# 基于非结构化文本增强关联规则的知识推理方法

李智星[1,2] 任诗雅[1,2] 王化明[1,2] 沈 柯[1]
(重庆邮电大学计算机科学与技术学院 重庆 400065)[1] (计算智能重庆市重点实验室 重庆 400065)[2]

**摘 要** 知识图谱用一种结构化的方式存储实体、实体的属性以及实体之间的关系。由于知识图谱中的知识易于被计算机处理,因此它在许多自然语言处理任务中都起着至关重要的作用。虽然从绝对数量来看,现有的知识图谱已经包含了海量的三元组事实,但是与真实世界中存在的知识相比它远远不够。因此,如何完善知识图谱成为目前的研究热点。现有的研究方向主要分为内部推理和外部抽取两类,然而这些方法仍有很大的提升空间:一方面,由于知识图谱内部知识存在错误或缺失,可能会在推理时产生错误的扩散;另一方面,现有的知识抽取方法主要集中于对实体类型、关系等知识的抽取,从而导致抽取的知识不够全面。鉴于此,提出了一种基于非结构化文本增强关联规则的知识推理方法。该方法从非结构化文本表述中抽象出文本表述模式,并以词语分布袋的形式对其进行表示,进而结合知识图谱已有的知识构建关联规则。与传统关联规则的区别在于,该方法得到的关联规则可以通过与非结构化文本匹配的方式来完成知识推理。实验结果表明,与传统方法相比,该方法可以高效地从非结构化文本中推理出数量更大且质量更高的三元组知识。

**关键词** 知识图谱完善,关联规则,知识推理,文本增强,三元组知识
**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/jsjkx.181001939

## Knowledge Reasoning Method Based on Unstructured Text-enhanced Association Rules

LI Zhi-xing[1,2] REN Shi-ya[1,2] WANG Hua-ming[1,2] SHEN Ke[1]
(Coolege of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[1]
(Chongqing Key Lab of Computation Intelligence, Chongqing 400065, China)[2]

**Abstract** Knowledge bases (KBs) store entities, entity attributes and relations between entities in a structured manner. Because the knowledge in the KBs can be easily processed by computers, KBs play a vital role in many natural language processing (NLP) tasks. Although current KBs contain massive triple knowledge from the perspective of absolute quantity, they are far less than the knowledge existing in real world. Therefore, many researches focus on how to enrich the knowledge base with more high-quality knowledge. Internal reasoning and extracting from external resources are two main kinds of methods for knowledge base completion, but they still need to be improved. On the one hand, since the knowledge in KBs are not perfect and some errors exist, reasoning on such error knowledge will cause error propagation. On the other hand, existing extracting methods usually focus on limited relations and properties and thus cannot find comprehensive knowledge from external resources such as texts. In light of this, this paper proposed a knowledge reasoning method based on unstructured text-enhanced association rules. In this method, the text representation pattern is abstracted from the unstructured text firstly, then it is represented in the form of a bag of distribution, and the association rules can be mined through combining the knowledge of KBs. The difference from the traditional association rules is that the association rules obtained by the proposed method can directly match unstructured texts for knowledge reasoning. Experimental results show that the proposed method can efficiently infer triple knowledge from unstructured text with higher quality and larger quantity compared with traditional methods.

**Keywords** Knowledge bases completion, Association rules, Knowledge reasoning, Text-enhanced, Triple knowledge

**到稿日期**:2018-10-18 **返修日期**:2019-02-01 **本文受国家重点研发计划项目**(2016QY01W0200),**国家自然科学基金青年项目**(61502066),**重庆市基础与前沿研究计划项目**(cstc2015jcyjA40018)**资助**。

**李智星**(1985-),**男**,**博士**,**副教授**,**主要研究方向为自然语言处理、机器学习**,E-mail:lizx@cqupt.edu.cn(**通信作者**);**任诗雅**(1994-),**女**,**硕士**,**主要研究方向为自然语言处理、知识图谱**;**王化明**(1995-),**男**,**硕士**,**主要研究方向为自然语言处理、多粒度计算**;**沈 柯**(1996-),**女**,**主要研究方向为自然语言处理**。

## 1 引言

知识图谱(Knowledge Graph/Base)[1]最早由谷歌发布,其主要作用是提高搜索引擎返回答案的质量以及用户查询的效率。由于知识图谱包含了大量的结构化知识以及特殊的存储结构,使得它在许多自然语言处理应用中起着至关重要的作用,例如问答系统[2]、实体链接[3]等。近年来,一些大型的知识图谱,例如 DBpeida[4],Wikidata[5],Yago[6],Freebase[7]等,受到了越来越多的关注。虽然这些知识图谱包含了数以千万计的实体以及数以亿计的三元组事实,但是与真实世界中存在的知识相比,它们仍然不够完善。因此,完善知识图谱(Knowledge Bases Completion,KBC)成为当前的研究热点,它主要是指将新的实体、关系、实体属性及属性值加入到知识图谱中。目前,完善知识图谱的方法主要集中在两个方面:1)使用知识图谱内部知识推理完善知识图谱;2)从非结构化文本中抽取新的知识来完善知识图谱。

