语
bmkg_instance:100	rdf:label	“中国合伙人”@zh
	bmkg_property:actor	bmkg_blanknode:10
	bmkg_property:actor	bmkg_blanknode:11
	bmkg_property:actor	bmkg_blanknode:12
bmkg_blanknode:10	bmkg_property:actor_name	“黄晓明”@zh
	bmkg_property:actor_id	bmkg:instance:200
	bmkg_property:character	“成东青”@zh
	bmkg_property:serial_num	**1**

注：bmkg_blanknode:10 是匿名节点 id，描述的是影片中国合伙人中的主演黄晓明的信息，加粗的一行表明其排在演员表的第 1 位。

源中筛选出影视信息。影视词汇的词频和共现率都很高，基于关键字过滤的方法能有效地抽取大部分影视数据。此外，利用文献[13]提出的基于智能结构化感知的实体抽取技术，能够充分感知数据中结构化知识，进一步地迭代抽取所需类别实体。

3) 属性对齐。该过程的主要任务是统一异构数源中属性描述词汇。双语影视本体的概念和属性很少，对属性进行频度统计，发现属性描述信息是一个长尾分布，常用的属性名其实非常少，因而，可以花费非常小的代价，人为构建同义词映射表实现属性对齐，确保不同数据源语义对齐方面的正确性。

4) 属性值处理。对属性值中的长文本进行初始分割，主要任务是识别文本中的词汇语义边界(如标点符号、空格、超链接、不同语言单词的交界等)，将文本分割为更小粒度的文本块，以减少后续命名实体识别的难度。

5) 实体类别识别。公开的影视数据集依赖于群体编辑，存在多种不同的概念层次结构，且概念语义粒度不一致，上下位关系紊乱，甚至会产生歧义。该步骤的目的是通过基于文本规则的方法，初步确定实体类别，例如，百科页面中“刘德华”可以通过职业属性来判断他属于演员、制片人等类别。在后续大规模实体匹配基础上，通过知识互补以及相应的推理机制，进一步对实体的类别信息进行完善。

经过上述 5 个步骤后，源数据转化为结构化 JSON 格式数据。

2.4　对象型属性实体链接

对象型属性，即取值范围是指定类型实体的属性。如演员表属性，其值是演员实体列表。命名实体通常指人名、机构名、地名以及其他所有以名称为标识的实体。对象型属性实体链接工作的任务是将对象型属性值中未标注的命名实体识别出来，并建立其到相应实体的知识链接。

2.4.1　属性值命名实体识别

命名实体识别过程通常包括两部分：确定实体类别实和体边界识别。对于前者，根据属性取值范围已经基本确定了实体类别。对于后者，英文的命名实体之间几乎都有明显的标识，比较容易识别，因此本研究主要针对中文命名实体边界的识别。

结构化好的数据源(如豆瓣)，其对象型属性值中命名实体基本已经标注出来。半结构化数据源(如百度百科)，许多命名实体并没有进行标注，属性值大多以文本形式存在，主要有 3 种情况：1)含有超链接信息的文本，即文本中将实体信息以超链接形式出现；2)有明显语义标记的文本，命名实体之间用一致的标点符号分隔，且没有歧义；3)没有明显语义边界的长文本，命名实体之间没有分隔符，或使用如空格、“-”等有歧义的分隔符。对于前两种情况，在语义信息抽取的属性值处理过程已经处理过，因此我们主要对第 3 种情形进行处理。

我们选用 ansj[①]作为中文分词工具，ansj 是基于条件随机场和 Google 语义模型的开源工具，在分词正确率以及分词速率方面有非常不错的表现。中文分词工具通常也带有命名实体识别功能，但一般仅识别人名、地名、机构名等通用类别的实体，且对合成词的识别效果不好。我们通过词典来改进命名实体识别的效果。一方面，结合我们收集和整理的大规模通用细胞词库，能够大大提高分词的正确率，并增大分词的粒度。另一方面，在分词序列的基础上，利用影视领域词表进行最大词块匹配，能够充分识别已登录的合成词，提升命名实体识别的效果。

