文章编号：1003-0077(2018)04-0080-07

基于位置的知识图谱链接预测

张宁豫[1]，陈 曦[2]，陈矫彦[3]，邓淑敏[2]，阮 伟[4]，吴春明[2]，陈华钧[2]

(1. 之江实验室，人工智能与未来网络技术研究院，浙江 杭州 311121；
2. 浙江大学，计算机科学与技术学院，浙江 杭州 310058
3. 牛津大学，计算机科学系，英国，OX1 3QR；
4. 浙江大学，控制科学与工程学院，浙江 杭州 310058)

摘 要：链接预测是知识图谱的补全和分析的基础。由于位置相关的实体和关系本身拥有丰富的位置特征，该文提出了一种基于位置的知识图谱链接预测方法。该方法首先通过分析实体和关系的语义特征对关系进行分类，然后提出了一种基于位置的实体和关系位置特征和规则的挖掘方法；其次，通过挖掘出的实体位置特征和规则，对实体和关系的向量化方法预测结果进行约束，得到最终的结果。该文通过对 WikiData、FB 和 WN 数据集的实验，证明该方法针对基于位置的关系和实体链接预测拥有较好的效果。

关键词：位置特征；知识图谱；链接预测

中图分类号：TP391 **文献标识码**：A

Location Based Link Prediction for Knowledge Graph

ZHANG Ningyu[1], CHEN Xi[2], CHEN Jiaoyan[3], DENG Shumin[2], RUAN Wei[4],
WU Chunming[2], CHEN Huajun[2]

(1. Artificial Intelligence and Future Network Technology Research Institute, Zhejiang Lab, Hangzhou, Zhejiang 311121, China;
2. College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China;
3. Department of Computer Science, Oxford University, Oxford OX1 3QR, UK;
4. College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310058, China)

Abstract: Link prediction is the basis of complement and analysis of knowledge graph. To leverage the rich location characteristics in the location-related entities and in their relationships, this paper presents a location-based knowledge graph link prediction method. This method first classifies relations by analyzing the semantic features of entities and relationships, and then proposes a method for mining features and rules based on location-based entities and relationships. Secondly, by mining the entity location features and rules, we construct the constrains on the prediction results of vectorization methods for entities and relatiouships, and get the final results. Based on the experiments of WikiData, FB and WN datasets, we proved that the method has good effect on location-based relationship and entity link prediction.

Key words: location features; knowledge graph; link prediction

0 引言

知识图谱例如 FreeBase、Yago 等是很多人工智能应用的重要数据来源。它包含了海量的实体和关系并以三元组的形式进行存储。然而，大多数知识库的数据都是缺失的。所以知识库补全，也就是对现有的知识库进行链接，预测新的关系和实体是一项重要的工作。

现有的知识图谱链接预测方法大多都是直接利用实体、关系本身或图的特征来进行链接预测。对于给定的知识图谱，实体和关系通常会被映射成低维的向量。通过定义一个打分函数来对每一对实体和关系的三元组进行预测。实体和关系的向量可以通过

收稿日期：2017-03-16 **定稿日期**：2017-05-04
基金项目：国家自然科学基金(61673338)

最大化已知正确三元组的打分函数来训练获得。

然而，在训练实体、关系向量与打分函数的过程中，这类方法并没有利用实体和关系本身隐藏的位置特征。此外，由于实体和关系向量化方法数据驱动特点，如果训练结果中某一类关系或者实体数据量很小，训练出的这一关系或实体的向量针对打分函数可能会导致过拟合等问题。

事实上，现有的知识库中储存着海量的位置相关的实体和关系。例如，在三元组（鲁迅，WasBornIn，绍兴）中，实体“绍兴”有明确的位置特征。利用实体“绍兴”的属性可以获得位置特征，进而可以推测实体“鲁迅”隐含的位置特征，利用位置的隐含特征构造规则约束。例如，在判断三元组（鲁迅，WasBornIn，浙江）是否成立时，利用实体“鲁迅”的位置特征和空间位置的规则判断，可以约束判断的最终结果。

在本文中，我们提出了一种针对位置关系的基于向量化和规则的链接预测方法。位置相关的关系指的是三元组中至少含有一个实体，其属性或者本身含义带有位置的特点。例如，至少有一个实体是一个地名、一个区域名称、一个兴趣点名称等。