知识图谱以一种结构化的形式存储知识并且其本身包含了大量的知识,鉴于这种特性,使用知识图谱内部知识推理完善知识图谱成为目前完善知识图谱的主流方向之一。其主要有两种方法:1)利用表示学习的方式,将知识图谱中的实体和关系嵌入到一个低维的向量空间,然后利用一个评价指标计算三元组事实成立的概率;2)利用逻辑推理的方式,从知识图谱中学习类似 $rel_1(e_1,e_2)\wedge rel_2(e_2,e_3)\rightarrow rel_3(e_1,e_3)$ 形式的规则。然而,这些方法仅仅对知识图谱中存在的实体起作用,并不能增加新的实体信息。而且由于知识图谱存在错误的信息,基于知识图谱内部的知识图谱完善还可能造成错误传播等问题。

完善知识图谱需要从外部资源获取新知识。随着互联网的发展,网络上的文本信息急剧增加,如网络新闻、产品说明、用户评论等。这些信息包含了大量的碎片化知识,如何有效抽取这些碎片化的知识并将其与现有知识图谱进行整合是目前研究的一个热点。由于这些文本信息大都以非结构化自然语言的形式存在,计算机无法直接对其进行有效处理,因此如何理解并利用这些信息是一个非常具有挑战性的问题。现有的方法所抽取到的知识往往是非结构化文本中所包含知识的一小部分。例如,对于非结构化文本"Lisa is widow of film director Donen",利用常识可以推理出 Lisa 和 Donen 的类型(人类)、性别,Donen 的职业(电影导演),Lisa 和 Donen 的关系(配偶)以及 Donen 已经去世了等知识。但仅仅将关系分类算法应用于该文本,则只能得到 Lisa 和 Donen 的关系以及 Donen 的职业等一小部分知识。虽然一些研究也关注于联合抽取实体间的关系及实体的类型,但是这些方法所抽取的类型十分有限。

针对以上问题,本文提出了一种基于非结构化文本增强关联规则的知识推理方法,用以完善知识图谱。该方法从非结构化文本表述中抽象出文本表述模式,并结合知识图谱已有的知识构建规则。规则的形式主要包含两种:一种形式的规则前件包含非结构化文本模式,后件包含三元组事实;另一种形式的规则前件包含非结构化文本模式及三元组事实,后件包含三元组事实。使用该规则可以将非结构化文本中包含的三元组事实推理出来用以完善知识图谱。实验结果证明了本文方法的有效性。

## 2 相关工作

### 2.1 文本模式建模

文本模式建模的任务是从一系列相似的文本中抽象出一个统一的表达形式。传统的词袋模型(Bag of Words,BoW)以词语出现的频率来表示文本,虽然它在语言建模和文本分类等任务上取得成功,但是其稀疏性、忽略了单词的上下文信息、忽略了单词的位置信息等缺点对自然语言处理任务还是有较大的影响。为了缓解远程监督产生的噪声以及提升关系分类的性能,REHESSION[8]定义了一些关系的常用表达模板。但是由于非结构化文本的多样性,使用这些特定的模板建模文本是不现实的。RLSW[9]提出了一种词语分布袋(Bag of Distribution,BoD)模型来对相似的文本建立一个统一的表达模式。它使用 Beta 分布来拟合每个单词在一个文本集中的位置,此后该文本集可用一系列的 Beta 分布来表示。相较于其他方法而言,BoD 具有可以根据任何文本集生成模板的特性,因此其更适用于对非结构化文本进行模式建模,并为后续的规则挖掘提供服务。然而,BoD 只关注主语和宾语之间的信息,忽略了主语和宾语的前后信息,而这些被忽略的信息可能含有重要知识。