2.4.2　实体链接

实体链接的核心是计算命名实体和候选实体的相似度，选择相似度最大的候选实体作为链接的目标实体[14]，选择合适的文本语义特征来计算实体相似度是实体链接的关键性问题。文献[15]是在维基百科数据集上的知识链接补全工作，它采用文档中丰富的出入链信息作为基本元素来计算文档相似度，在此基础上，通过加权的 7 个文本语义特征来计算实体的语义相似度。

本文借鉴上述加权思想，并根据实际情况做了一些改进：一是百度百科的链接质量不高，基于出入链的文档相似度计算方法不再适合，需要重新定义；二是考虑到影视领域特征，重新提炼文本特征计算实体相似度。

定义 1　文档相似度。我们采用基于向量空间的文档相似度计算方法，将文档表示为两种向量形式：一种是 TF-IDF 向量，标记为 V_t；一种是 Word2Vec 向量，标记为 V_w。V_w 是通过整个百度百

① https://github.com/ansjsun/ansj_seg/

科语料库学习出 Word2Vec[16]词向量，然后计算文档中词向量的平均值而得到。给定百度百科中两个实体文档，根据不同的文档向量表示方式，文档相似度定义如下：

$$r(a,b)=\begin{cases}V_{\mathrm{t}}(a)\bullet V_{\mathrm{t}}(b),\\ V_{\mathrm{w}}(a)\bullet V_{\mathrm{w}}(b),\end{cases} \tag{1}$$

其中，$V_{\mathrm{t}}(\mathrm{a})$，$V_{\mathrm{t}}(\mathrm{b})$，$V_{\mathrm{w}}(\mathrm{a})$，$V_{\mathrm{w}}(\mathrm{b})$ 分别为实体 a 和 b TF-IDF 和 Word2Vec 向量。

定义 2 语义相似度。假设 B 是一个实体集合，实体与 B 之间的语义相似度定义为

$$\mathrm{SR}(a,B)=\frac{1}{|B|}\sum_{b\in B}r(a,b)\text{。} \tag{2}$$

定义 3 实体相似度。文档 C 对应的实体记为 a，词汇全集记为 C_{text}，m 是属性 p 中某一命名实体，p 的属性名领域词集记为 $C_{\mathrm{attr\_name}}(m)$，属性值领域词集为 $C_{\mathrm{attr\_value}}(m)$，影视领域词汇全集为 C_{domain}，相应的向量分别记为 $V_{\mathrm{attr\_name}}(m)$，$V_{\mathrm{attr\_value}}(m)$，$V_{\mathrm{domain}}$，正文和属性框的出链实体集合分别为 O_{article}，页面入链集合为 I_{all}，b 是 m 的候选实体。如表 3 所示，定义了 7 个特征相似度，有两种文档向量形式，计算可得到 14 个特征相似度。于是，实体相似度定义如下：

$$\mathrm{e\_sim}(m,b)=\sum_{i}^{14}w_i * f_i\text{。} \tag{3}$$

其中，特征权重值可以通过 logistic 线性回归模型进行学习。采用十折校验法进行评测，当仅用 TF-IDF 向量计算 7 个特征时，模型正确率为 82.1%，仅用 Word2Vec 向量时为 78.2%，使用全部特征时，正确率提高到 88.2%。

通过建立如相似度阈值、关键词过滤、时间过滤等规则，对模型结果进行修正，进一步提高结果的正确性。采用基于随机采样的人工评测法进行估算，链接的平均正确率在 95%以上。

表 3 特征相似度

Table 3 Feature Similarity

相似度类别	公式
文档相似度	$f_1=r(b,a)$
领域相似度	$f_2=r(b,C_{\mathrm{attr\_name}}(m))$ $f_3=r(b,C_{\mathrm{attr\_value}}(m))$ $f_4=r(b,C_{\mathrm{domain}})$
出入链相似度	$f_5=SR(b,O_{\mathrm{article}})$ $f_6=SR(b,O_{\mathrm{infobax}})$ $f_7=SR(b,I_{\mathrm{all}})$