首先，针对基于位置的三元组，我们根据其特点把基于位置的关系分成了三类：包含关系、相邻关系和相交关系。包含关系是两个实体本身的地理坐标范围是相互包含的，例如 LoactedIn。相邻关系是指两个实体本身的地理坐标范围是相互分离的，但在一定距离内，例如 NearBy。相交关系是指两个实体本身的地理坐标范围是相互交叉的，例如 HasSameHometown。针对不同的实体，我们提取出不同的隐藏位置特征。针对不同的关系类型，我们提取不同的规则。实体的隐藏位置特征主要由实体本身的位置（如经纬度或地名）和它的辐射范围组成。规则主要分成两类：一类是通用规则。例如，两个实体间拥有 NearBy 关系必然会存在 HasNeighbour 关系，同时 NearBy 关系的实体必须是属于 Location 类型的。另一类是位置规则。例如，实体 h 和实体 t 的隐藏位置特征是后者包含前者，则两个实体间有可能存在包含这类的关系。最后，我们利用规则对向量化方法结果进行约束，得到最终的结果，如图 1 所示。

我们的方法有以下优点：(1)规则的使用降低了计算空间并提高了准确度；(2)保留了向量化方法的优点，同时加入了隐藏的位置信息；(3)它是一个通用的框架，能够适用各种通用的向量化方法和规则。

综上所述，本文的贡献如下：

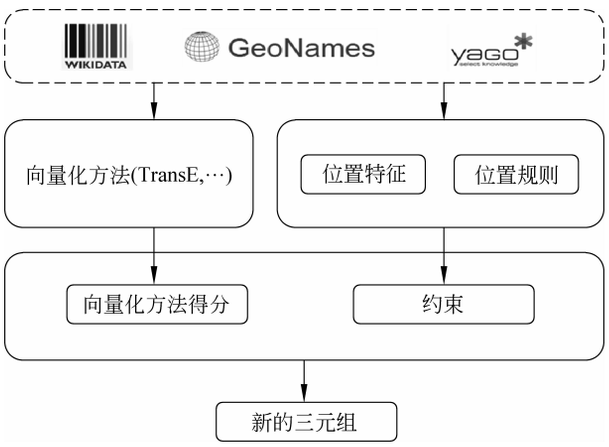

图 1　基于位置的向量化和规则链接预测方法

(1) 针对基于位置的三元组，我们提出了挖掘实体和关系位置特征的方法。

(2) 提出了一种针对位置关系的基于向量化和规则的链接预测方法。

(3) 利用 WikiData、FB 和 WN 的数据集进行实验，证明针对位置相关的链接预测，本方法比其他方法准确度有所提高。

1　相关工作

知识图谱的链接预测通常是指给定一组三元组，预测其成立的可能性。根据 Nickel Maximilian[1] 的研究，知识图谱链接预测通常分为三大类：(1)通过实体和关系的隐含特征将其转换成低维向量的方法[2-3]；(2)基于图特征的方法[4-5]；(3) 基于马尔科夫概率图利用一阶谓词逻辑[6]或者软逻辑（probabilistic soft logic）[7]来预测。

基于向量化的知识图谱链接预测方法的核心是用向量来表达实体和关系隐藏的特征。RESCAL[2] 和 TransE[8] 是两个典型的方法。它们通过最小化结构风险或边界误差来学习隐藏的向量。然而，在学习和预测的过程中，这类方法都没有利用潜在的位置特征和应用规则。TRESCAL[9] 将规则和 RESCAL 整合在了一起，但它仅能使用单一规则（例如某种关系的实体必须是特定的类型）。Rocktäschel 等[10] 提出了将一阶谓词逻辑映射成低维向量。但是他们的方法中规则并没有直接起到链接预测的作用，也没有降低预测的复杂度。Wang Q 等[11] 提出了一种基于整数线性规划（ILP）的方法，将向量化结果和规则整合起来进行链接预测，但是他们并没有利用潜在的位置特征和基于位置的规