### 2.2 基于知识图谱内部知识推理完善知识图谱

考虑到知识图谱内部的结构特征及其内部包含的大量知识,基于知识图谱内部知识推理完善知识图谱的研究吸引了较多的关注。其主要分为两个研究方向:1)表示学习的方法;2)逻辑推理的方法。其中,表示学习的方法的主要思想是将知识图谱内部的实体和关系嵌入到一个低维的向量空间,然后利用一个评价指标计算三元组事实成立的概率,例如 TransE[10]认为知识图谱中正确三元组事实的向量表示会满足 $h+r=t$,即头实体的向量表示加上关系的向量表示应该等于尾实体的向量表示,然后定义目标函数,最后得到实体和关系的向量表示。通过这样的方式将知识图谱内部的实体和关系转化为向量后,可以推理出更多的三元组事实以完善知识图谱。然而 TransE 在处理复杂关系时遇到了困难,例如一对多关系、多对一关系、多对多关系。TransH[11]克服了 TransE 这一缺点,对于每一个关系,他都假设其落在一个超平面上,类似于 TransE 模型在该超平面上进行,这样的方式使得同一个实体向量在不同关系下有不同的表示。另一种逻辑推理的方法主要是从知识图谱中学习类似 $rel_1(e_1,e_2)\wedge rel_2(e_2,e_3)\rightarrow rel_3(e_1,e_3)$ 形式的规则。例如,AMIE[12]基于开放世界假设(Open World Assumption,OWA)挖掘规则,提出了一种新的方法模拟负例,然而它的搜索策略在大型知识图谱的使用中受限,因此 AMIE+[13]提出了一系列的修剪策略及查询重写技术,使其模型可以在大型的知识图谱中更有效地挖掘规则。RDF2Rules[14]通过挖掘知识图谱中的频繁模板(Frequent Predicate Cycles,FPCs)来生成规则。但由于知识图谱自身的不完整性和不准确性,这些方法存在着一些潜在的问题,而且仅对知识图谱中存在的实体和关系起作用。

### 2.3　结合非结构化文本完善知识图谱

知识图谱的完善主要是指将新的实体、关系、实体属性及属性值加入到知识图谱中。碎片化的非结构化文本中包含大量知识且更新速度迅猛，因此如何从非结构化文本中抽取知识用以完善知识图谱也得到了较多的关注。例如，一些关系分类的方法[15-16]使用远程监督的方式对齐知识图谱和自然文本，然后再利用各种算法辨别实体间的关系。目前，还有一些比较流行的知识抽取方法使用联合抽取的方式，同时得到实体间的关系及实体的类型，然而这些方法所抽取的类型十分有限，例如文献[17-18]仅能得到类似人类、地点等较粗粒度的实体类型，文献[19]所抽取的实体类型也被局限在实体类型层次中。

不同于以上的研究，本文主要通过非结构化关联规则来搭建知识图谱和非结构化文本之间的桥梁。使用这些规则可以直接从非结构化文本中推理出新的三元组知识，因此这些规则具备从大规模非结构化文本中发现和整合碎片化知识的能力。

## 3　问题定义

给定一系列的句子 $S$ 和事实集合 $F$。这些句子包含实体 $e\in E$ 以及实体所对应的三元组事实 $f(e_1,e_2)=\langle e_1,rel,e_2\rangle\in F$。这里，定义要挖掘的一阶关联规则的形式如下：

$$(ptn,e_1,e_2)\rightarrow f(e_1,e_2)$$

其中，$ptn$ 是用来拟合 $S$ 的一个文本模式，$f(e_1,e_2)$ 是 $e_1$ 或 $e_2$ 相关的三元组事实。由于 $ptn$ 是从非结构化文本中得到的，因此被命名为非结构化文本增强的关联规则。例如，给定规则 $(ptn',e_1,e_2)\rightarrow\langle e_1,gender,male\rangle$，且文本“$Jim\_e_1$ is the son of $Cameron\_e_2$”与 $ptn'$ 匹配，则可以推测出三元组 $\langle Jim,gender,male\rangle$。更进一步地，如果知识图谱中包含关于 $e_1$ 和 $e_2$ 的三元组事实，那么就可以推断出更多的三元组事实。例如 $\langle Cameron,Citizenship,American\rangle$ 在知识图谱中存在，那么就可以很容易地推断出 $\langle Jim,Citizenship,American\rangle$。因为在大多数的情况下，一个孩子的国籍与他父母的相同。因此，结合知识图谱中的知识，可以将一阶规则扩充为二阶规则：

$$(ptn,e_1,e_2)\wedge f(e_1,e_2)\rightarrow f'(e_1,e_2)$$

## 4　文本增强的关联规则挖掘

本文的主要目的是构建关联规则，并从非结构化文本中抽取三元组事实。由于非结构化文本的多样性，直接点对点地匹配规则是不现实的。普遍来说，表达相似的文本可能包含了相同的三元组事实。例如“Jim is the son of Cameron”和“Kimi is the first son of Jack”的主语都是男性，主体和客体都是父子关系。因此，一个很自然的想法是将表达相似的句子聚类来挖掘三元组事实，然后将这些句子整合成一个统一的模式放入到规则中。根据以上描述，本文方法的整体框架如图 1 所示，其主要由 3 个部分组成：关系文本聚类、文本模式建模以及非结构化关联规则挖掘。