2.5 大规模实体匹配

为了实现不同语言异构影视数据源的知识复用和融合，我们结合影视领域的实际情况，研究了基于 SF 的实体匹配算法，在中英文数据源之间进行大规模实体匹配工作。

2.5.1 基于 Similarity Flooding 的实体匹配算法

近几年来出现的比较优秀的大规模实体匹配算法大都借鉴了 SF 算法[9]的核心思想，并且在各自的应用场景中都取得不错的效果。如图 2 所示，SF 算法以两个图作为输入，输出对应结点的映射。SF 算法的主要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 广播。

文献[9]中，SF 是在小规模的异构本体 schema 数据集上实现的。根据相似度传播图的构建方法，图规模会随节点数量呈几何倍数增长。从表 8 的统计数据可以看出，影视作品和影视人规模皆在 10 万以上，按照原有算法，相似度传播图将达到 100

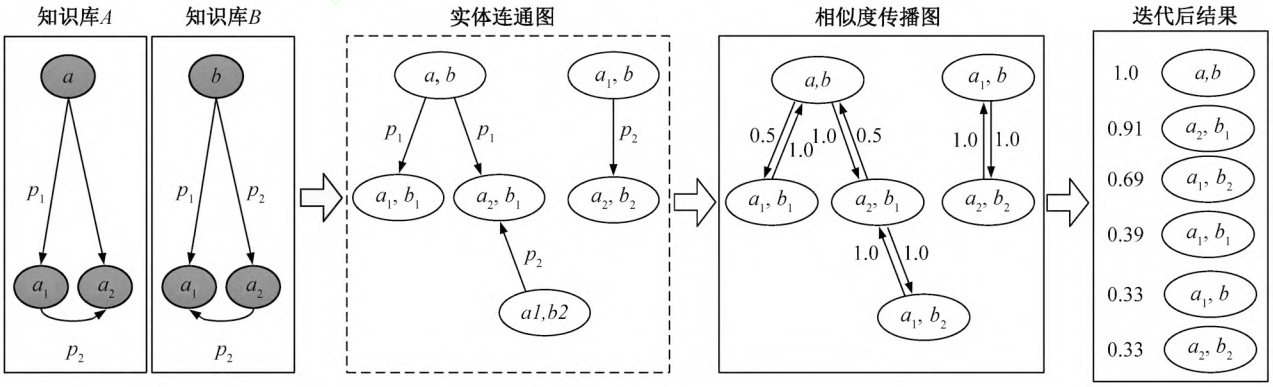

图 2 基于 Similarity Flooding 的实体匹配算法示意图[9]

Fig. 2 Example of the entity matching algorithm based on Similarity Flooding[9]

亿的规模，其计算量非常可观，所以必须减小图规模，算法才具有可行性。

在实际构图过程中，预先对实体对进行剪枝，具体步骤为：1)排除不同类别的实体对；2)排除不同上映年份的影视作品实体对和不同出生年份的影视人物实体对；3)计算候选实体对的相似度，剔除相似度低于一定阈值的实体对。剪枝之后，相似度传播图中的节点数量下降到 300 万左右，大大减少了算法的计算量。

除选择合适的匹配框架外，如何提炼数据中合适的内容特征和结构特征来计算实体之间的相似度，使相似度能够具有足够大的区分度，也是实体匹配任务的关键性问题。

2.5.2　实体相似度

实体的相似度主要考虑两个问题：实体主题词相似度(代表实体的标题信息)和属性相似度(代表实体的结构化信息)。

1) 实体主题词相似度。

实体的主题词，又称为实体标题词、标签词，是表达实体的核心词汇。除标题词外，影视实体通常还有一些别名，例如，影片“中国合伙人”的主题词和别名如表 4 所示。

归并实体别名、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，能够显著提高实体匹配的召回率。主题词相似度定义如下：

$$\mathrm{n\_sim}(a,b)=\max\left(\frac{lcslen(x,y)}{\max(len(x),len(y))}\right),\qquad(4)$$