则。基于图的方法核心是挖掘知识图谱图结构所有的特征。Lü Lin[12]挖掘节点之间的相似度来进行链接预测。Path ranking algorithm(PRA)[13]是利用节点之间不同通路包含的特征来进行预测,也可以提炼出规则来约束结果。但是,基于图特征的方法通常适合局部的链接预测,不一定能挖掘出全局的隐藏特征。我们方法的不同点在于提供了一个通用的利用位置特征和规则的预测框架,可以整合各种向量化方法和规则。

在马尔科夫网络中,规则已经被大量使用,代表性的研究有利用一阶谓词逻辑[6]和软逻辑(probabilistic soft logic)[7]。本文利用规则来约束向量化方法的结果,将整合问题变成一个整数规划问题。此外,我们挖掘出了隐藏的位置特征,构造了位置特征的规则。

2 方法

2.1 定义

定义 1(实体位置特征) 如果实体 e 能够在当前知识库或外部数据库如 Yago、GeoNames、LinkedGeoData 和 WikiData 中匹配到相应的位置(经纬度)和大致范围或所属上级的范围,则 e 有位置特征 $f_e=[\text{lng},\text{lat},D]$,lng 是经度,lat 是纬度,$D$ 是一个描述实体包含范围的数值,通常情况由实体本身的行政地域半径或上级所属区域半径最小值确定。

定义 2(位置相关三元组) 三元组(h, r, t)的实体 h、t 中至少有一个实体含有位置特征。

定义 3(包含关系) 实体 h 和 t 的位置特征存在 $\sqrt{|(h_{\text{lng}}-t_{\text{lng}})^2-(h_{\text{lat}}-t_{\text{lat}})^2|}<|h_D-t_D|$,则两者存在包含关系 HasContain$(h,t)$。

定义 4(相邻关系) 实体 h 和 t 的位置特征存在 $\sqrt{|(h_{\text{lng}}-t_{\text{lng}})^2-(h_{\text{lat}}-t_{\text{lat}})^2|}\geqslant|h_D+t_D|$,则两者存在相邻关系 HasAdjacent$(h,t)$。

定义 5(相交关系) 实体 h 和 t 的位置特征存在 $|h_D-t_D|\leqslant\sqrt{|(h_{\text{lng}}-t_{\text{lng}})^2-(h_{\text{lat}}-t_{\text{lat}})^2|}<|h_D+t_D|$,则两者存在相交关系 HasIntersect$(h,t)$。

2.2 框架

如图 2 所示,我们的系统由两部分组成:(1)位置特征和规则挖掘。首先对三元组中实体进行位置特征提取,然后对基于位置的三元组的关系进行自动识别或者人工标注分类,最后提取出其他可能存在的位置特征和规则。(2)基于向量化和规则的链接预测。首先对三元组利用向量化方法进行训练,然后利用规则对结果进行约束。

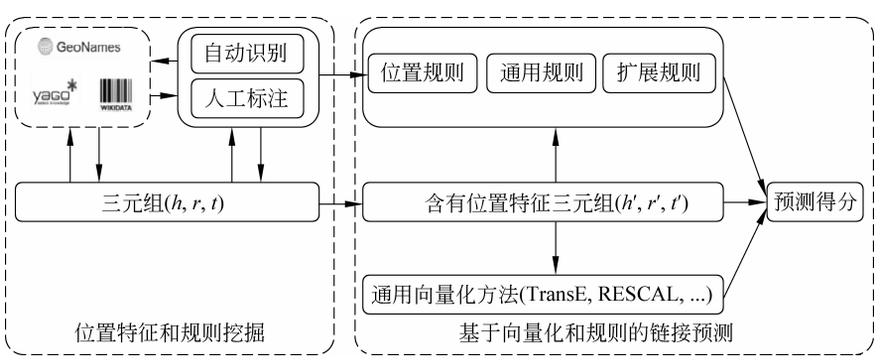

图 2 框架系统的组成

2.3 隐含的位置特征和规则挖掘

给定一个基于位置的三元组(h,r,t),首先我们需要提取出三元组中实体可以直接获得的位置特征。例如,三元组(鲁迅,WasBornIn,绍兴)中,通过对实体“鲁迅”和“绍兴”的类型和本地数据库以及外部数据库 Yago、GeoName、LinkedGeoData 和 WikiData 的匹配得到,实体“绍兴”是一个地名。我们可以获得该实体的经纬度、面积、相邻城市等信息。通过近似计算(利用面积或相邻区域经纬度),我们可以获得实体“绍兴”的位置特征。然后我们需要获得关系“WasBornIn”的类别,即它属于包含、相邻、相交哪一类。一般地说,有两种做法:(1)自动识别。遍历所有三元组中两个实体都含有位置特征的三元组,通过反向计算实体位置特征的差异,推导出此三元组拥有的关系,对常见的如 LocatedIn、Nearby 等