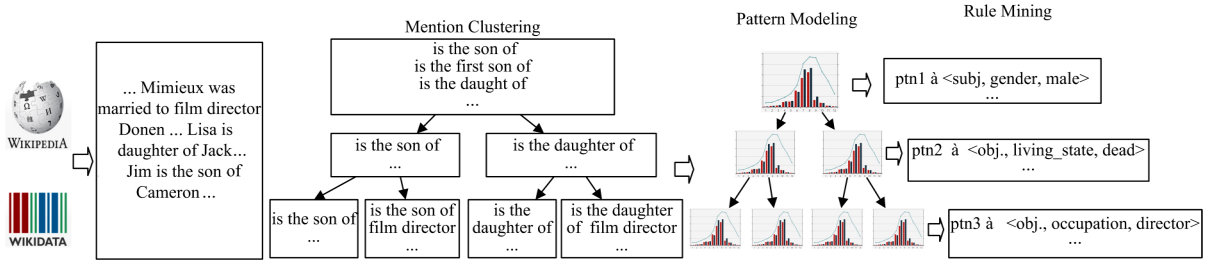



**图 1　基于非结构化文本增强关联规则的知识推理方法的整体框架**

Fig. 1　Knowledge reasoning framework based on unstructured text-enhanced association rules

### 4.1　关系文本聚类

#### 4.1.1　关系文本收集

一般来说，相同关系或属性的文本表达方式更相似。本文利用远程监督的方式收集相似文本，具体步骤如下：1）从 Wikidata 中收集预先定义关系所对应的实体对 $(e_1,e_2)$。2）爬取 $e_1$ 对应的 Wikipedia 文章，匹配文章中包含实体对的句子。实体对的匹配主要包含完全匹配、同义词匹配、部分匹配以及人称代词匹配等。3）对于每个句子，截取 $e_1$ 和 $e_2$ 前后的 3 个单词以及它们中间的单词作为关系文本。

#### 4.1.2　关系文本聚类

同种关系的文本虽然表达十分相似，但可能代表了相反或完全不相同的意思，例如“$e_1$ is father of $e_2$”和“$e_1$’s father is $e_2$”，若仅仅使用词汇信息而忽略词汇顺序则无法区分 $e_1$ 和 $e_2$ 到底谁是父亲。LSWMD[9] 是 Yang 于 2017 年提出的一种计算句子之间相似度的算法，该算法在计算句子之间的相似度时使用了单词的语义信息及语法信息，具体公式如下：

$$loc(w_i)=\frac{1}{n}*(i-0.5) \tag{1}$$

$$d(w_1,w_2)=\alpha*ed(w_1,w_2)+(1-\alpha)*|loc(w_1)-loc(w_2)| \tag{2}$$

其中，$w_i$ 是关系文本中的单词，$ed(w_1,w_2)$ 是 $w_1$ 和 $w_2$ 之间的欧几里得距离。

由于本文所截取的关系文本包含了更多的单词，因此重新定义 $loc(w_i)$ 如下：

$$loc(w_i)=2*\frac{i-idx(e_1)}{idx(e_2)-idx(e_1)}-1 \tag{3}$$

关系文本的相似度计算完成后，使用聚类算法完成关系文本的聚类。本文使用基于密度峰值的聚类算法[20]。

### 4.2　文本模式建模

一个类簇由语义和语法都相近的关系文本组成，为了将这些类簇放入到非结构化关联规则中，需要将其表示为一个统一的文本模式。传统的词袋模型只使用了单词的词频信息，忽略了单词的位置。文献[9]提出的词语分布袋（Bag of

Distribution,BoD)模式可以用来表示一个类簇。该方法使用 Beta 分布拟合单词在一个类簇的位置分布,再根据单词的频次排序,用高频单词的 Beta 分布表示该类簇。但是 BoD 仅仅对主语和宾语之间的单词进行建模,忽略了主语和宾语前后的单词,这可能会导致重要的信息丢失,因此本文提出了一个改进的 BoD(BoD* )来建模关系文本。具体步骤为:1)计算类簇中每个单词出现在关系文本中的位置集合;2)每个单词的位置集合用一个高斯分布来拟合。一个类簇所对应的 BoD* 模式可以表示为如下形式:

$$BoD^*(c)=\{(\mu_i,\sigma_i,p_i)\mid w_i\in W_c\}$$

其中,$c$ 表示一个类簇,$W_c$ 是 $c$ 中出现过的所有单词,$p_i$ 是 $w_i$ 出现在 $c$ 中的频次,$\mu_i$ 和 $\sigma_i$ 是单词 $w_i$ 对应位置集合的均值和标准差。

通过上述方法,可以将一个类簇表示为 $BoD^*(c)$,因此初始化的关联规则 $(ptn,e_1,e_2)\rightarrow f(e_1,e_2)$ 可以转化为 $BoD^*(c)\rightarrow f(e_1,e_2)$。对于任意的非结构化文本 $S$,它包含的实体对 $(e_1',e_2')$ 已知,如果它能匹配到一个合适的 $BoD^*(c)$,那么该 $BoD^*(c)$ 所对应的规则后件 $f(e_1,e_2)$ 就可以赋值给实体对 $(e_1',e_2')$。$f$ 的选择将在下一节进行讨论。