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

2) 属性相似度。

不同类别属性，其相似度公式也不一样，通常有如下几种情况。

a) 二值型：

$$\mathrm{p\_sim}(x,y)=\begin{cases}1, & \text{当 } x==y,\\ 0, & \text{否则。}\end{cases}$$

b) 字符串型：

$$\mathrm{p\_sim}(x,y)=\frac{\mathrm{lcslen}(x,y)}{\max(\mathrm{len}(x),\mathrm{len}(y))}。$$

c) 数值型：

$$\mathrm{p\_sim}(x,y)=|x-y|/\max(x,y)。$$

d) 列表型：如演员表、代表作品等属性，其属性值通常是由多个实体组成的列表，相似度定义为

$$\mathrm{p\_sim}(x,y)=x\cap y/x\cup y。$$

3) 实体相似度。

综上所述，我们定义实体相似度为

$$\mathrm{e\_sim}(a,b)=\frac{1}{N+1}(\mathrm{n\_sim}(a,b)+\sum_{i=1}^{N}\mathrm{p\_sim}_i(a,b)),$$

其中 N 为属性数量，$\mathrm{p\_sim}_i(a,b)$ 为相应的属性相似度。

2.5.3　跨语言实体匹配

对于相同语言的知识库(如豆瓣和百度百科)，可以直接采用基于 SF 的匹配算法。对于跨语言实体匹配而言，关键在于建立不同语言实体之间的联系，克服相似度计算的语言障碍。文献[10]以中文维基为桥梁，基于维基百科页面中的多语言等价链接信息以及页面出入链信息来计算相似度，绕过了不同语言文本之间相似度的计算。

与文献[10]相同的是，通过影视数据源中普遍存在的 Imdb 链接，可以得到大量的等价实体。Imdb 链接具有全球唯一性，具有相同 Imdb 链接的实体是等价的，统计数据如表 5 所示。不同之处在于，我们所匹配的知识库是异构的，页面的内部链接不具有共指性，不能采用基于页面出入链的方法来计算相似度。但是，豆瓣和百度百科提供了大量的英文别名信息，如表 6 所示。双语词对的平均覆盖率在 60%以上，基于这些信息构建大规模双语映射词典，可以将部分命名实体映射为统一语言的文本。

表 4　影片主题词示例

Table 4　Film Name Example

数据源	标题词	别名
豆瓣电影	中国合伙人	三个中国先生 中国先生 海阔天空(台湾片名)
百度百科	中国合伙人(2013 年陈可辛执导电影)	海阔天空(台) 三个中国先生 American dreams in China

表 5　Imdb 链接统计

Table 5　Imdb links statics

数据源	Imdb 链接	链接率/%
中文数据源	161006	81.3
英文数据源	18696	11.7
匹配	13674	—

表 6　双语词对统计表

Table 6　Bilingual alignment word pairs statics

数据源	双语词对数	实体数	双语词对覆盖率/%
豆瓣电影	134998	197939	68.2
百度百科	56234	111862	50.2

表 7　实体统计表

Table 7　Entities statistics

数据源	影视作品数	影视人数
豆瓣电影	127406	70534
百度百科	69861	42012
LinkedMdb	85825	103348
DBpedia	98858	157756

事实上，在影视领域中，由于知识结构简单一致、信息量丰富，要判断两个实体是否相似，只需要使用实例的一部分信息即可，如判断两部电影是否相似，只要匹配影片名、年份、演员、导演、编剧、制片人等信息中 3~4 个，其正确率都在 95%以上。鉴于这种领域特点，即便只有六成多命名实体对覆盖率，基于部分文本相似度计算公式也有非常大的区分度。另外，我们还从其他如 Wikipedia, Freebase 等知识库中抽取更多的双语词对来提升映射词典的覆盖率，尽量避免由词典覆盖率不足带来的相似度矩阵稀疏性问题。通过这种部分映射的方法解决了跨语言实体相似度计算问题后，其他步骤与同语言实体匹配相同。