关系，此方法可以方便地判别；(2)人工标注。事实上，基于位置的关系总数并不多，再者，通常整个知识图谱需要预测的关系数量级也不是很大，远小于实体个数数量级。所以可以采取人工标注的方法来解决额外的关系分类问题。最后，我们通过已经获得的关系“WasBornIn”属于包含关系，判断实体“鲁迅”隐藏位置特征，该特征和实体“绍兴”的位置特征存在包含关系。这个知识可以作为规则，为后续的未知链接预测做约束。

具体地说，对于任意三元组(h,r,t)，如果只有实体t可以直接获得位置特征$f_t=[t_{\text{lng}},t_{\text{lat}},t_D]$，根据关系$r$我们可以推测实体$h$隐含的位置特征。如果$r$属于包含关系，则$h$可能存在隐含位置特征$[t_{\text{lng}},t_{\text{lat}},t_D-\mu]$，其中$0<\mu<t_D$。如果$r$属于相交关系，则$h$可能存在隐含位置特征$[h_{\text{lng}},h_{\text{lat}},h_D]$，其中$|h_D-t_D|\leqslant\sqrt{|(h_{\text{lng}}-t_{\text{lng}})^2-(h_{\text{lat}}-t_{\text{lat}})^2|}<|h_D+t_D|$，也就说是$h$位于一个环状区域范围内。如果$r$属于相邻关系，则$h$可能存在隐含位置特征$[h_{\text{lng}},h_{\text{lat}},h_D]$，其中以上变量满足条件$\sqrt{|(h_{\text{lng}}-t_{\text{lng}})^2-(h_{\text{lat}}-t_{\text{lat}})^2|}\geqslant|h_D+t_D|$。反之，如果实体$h$含有隐藏位置特征，以此来推导$t$，也是如此。事实上，对于相交和相邻关系，大多数三元组的两个实体本身都可以直接获取位置关系。以上的隐藏特征都是近似特征。

由此，我们可以获得海量的实体隐藏位置特征和规则。事实上，可以获得以下规则：

规则 1(实体类型匹配)　特定的关系拥有特定类型的实体。例如，关系 LocatedIn 拥有的两个实体一定是 Location 类型的；关系 WasBornIn 拥有的两个实体一定是一个是 Person 类型，一个是 Location 类型。

规则 2(参数个数匹配)　一对多和多对一的关系中特定实体的数目有一定限制。例如 CityLocatedInCountry 是一个多对一的关系。给定一个城市实体，在知识图谱中最多存在一个国家实体与之对应。

规则 3(相似关系匹配)　如果关系r_1和r_2存在一定的牵连或同属于同一个类型(同是包含类型)，在不违背规则 1、2 的前提下，则拥有r_1关系的实体可能存在r_2关系。例如，CityCapitalOfCountry->CityLocatedInCountry。

规则 4(位置包含关系)　如果两个实体的位置特征存在包含关系，则两个实体可能存在包含关系。例如，实体“鲁迅”和实体“浙江”的位置关系存在包含关系，则两个实体很大程度上存在包含关系。

规则 5(位置相邻关系)　如果两个实体的位置特征存在相邻关系，则两个实体可能存在相邻关系。例如，实体“西湖”和实体“浙江大学”的位置关系存在相邻关系，则两个实体很大程度上存在相邻关系。

规则 6(位置相交关系)　如果两个实体的位置特征存在相交关系，则两个实体可能存在相交关系。例如，实体“金庸”和实体“徐志摩”的潜在的位置特征存在相交关系，则两个实体可能存在相交关系。

规则 7(位置包含传导)　如果实体e_2的位置特征包含实体e_1的位置特征，实体e_3的位置特征包含实体e_2的位置特征，则实体e_3和e_1存在包含关系。包含关系可以一直连续传递，相邻和相交关系不能传递。例如，实体“鲁迅”和实体“浙江”存在包含关系，实体“浙江”和实体“中国”存在包含关系，则实体“鲁迅”和实体“中国”存在包含关系。