### 4.3 非结构化文本增强的关联规则挖掘

#### 4.3.1 规则构建与挖掘

由于每个类簇包含的是语法和语义都相似的关系文本,因此这些关系文本包含的实体可能拥有相同的三元组事实,该三元组事实由实体、实体在知识图谱中存在的属性及对应的属性值组成。本文所使用到的属性如表 1 所列。需要注意的是,为了找到三元组事实之间的相似性,实体对被替换为 $e_1$ 和 $e_2$。如果仅仅使用相同的三元组事实作为规则的后件,可能导致规则的数量太少,因此针对每个规则可以计算支持度和置信度,保留支持度和置信度大于预先设定好的阈值的规则。对于支持度和置信度的计算,受到关联规则挖掘算法思想的启发,将类簇中的每个关系文本对应的三元组事实表示成一个事务(*tranction*)的形式。为了与定义的一阶规则和二阶规则对应,从事务集中挖掘频繁一项式和频繁二项式。针对每个类簇挖掘的频繁一项式 $f$,规则 $BoD^*(c)\rightarrow f$ 将被加入到一阶规则集中。同理,针对每个类簇挖掘的频繁二项式 $(f,f')$,规则 $BoD^*(c)\wedge f\rightarrow f'$ 和 $BoD^*(c)\wedge f'\rightarrow f$ 将被加入到二阶规则集中。一阶规则 $(BoD*(c)\rightarrow f)$ 对应的支持度与置信度的计算公式如下:

$$sup(r)=\frac{|\{t\mid f\in t\ \&\ t\in T\}|}{|T|}\tag{4}$$

$$conf(r)=\frac{\{|t|\ f\in t\ \&\ t\in T\}|}{|\{t\mid f\otimes t\neq\emptyset\ \&\ t\in T\}|}\tag{5}$$

其中,$T$ 是所有的事务集集合,$t$ 中包含一些三元组事实;$f\otimes t\neq\emptyset$ 表示 $f$ 对应的属性必须存在于 $t$ 中。

二阶规则 $(BoD*(c)\wedge f\rightarrow f')$ 对应的支持度与置信度的计算公式如下:

$$sup(r)=\frac{|\{t\mid f,f'\in t\ \&\ t\in T\}|}{|T|}\tag{6}$$

$$conf(r)=\frac{|\{t\mid f,f'\in t\ \&\ t\in T\}|}{|\{t\mid f'\otimes t\neq\emptyset\ \&\ t\in T\}|}\tag{7}$$

表 1 规则挖掘中使用的属性及其在 Wikidata 中对应的 ID

Table 1 Properties and their IDs in Wikidata used in rules mining

| 属性名 | ID | 属性名 | ID |
|---|---|---|---|
| country of | P27 | genre | P136 |
| place of birth | P19 | religion | P140 |
| sex or gender | P21 | headquarters | P159 |
| place of death | P20 | country of origin | P495 |
| instance of | P31 | date of birth | P569 |
| country | P17 | inception | P571 |
| capital | P36 | publication date | P577 |
| occupation | P10 | Capital of | P1376 |
| place of interment | P11 | date of death | P570 |

表 1 中,$idx(e_1)$ 和 $idx(e_2)$ 分别是 $e_1$ 和 $e_2$ 在关系文本中的位置。

#### 4.3.2 细粒度规则挖掘

规则挖掘取决于其对应的支持度和置信度大小,一些规则的支持度和置信度可能在更细粒度的类簇中拥有较高的值。因此,为了挖掘更多、更细粒度的规则,本文使用一个自顶向下的层次聚类算法以产生不同粒度的类簇。在不同粒度的类簇进行规则挖掘时,孩子类簇可以利用继承父亲类簇规则后件的形式形成新的规则。当出现重复后件时,保留一个即可。

#### 4.3.3 冲突消解

基于相同的 $BoD^*(c)$ 可能会产生一些冲突的后件,比如 $BoD^*(c)\rightarrow Male(e_1,gender)$ 和 $BoD^*(c)\rightarrow Female(e_1,gender)$ 表示 $e_1$ 的性别既是男性又是女性,但该事实显然是不成立的。在实验中,如果两个规则存在冲突,则保留拥有更高置信度的规则。

## 5 实验

### 5.1 数据集描述及实验设置

实验中所涉及的关系是 Wikidata 中包含实体对较多的关系,实体所对应的 Wikipedia 页面内容被作为自然语言文本资源。数据集被分为训练集和测试集,比率为 7∶3。表 2 列出了更详细的数据,其中有效实体对是指至少有一个关系文本可以从 Wikipedia 文章中挖掘的实体对数量。