我们在上述 4 个知识库之间进行实体匹配，首先是同种语言数据源的实体匹配，然后根据匹配的实体进行数据源合并，最后将合并后的中英文数据源进行实体匹配。考虑实际数据情况，实验仅对知识库中主要实体进行匹配，统计数据如表 7 所示。

在以上 4 个异构数据源之间，我们进行 3 次不同的实体匹配: 1) 百度百科与豆瓣电影之间的中文实体匹配; 2) LinkedMdb 和 DBpedia 之间的英文实体匹配; 3) 在前面两步基础上，合并中英文数据集之间的跨语言实体匹配。

为了分析 SF 传播算法的性能，分别使用传播前后的实体相似度作为标准，考察不同阈值下的实体匹配结果。由于数据规模较大，且难以确定标准的数据集，所以采用随机抽样的人工评估方法。匹配结果如表 8 所示。

从表 8 可以发现:

1) 阈值对结果的正确率和正确匹配的数量影响很大。当阈值为 0.9 时，正确率很好，但是匹配数很少; 当阈值取 0.6 时，匹配数量大幅增加，而正确率却下降很快。

2) 使用 SF 传播算法后，匹配的正确率有了显著的提升。这是因为传播算法能够有效地降低错误匹配实例的相似度。例如，电影实体银行与 The Champion 间的相似度高达 0.8255, 因为二者均为卓别林于 1915 年导演的电影，进行 3 次 SF 算法迭代后，相似度降低到 0.6564。

3) SF 传播算法的召回率有所降低。由于相似度传播图的稀疏性(即节点的平均入度较小，导致部分节点的相似度无法得到充分传播)，会降低部

表 8　实体匹配结果

Table 8　Entity matching result

数据源语言	不使用 SF 传播算法				使用 SF 传播算法			
	阈值	匹配数	错误数	准确率/%	阈值	匹配数	错误数	准确率/%
中文	0.90	8362	8	99.90	0.90	8297	2	99.99
	0.75	18601	1670	89.02	0.75	17972	908	94.95
	0.60	34020	9703	71.47	0.60	29164	3129	89.27
英文	0.90	9065	12	99.87	0.90	8979	3	99.97
	0.75	19021	1936	89.82	0.75	17885	1095	93.88
	0.60	36659	9886	73.93	0.60	31431	3850	87.75
跨语言	0.90	8715	10	99.88	0.90	8647	2	99.99
	0.75	17022	2112	87.59	0.75	16262	1183	92.78
	0.60	33460	9112	72.17	0.60	28342	3767	86.71

分正确匹配实体对的相似度，从而使召回率有所降低。

另外，随着迭代次数的增多，引入错误的影响会随着相似度的传播而不断放大。因此，选择合适的迭代次数，对结果影响也比较大。

3　双语影视知识图谱共享平台

知识图谱是利用信息可视化技术构建的一种知识之间的关系网络图。我们建立了知识图谱共享平台①,目的是为了在概念、属性、实例等多个维度对BMKG 进行展示，并将实体之间的相互链接关系以可视化的形式表现出来。网站基于 Apache 开源框架进行开发，并采用 Virtuoso 作为数据库服务器，主要提供 3 个方面的功能：1)双语影视本体的基本信息，提供知识 Schema 和知识库的统计信息；2)数据查询接口，包括 SPARQL 终端查询接口、分类索引查询接口以及复合查询接口；3)知识网络的可视化，将实体之间链接关系以可视化的方式展现出来。