此外，如果未知的一对一关系的三元组中，其中一个实体和关系存在于已知三元组正样本中，那这个三元组很可能是不成立的。对于一些特殊的实体，可以通过几重的关系链传递估计出位置特征的信息。例如，三元组(鲁迅，说，中文)，实体“中文”的位置特征可以通过关系如“中国人说中文”、“中国人出生在中国”、“绍兴位于浙江”、“浙江位于中国”和“绍兴位于中国”等估计得到，其位置特征大致和实体“中国”的位置特征接近，从而估计出实体“中文”的位置特征。

2.4　基于向量化和规则的链接预测

给定一个知识图谱，其包含n个实体，m个关系。我们可以获得三元组集合$O=\{h,r,t\}$。向量化方法的目的在于：(1)通过隐含的特征把实体和关系映射到一个向量；(2)利用训练好的向量来预测新三元组成立的可能性。本文中我们利用了三种成熟的向量化方法：RESCAL、TRESCAL、TransE。

RESCAL 将每个实体e_i当成一个向量$e_i\in R^d$，每个关系r_k都是一个矩阵$R_k\in R^{d\times d}$。给定一个三元组(e_i,r_k,e_j)，它的打分函数如式(1)所示。

$$f(e_i,r_k,e_j)=e_i^TR_ke_j \tag{1}$$

$\{e\}$和$\{r_k\}$是通过最小化下面的结构损失函数来获得的，如式(2)所示。

$$\min_{\{e_i\},\{R_k\}}\sum_k\sum_i\sum_j(y_{ij}^{(k)})-f(e_i,r_k,e_j))^2+\lambda R \tag{2}$$

其中，如果三元组(e_i,r_k,e_j)成立，则$y_{ij}^{(k)}$等于1，反之为0。R是正则项。λ是正则化参数，控制正则化和损失函数之间的平衡。

TRESCAL 是 RESCAL 算法的一个扩展，需要

对给定关系的实体类型进行约束。例如，给定关系 r_k 和分别包含特定类型的实体集合 H_k，T_k，则问题变成优化问题，如式(3)所示。

$$\min_{\{e_i\},\{R_k\}} \sum_k \sum i \in H_k \sum_{j \in T_k} (y_{ij}^{(k)}) - f(e_i, r_k, e_j))^2 + \lambda R \tag{3}$$

TransE 将三元组(e_i, r_k, e_j)映射成以下的三个向量 $e_i, r_k, e_j \in R^d$，它使用以下的打分函数来计算三元组成立的可能性，如式(4)所示。

$$f(e_i, r_k, e_j) = || e_i + r_k - e_j || \tag{4}$$

其中$\{e_i\}$、$\{r_k\}$ 是通过优化式(5)的边缘损失函数(正确样本得到更高的得分，错误样本得分更低)来得到：

$$\min_{\{e_i\},\{r_k\}} \sum_{t^+ \in O} \sum_{t_- \in N} [\gamma - f(e_i, r_k, e_j) + f(e'_i, r_k, e'_j)]_+ \tag{5}$$

其中 t^+ 是正样本，O 是正样本的集合，t_- 是负样本，N 是负样本的集合。在替换过程中我们未采用随机替换，而是替换之后确保新的三元组在原始的数据集中存在确定的关系，但关系不是 r_k，这很大程度上确保了样本是负样本。我们利用随机梯度下降的方法来求解优化问题。

利用上述方法，对未知的三元组，打分高的一般情况下成立的可能性较高，反之较低。我们将向量化方法得分的输出记为 $y_{ij}^{(k)} = f(e_i, r_k, e_j)$，每个实体的位置特征记为 f_i、f_j，标记相交关系集合 $R_{\text{intersect}}$ 含三元组 s 对，相邻关系集合 R_{adjacent} 含三元组 p 对，包含关系集合 R_{contain} 含三元组 q 对，标记一对多、多对一、一对一关系集合 R_{1-M}，R_{M-1}，R_{1-1}，标记特定关系所属实体种类的集合 H_k、T_k。用逻辑变量 $x_{ij}^{(k)}$ 来标记这个三元组成立的最终可能。根据文献[11]我们把规则约束向量化结果的问题定义为一个整数规划的问题如式(6)所示①。