表 2 实验数据集

Table 2 Experiment datasets

| 关系 | 全部实体对 | 有效实体对 | 句子 |
|---|---|---|---|
| Human/spouse | 20000 | 12291 | 14609 |
| Song/composer | 3250 | 2134 | 2422 |
| Human/mother | 20072 | 9639 | 10778 |
| Film/director | 19233 | 14520 | 15422 |
| Human/father | 30000 | 18847 | 23434 |
| Human/child | 30000 | 13049 | 15874 |
| Film/castmember | 10000 | 5943 | 6852 |
| Song/performer | 11257 | 8773 | 9417 |

本文使用层次聚类的方式来挖掘多粒度的规则。因为在实验中发现如果只使用一层聚类,总会有一个类簇包含大量的关系文本,而有些类簇包含十分少的关系文本。随着聚类层次的加深,每个类簇包含的句子(事务)越来越少,这使得类簇表达能力不强,因此类簇深度不宜太深。实验中,类簇深度被设置为 3。

所有实验都在 PC 机上运行，其配置为 Intel(R) Core (TM) i7-6850K CPU @ 3.6 GHz，64 GB 内存，UBUNTU 16.04 操作系统，Python3.6 编程语言。

### 5.2　对比方法

本文的主要目的是推理出非结构化文本中包含的三元组事实，因此在实验中使用三元组推理来评估非结构化关联规则的性能。给定一个关系文本 $m=w_1,w_2,\cdots,w_n$，采用以下 4 种方法来推理三元组事实。

1) $BoD^*(c)\rightarrow f(e_1,e_2)$：该方法使用 $BoD^*$ 作为规则的前件，式(8)用于计算 $m$ 的规则类型：

$$\hat{c}=\arg\max_{c}\sum_{i=1}^{n}p_{w_i}*\int_{h}^{t}\frac{1}{\sigma_{w_i}\sqrt{2\pi}}e^{-\frac{(x-\mu_{w_i})^2}{2\sigma_{w_i}^2}}dx \tag{8}$$

其中，$p_{w_i}$，$\mu_{w_i}$，$\sigma_{w_i}$ 分别对应 $w_i$ 在 $BoD^*(c)$ 中的概率、均值、标准差；$t=0.5*(loc(w_{i-1})+loc(w_i))$；$h=0.5*(loc(w_i)+loc(w_{i+1}))$。若 $BoD^*(c)$ 中不包含 $w_i$，则 $p_{w_i}=0$。

2) $BoD(c)\rightarrow f(e_1,e_2)$(RLSW[9])：该方法使用 $BoD$ 作为规则的前件，式(9)用于计算 $m$ 的规则类型：

$$\hat{c}=\arg\max_{c}\sum_{i=1}^{n}p_{w_i}*\int_{b}^{e}Beta(\alpha_{w_i},\beta_{w_i}) \tag{9}$$

其中，$\alpha_{w_i}$，$\beta_{w_i}$ 分别为使用 Beta 分布拟合 $w_i$ 位置分布时的参数。

3) $BoW(c)\rightarrow f(e_1,e_2)$：使用传统的词袋模型作为规则的前件，利用该方式也可以得到与 $m$ 最相近的规则。

4) 多标签分类(magpie[1])：为了从多方面的角度验证提出方法的有效性，三元组推理被转化为多标签分类问题，每个三元组对应一个标签。magpie 是一个高效的基于深度学习的多标签文本分类工具，它实现了文献[21-22]中的方法。在对比实验中，我们将三元组知识推理转换为多标签分类问题，并采用 magpie 作为多标签分类的代表算法进行对比。

在方法 1)－3)中，对于一阶规则，只需根据相应的公式就可得到关系文本 $m$ 最符合的规则，再根据该规则的后件预测得出三元组事实。然而，对于二阶规则，因其前件包含了 $BoD^*(c)$ 和 $f$，所以还需判断 $f$ 是否存在于知识图谱中才能找到最符合的规则。对于每一个预测的三元组事实，我们将规则的置信度赋给三元组事实，并将其作为该预测的置信度。方法 4)可以根据算法的输出得到每个被预测的标签(即三元组事实)的置信度。将置信度进行倒序排序，计算被推理三元组事实的正确率。利用这种方式，每种方法都可以得到相应的 P/R 曲线；同时，也可以知道支持度和置信度的设置仅仅影响 P/R 曲线后半段的形状。因此，在实验中支持度阈值和置信度阈值被设置为支持度最大值和置信度最大值的 0.8 倍。

### 5.3　实验结果分析

#### 5.3.1　规则的数量

随着聚类层次的增加，类簇的粒度从粗变细，规则对应的置信度和支持度也随之增加，更多的规则能够被保留。表 3 列出了类簇深度对规则数量的影响。总的来说，规则的数量是随着聚类深度的加深而增加的，而增加的幅度取决于数据集。实验中，由于实体的类型(instance_of)在每个类簇中都有较高的置信度和支持度，没有对比价值，因此本文不予讨论。