4　结论

本文提出了一种融合多个异构数据源的双语影视知识图谱的构建流程，并对整个过程中所遇到主要问题和挑战以及解决方法加以描述，旨在构建语义一致、结构一致的中英文双语影视本体知识库。首先，我们构建了双语影视本体 BMO，为中英文影视知识的提供一个规范性的描述框架，并通过 5 个影视结构化抽取过程，统一了各个数据源语义描述；在实体链接问题上，我们总结了多种属性相似度的计算方法，并基于两种不同向量模型来表示文档向量，使实体的相似度特征增加一倍，显著提升了实体链接的效果；在大规模实体匹配方面，我们利用简单的相似度传播模型进行大规模的实体匹配，实验结果表明，对于结构化较好的影视知识，使用传统的相似度传播算法模型，能够取得非常好的效果，另外，我们利用数据源中存在的影视中英文别名关系，构建不同语言同义词之间的映射对，克服了计算实体之间相似度上的语言障碍，实现了跨语言实体匹配。当然，由于所采用数据源的限制，BMKG 能够建立的影视知识属性和概念还比较少，影视知识的描述也不够丰富，这也在一定程度上影响了实体链接和实体匹配的效果。大规模实体链接和实体匹配技术都是非常具有挑战性的工作，如何充分利用知识库中的知识，改进模型的效果，是未来需要研究的课题。

事实上，构建本体知识库是一项长期性的、系统性的复杂工作，需要不断改进和完善。BMKG 有待改进的地方还很多，比如寻求质量更好的中英文影视知识源来扩展知识库；建立更多种类的链接关系(例如人物的合作者关系、影视系列关系等)，解决不同数据源之间知识冲突；建立知识库的自动更新机制；增加影视评论知识等等。本体知识库的构建没有一个通用的构建流程，本文提出的方法对需要融合多个数据源的领域本体知识库的构建以及在限定领域中进行大规模实体链接和实体匹配具有一定借鉴意义。

总体来说，BMKG 是融合了 4 个异构优质的影视数据源的高质量 RDF 影视本体知识库，填补了国内在中文影视本体知识库方面的空白，该知识库为影视信息的挖掘和利用提供重要的语料基础，同时，对扩大中文影视信息的国际化影响也具有重要意义。

① http://166.111.68.66:10080/KegMovieKB/KegMovie_Index.html

参考文献

[1] Miller G A. WordNet: a lexical database for English. Communications of the ACM, 1995, 38 (11): 39–41

[2] Hassanzadeh O, Consens M. Linked movie data base // Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), Madrid, 2009: 1–5

[3] 宣腾．区域医疗本体知识库构建及其语义应用[D]. 成都：电子科技大学, 2013

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

[5] Lehmann J, Robert I, Max J, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal, 2014, 5: 1–29

[6] Suchanek, Fabian M, Serge A, et al. Paris: probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment, 2011, 5(3): 157–168

[7] Lacoste J S, Palla K, Davies A, et al. Sigma: simple greedy matching for aligning large knowledge bases // Proceedings of the 19th ACM SIGKDD international

conference on Knowledge discovery and data mining. Chicago, 2013: 572–580

[8] Li Juanzi, Jie Tang, Yi Li, et al. Rimom: a dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8): 1218–1232

[9] Melnik S, Hector G M, Erhard R. Similarity flooding: a versatile graph matching algorithm and its application to schema matching // Proceedings of 18th International Conference on Data Engineering. San Jose, 2002: 117–128

[10] Wang Zhichun, Li Juanzi, Wang Zhigang, et al. Cross-lingual knowledge linking across wiki knowledge bases // Proceeding of the 21st international conference on World Wide Web. New York, 2012: 459–468

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

[12] 刘颖. 基于 Web 结构的表格信息抽取研究[D]. 安徽:合肥工业大学, 2012

[13] 曾道建, 来斯惟, 张元哲, 等. 面向非结构化文本的开放式实体属性抽取. 江西师范大学学报: 自然科学版, 2013, 37(3): 279–283

[14] 赵军, 刘康, 周光有, 等. 开放式文本信息抽取. 中文信息学报, 2011, 25(6): 98–110

[15] Xu Mengling, Wang Zhichun, Bie Rongfang, et al. Discovering missing semantic relations between entities in Wikipedia // The Semantic Web–ISWC 2013. Berlin, 2013: 673–686

[16] Mikolov T, Kai C, Greg C, et al. Efficient estimation of word representations in vector space [J/OL]. (2013)[2015–05]. http://arxiv.org/pdf/1301.3781.pdf