$$\begin{aligned}
&\max_{x_{ij}^{(k)}} \sum_k \sum_i \sum_j y_{ij}^{(k)} x_{ij}^{(k)} \\
&s.t. \quad R1.\ x_{ij}^{(k)} = 0, \forall k, \forall i \notin H_k, \forall j \notin T_k, \\
&\quad R2.\ \sum_i x_{ij}^{(k)} \leqslant 1, \forall k \in R_{1-M}, \forall j; \sum_i y_{ij}^{(k)} \leqslant 1, \forall k \in R_{M-1}, \forall i; \\
&\quad\quad \sum_i y_{ij}^{(k)}, \sum_j y_{ij}^{(k)} \leqslant 1, \forall k \in R_{1-1}, \forall i, \forall j; \\
&\quad R3.\ x_{ij}^{(k_1)} \leqslant x_{ij}^{(k_2)}, \forall r_{k_1} -> r_{k_2}, r_{k_1}, r_{k_2} \in R_{\text{contain}}, r_{k_1}, r_{k_2} \in R_{\text{adjacent}}, r_{k_1}, r_{k_2} \in R_{\text{intersect}}, \forall i, j, \\
&\quad R4.\ \sum_k y_{ij}^{(k)} \geqslant q\delta_1, \forall k \in R_{\text{contain}}, \forall f_i, f_j \text{HasContain}(e_i, e_j), \\
&\quad R5.\ \sum_k y_{ij}^{(k)} \geqslant p\delta_2, \forall k \in R_{\text{adjacent}}, \forall f_i, f_j \text{HasAdjacent}(e_i, e_j), \\
&\quad R6.\ \sum_k y_{ij}^{(k)} \geqslant s\delta_3, \forall k \in R_{\text{intersect}}, \forall f_i, f_j \text{HasIntersect}(e_i, e_j), \\
&\quad R7.\ x_{it}^{(k)} \leqslant x_{ij}^{(k)}, \forall k \in R_{\text{contain}}, \forall f_i, f_t \text{HasContain}(e_i, e_t), \forall f_t, f_j \text{HasContain}(e_t, e_j), \\
&\quad R8.\ x_{ij}^{(k)} = 0, \forall i, k \in O, \forall j \notin O, \forall k \in R_{1-1}
\end{aligned} \tag{6}$$

其中 $x_{ij}^{(k)} \in \{0,1\}$，$\forall i, j, k, O$ 是正样本集合。通过解答上述问题求得最终的得分 $x_{ij}^{(k)}$。

我们的方法优势如下：(1) 在向量化方法的前提下，利用位置和通用规则，使含有显性和隐性位置特征的三元组链接预测准确率有明显的提高；(2)这是一个通用的框架，向量化方法和规则都可以灵活变化。

3 实验

实验的具体流程如下：(1)位置特征和规则挖掘；(2)基于向量化和规则的链接预测；(3)分析位置特征和规则对结果的影响。

3.1 数据集

在实验中我们使用了三个数据集：WikiData-500K、WN-100K、FB-500K，分别从 WikiData[14]、WordNet[15]、FreeBase[16]获取。WikiData 是目前较大的一个开放的知识图谱。WikiData 包含有 human、taxon、administrative territorial、architectural structure、event、chemical compound、film、thoroughfare、astronomical object 等类型的实体组成的三元组信息。据我们统计有至少 19.8%的三元组中至少有一个实体

含有位置信息(事件、行政区划、地点等)[①]，可以直接通过 API 获取。我们由此构建了 WikiData-500K 数据集。WN-100K 和 FB-500K 都是由不同学者发布出的三元组数据集。我们从 WN-100K、FB-500K 筛选出位置相关的三元组来进行训练。具体地说，在完整知识库中至少 30% 的三元组都满足条件要求。此外，我们还利用 Yago[②]、GeoNames[③]、LinkedGeoData[④] 和 WikiData 对所有数据中的实体进行位置信息匹配，以获得实体本身的位置特征。我们过滤了数据集中出现次数少于三次的实体，并采用了文献[8]的方法来判断实体的关系是一对多还是多对一来制定规则。此外，我们制定了一些同类匹配的规则。实验数据集如表 1 所示。