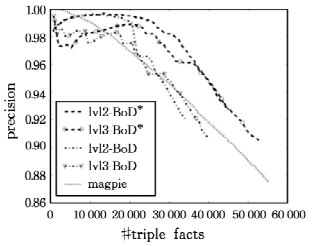
表 3　在聚类深度变化情况下的规则数量对比

Table 3　Comparison of number of rules when clustering depth changes

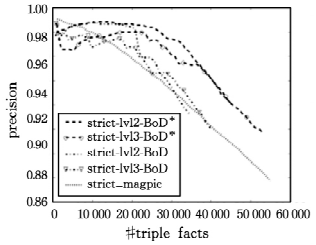
| 关系 | 一阶规则 | | | 二阶规则 | | |
|---|---|---|---|---|---|---|
| | lvl1 | lvl2 | lvl3 | lvl1 | lvl2 | lvl3 |
| spouse | 0 | 2 | 37 | 8 | 84 | 575 |
| composer | 1 | 10 | 20 | 2 | 21 | 48 |
| mother | 3 | 33 | 128 | 9 | 10 | 396 |
| director | 2 | 20 | 84 | 13 | 15 | 660 |
| father | 3 | 34 | 195 | 9 | 10 | 571 |
| child | 2 | 24 | 163 | 5 | 6 | 408 |
| castmember | 0 | 9 | 95 | 11 | 16 | 976 |
| performer | 1 | 11 | 34 | 0 | 8 | 35 |

#### 5.3.2　规则的质量

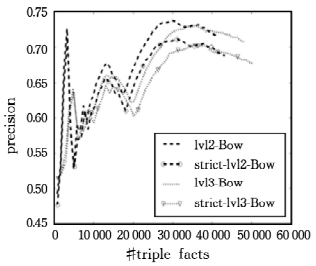
规则的后件包含三元组事实，如果一个关系文本符合规则的前件，那么该关系文本包含的三元组事实可以被预测。图 2 描述了规则的前件分别使用 $BoD^*$，$BoD$，$BoW$ 时预测三元组事实的结果。




(a)




(b)




(c)

图 2　三元组推理 P/R 曲线

Fig. 2　P/R curves of reasoning triple facts

为了从多方面表现本文方法的有效性，图 2 还描述了将三元组推理作为多标签分类时的预测结果。由于知识图谱的

1) https://github.com/inspirehep/magpie

不完整性造成了一些实体对应的属性缺失，在图 2 中有两种方法被用来评估不能判断正确与否的三元组事实：忽略缺失属性的三元组事实和将缺失属性的三元事实看作错误预测(strict)。总的来说，相比其他的方法，$BoD^*$ 的性能最优。虽然 $BoW$ 推理的三元组事实的数目与 $BoD^*$ 相当，但是它的准确率不够突出。相比 $BoD^*$ 而言，$BoD$ 仅仅关注了主语和宾语之间的单词，丢失了一部分重要的信息，从而导致它的性能低于 $BoD^*$。值得注意的是，虽然本文提出的方法没有复杂的训练阶段，对比基于深度学习的多标签分类，仍取得了更好的结果，这表明本文提出的方法具有更好的泛化性能。

### 5.4 实例分析

#### 5.4.1 规则挖掘

表 4 列出了一阶规则和二阶规则的输出样例。这里可以发现一些有趣的规则，比如 $Occupation(e_2, guitarist) \rightarrow Gender(e_2, male)$ 表示在某个类簇中大多数的吉他手都是男性；$Public\_date(e_1, before\text{-}90s) \rightarrow Gender(e_2, male)$ 表示于 1990 年之前出版的书籍大多数都是男性写的；$Female(e_2, gender) \rightarrow Male(e_1, gender)$ 表示大多数配偶都是异性等。

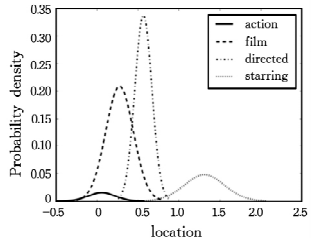
表 4 规则样例

Table 4 Examples of rules

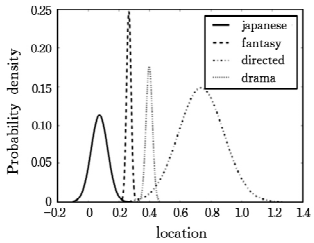
| $BoD^*$ $(\mu, \sigma, p)$ | 一阶规则 | 二阶规则 |
|---|---|---|
| (American,0.277,0.470,0.010),<br>(composed,0.368,0.293,0.013),<br>(written,0.408,0.951,0.034),<br>(music,0.453,0.379,0.009),… | Occupation($e_2$,songwriter),<br>Citizenship($e_2$,America),<br>Instance_of($e_1$,song),… | Country_of_origin($e_1$,America)→Citizenship($e_2$,America),<br>Occupation($e_2$,guitarist)→Gender($e_2$,male),<br>Public_date($e_1$,before-90s)→Gender($e_2$,male),… |
| (married,0.043,0.388,0.106),<br>(actor,0.582,0.144,0.014),<br>(actress,0.642,0.229,0.032),<br>(with,1.041,1.511,0.019),… | Male($e_1$,gender),<br>Female($e_2$,gender),<br>Occupation($e_2$,actor),… | Occupation($e_1$,actor)→Occupation($e_2$,film_actor),<br>Citizenship($e_1$,America)→Citizenship($e_2$,America),<br>Female($e_2$,gender)→Male($e_1$,gender),… |

#### 5.4.2 多粒度的 BoD*

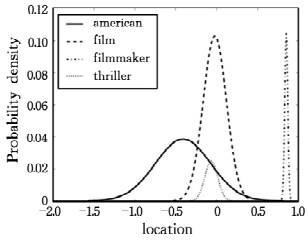
图 3 显示了不同粒度的 $BoD^*$ 结果。




(a)Coarse-grained cluster




(b)Fine-grained cluster




(c)Fine-grained cluster

图 3 类簇粒度样例

Fig. 3 Examples of cluster granularity

图 3(a)表示一个较粗粒度的类簇，它描述了频率较高的 4 个单词在类簇中的位置分布。从图 3(a)中可以发现，该类簇包含的知识是关于电影、导演等内容。图 3(b)和图 3(c)表示较细粒度的类簇。从图 3(b)中可知，该类簇包含的知识与日本、戏剧等有关，而图 3(c)表示的类簇包含的知识与美国电影有关。因此，当一个句子匹配到更细粒度的类簇时，可以推理出更详细的知识。

结束语 本文提出了一种基于非结构化关联规则的知识推理方法，使用该方法产生的非结构化规则可以直接推理出非结构化文本中包含的三元组事实，实验结果表明了该方法的有效性。然而，本文仍有许多不足之处，需要不断完善与改进。后续工作主要体现在以下几个方面：1)数据集的规模以及关系的数量需要进一步扩展；2)$BoD^*$ 模式的生成对于英语数据集相对容易，在未来的工作中，将尝试在更多的语言上挖掘 $BoD^*$ 模式。最后，目前 $BoD^*$ 模式是只针对类簇中的单词建立的，为了增加 $BoD^*$ 模式的泛化能力，未来将使用单词向量代替单词。

## 参 考 文 献

[1] QI G L, GAO H, WU T X. The Research Advances of Knowledge Graph[J]. Technology Intelligence Engineering, 2017, 3(1):4-25. (in Chinese)
漆桂林，高桓，吴天星. 知识图谱研究进展[J]. 情报工程，2017，3(1)：4-25.

[2] YIH W T, CHANG M W, HE X, et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base[C]// Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing. 2015:1321-1331.

[3] LU W, WU C. Literature Review on Entity Linking[J]. Technology Intelligence Engineering, 2015, 34(1):105-112. (in Chinese)
陆伟，武川. 实体链接研究综述[J]. 情报学报，2015，34(1)：105-112.

[4] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus

for a web of open data[M]// The semantic web. Springer, Berlin, Heidelberg, 2007: 722-735.

[5] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

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

[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. ACM, 2008: 1247-1250.

[8] LIU L, REN X, ZHU Q, et al. Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 46-56.

[9] YANG X, REN S, LI Y, et al. Relation Linking for Wikidata Using Bag of Distribution Representation[C]// National CCF Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2017: 652-661.

[10] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Advances in Neural Information Processing Systems. 2013: 2787-2795.

[11] WANG Z, ZHANG J, FENG J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]// AAAI. 2014, 14: 1112-1119.

[12] GALÁRRAGA L A, TEFLIOUDI C, HOSE K, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C]// Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 413-422.

[13] GALÁRRAGA L, TEFLIOUDI C, HOSE K, et al. Fast rule mining in ontological knowledge bases with AMIE $ $ + $ $ + [J]. The International Journal on Very Large Data Bases, 2015, 24(6): 707-730.

[14] WANG Z, LI J. RDF2Rules: Learning Rules from RDF Knowledge Bases by Mining Frequent Predicate Cycles[DB/OL]. (2015-12-24)[2018-08-20]. https://arxiv.org/abs/1512.07734.

[15] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[16] LIN Y, SHEN S, LIU Z, et al. Neural relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 2124-2133.

[17] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014, 1: 402-412.

[18] MIWA M, SASAKI Y. Modeling joint entity and relation extraction with table representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1858-1869.

[19] REN X, WU Z, HE W, et al. Cotype: Joint extraction of typed entities and relations with knowledge bases[C]// Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee. 2017: 1015-1024.

[20] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[21] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1746-1751.

[22] BERGER M J. Large scale multi-label text classification with semantic word vectors[R]. Stanford University, 2015.