表 1　实验数据集

数据集	实体	关系
WikiData-500K	位置相关实体 14 561 所有实体 15 321	231
WN-100K	位置相关实体 5 325 所有实体 38 696	231
FB-500K	位置相关实体 5 612 所有实体 14 951	11

3.2　特征和规则挖掘

我们的任务是提取出实体隐含的位置特征。首先，对数据集中所有的实体进行位置信息匹配。利用外部数据集拥有的准确地理位置信息匹配数据集中实体，大约 40% 的实体能匹配到准确的位置特征。然后，我们对数据集中拥有的关系进行分类。利用自动分类方法标记了约 63% 的关系，剩下的关系采用人工标记。事实上，有约 5% 的关系是有歧义的，我们将它们默认归到包含关系类。最后利用位置特征和关系类型挖掘剩下的实体隐藏位置特征。

3.3　链接预测

我们的任务是补全位置相关的三元组(h,r,t)，也就是说，给定 h 和 t 预测 r 或者给定 h 和 r 预测 t，或者给定 r 和 t 预测 h。本节中测试了 RESCAL、TRESCAL、TransE，并把利用基于位置的规则来约束向量化结果的方法命名成 l-RESCAL、l-TRESCAL、l-TransE。

对每个数据集，我们把基于位置的三元组按照 4∶1 的比例划分成训练集和测试集。对每一个实体我们都获得其所属类型。对于测试三元组，通过计算命中@10(正确命中结果排前十所占的比例)来衡量。在具体实验中，RESCAL、TRESCAL 的正则化参数 $\lambda=0.1$，我们迭代训练了十次。在向量化训练过程中，我们将维度分别设置成 10，20，50，100 来选择最优的参数。然后利用集成学习的方法获得三种向量化方法的最优结果。在规则约束的过程中，$\delta_1=0.7$，$\delta_2=0.6$，$\delta_3=0.4$，我们使用 lp solve[⑤] 来解整数规划问题。我们对规则约束重复进行了 20 次取平均值，以获得最优的结果。

表 2 展示了不同数据集下不同关系进行关系预测的结果。可以看出，利用基于位置的规则方法对特定的关系有显著的提高。RESCAL 和 TRESCAL 的提升幅度比 TransE 要高。

表 2　位置相关关系命中@10 结果/%

关系	RESCAL	l-RESCAL	TRESCAL	l-TRESCAL	TransE	l-TransE
CityLocatedInState	56.1	**67.1**	57.3	**59.3**	55.9	**58.4**
CityLocatedInCountry	61.3	**66.5**	62.4	**62.8**	63.5	**64.1**
CityCapitalOfCountry	45.3	**46.1**	47.2	**47.5**	46.1	**46.4**
NearBy	34.3	**35.2**	35.2	30.2	34.2	**35.3**
WasBornIn	61.3	**65.5**	63.2	60.2	63.2	**65.3**
HasSameHometown	45.5	**45.6**	44.2	40.2	44.2	**45.3**
总平均值	55.2	**61.2**	56.5	**61.7**	57.5	**59.9**

3.4　位置特征和规则分析

我们还对不同关系类型和不同实体进行了结果的比较，如表 3 所示。从结果可以看出，对我们的方法，包含关系获得的提升程度较高，其次是相

① www.wikidata.org
② www.mpi-inf.mpg.de
③ www.geonames.org
④ www.linkedgeodata.org
⑤ lpsolve.sourceforge.net/5.5/

邻关系和相交关系。事实上,包含关系的位置隐含特征区域较为狭小,因此对关系的确定限制较大,可以获得较好的结果;而相邻关系和相交关系(实体都可以直接获得位置特征除外)获取的隐藏位置区域较大,因此限制较为不准确。对实体而言,两个实体都可以直接获得位置关系的预测结果提升幅度最大,其次是单一实体的结果。有趣的是,对于两个都不能直接获得位置信息的实体,本方法仍能获得少量的提升。事实上,例如判断三元组(徐志摩,HasSameHometown,金庸)时,实体"徐志摩"和"金庸"的隐藏位置特征是可以获得的,利用人工标记关系"HasSameHometown"为相交关系,使用我们的方法可以获得准确度的提升。

表3 不同类型关系命中@10结果/%

关系和实体	RESCAL	l-RESCAL	TRESCAL	l-TRESCAL	TransE	l-TransE
包含关系均值	55.3	**60.1**	56.1	**57.2**	55.2	**56.8**
相邻关系均值	35.3	**37.9**	35.5	**37.8**	34.5	**35.4**
相交关系均值	41.2	**42.2**	40.2	39.2	39.8	**40.5**
两个实体含位置	57.2	**78.2**	55.6	**70.3**	60.2	**70.0**
单个实体含位置	55.8	**60.2**	50.2	49.2	48.5	**50.3**
两个实体都不含位置	48.4	**50.2**	49.6	**49.9**	51.6	49.3

4 结论

本文提出了一种针对位置关系的基于向量化和规则的链接预测方法。实体位置特征和规则的使用降低了计算空间,提高了基于位置链接预测的准确度。我们还对位置特征和规则进行了实验分析。

实验结果证明,对于特定类型的关系,位置特征和规则的利用可以使链接预测的准确度得到一定程度的提高。将来,我们计划:(1)分布式我们的方法,使得它能够适用于更大的数据集;(2)加入更加复杂的空间规则;(3)尝试在向量化训练的同时直接利用规则,以提高准确度。

参考文献

[1] Nickel M, Murphy K, Tresp V,et al. A review of relational machine learning for knowledge graphs[J]. Proceedings of the IEEE,2016,104(1): 11-33.

[2] 李阳,高大启. 知识图谱中实体相似度计算研究[J]. 中文信息学报, 2017,31(1): 140-146.

[3] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases [C]//Proceedings of Conference on Artificial Intelligence. 2011: 1923-1944.

[4] Lao N, Mitchell T, Cohen WW. Random walk inference and learning in a large scale knowledge base. [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 529-539.

[5] Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 601-610.

[6] Jiang S, Lowd D, Dou D. Learning to refine an automatically extracted knowledge base using Markov logic [C]//Proceedings of the 12th International Conference on Data Mining. 2012: 912-917.

[7] Pujara J, Miao H, Getoor L, et al. Knowledge graph identification [C]//Proceedings of International Semantic Web Conference. 2014: 542-557.

[8] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of Advances in Neural Information Processing Systems. 2013: 2787-2795.

[9] Chang K W, Yih S W, Yang B. Typed tensor decomposition of knowledge bases for relation extraction [C]//Proceedings of Conference on Empirical Methods on Natural Language Processing. 2014: 1568-1579.

[10] Rocktäschel T, Bosnjak M, Singh S,et al. Low-dimensional embeddings of logic[C]//Proceedings of the ACL 2014 Workshop on Semantic Parsing. 2014: 45-49.

[11] Wang Q, Wang B, Guo L. Knowledge base completion using embeddings and rules[C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence. 2015: 1859-1865.

(下转第129页)

孙晓(1980－)，博士，副教授，主要研究领域为自然语言处理，智能人机会话及相关机器学习算法的研究与开发。
E-mail：sunx@hfut.edu.cn

任福继(1959—)，博士，教授，主要研究领域为自然语言处理，人工智能，语言理解与交流，情感计算等。
E-mail：ren2fuji@gmail.com

张陈(1993—)，硕士，主要研究领域为自然语言处理，微博情感分析，异常检测.神经网络等。
E-mail：1725685823@qq.com

(上接第 86 页)

[12] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[13] Lao N, Cohen W W. Relational retrieval using a combination of path constrained random walks[J]. Machine Learning, 2010, 81(1): 53-67.

[14] Vrande či ć D, Krötzsch M. Wikidata: A free collaborative knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78-85.

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

[16] Bollacker K, Cook P, Tufts, P. Freebase: A shared database of structured general human knowledge [C]//Proceedings of the 21st AAAI Conference on Artificial Intelligence, 2007(7): 1962-1963.

张宁豫(1989—)，博士，主要研究领域为语义挖掘。
E-mail：zhangningyu@zju.edu.cn

陈矫彦(1988—)，博士，主要研究领域为语义挖掘。
E-mail：jiaoyanchen@zju.edu.cn

陈曦(1990—)，博士，主要研究领域为语义挖掘。
E-mail：xichen@zju.edu.cn