11: 获得 $F_1,F_2,\ldots,F_l$;
12: end for
13: for each in $F_1,F_2,\ldots,F_l$ do:
14: 根据连接规则得到候选关系路径 C_P;
15: 记录连接的种子实体对数 n;
16: if $n\geqslant\tau$ then
17: 把相应的关系路径添加到频繁关系路径集 F_P 中;
18: end for
19: for each $path\in F_P$ do
20: 添加实体类型得到元路径 p;
21: $P\leqslant p\cup P$;
22: $SP\leqslant n\cup SP$;
23: end for
24: return P, SP;
25: end procedure

3.2 元路径的权重学习

算法 FPMPG 产生了重要元的路径 P,但是针对实体集扩展问题,不同元路径的重要性是不同的.因此,如何对这些元路径进行整合就变得尤为重要.实体集扩展可以看作是建立一个排序模型进而对候选实体进行排序,取恰当的前 k 个结果作为扩展集合,本文设计的排序模型如公式(1)所示:

$$R(c_i,S)=\frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{l}w_k\times r\{(c_i,s_j)\mid p_k\} \tag{1}$$

其中,c_i 代表第 i 个候选实体;$S=\{s_1,s_2,\ldots,s_m\}$是种子集合;l 是元路径数目;$r\{(c_i,s_j)|p_k\}$表示在路径 p_k 下种子实体 s_j 和候选实体 c_i 的相关性;w_k 是元路径 p_k 的权重值,即需要学习的权重值.本文中,我们设计了两种权重学习方法:一种是启发式方法,一种是 PU learning 方法.下面详细介绍这两种方法.

3.2.1 启发式方法

直观上,若一条元路径连接的种子对越多,就越能反映种子实体之间的共同特征,也就越重要.基于这个思想,我们设计了一种与文献[9]类似的启发式方法,即,采用元路径连接的种子对数和所有种子对数的比值来衡量元路径的重要性.公式如下:

$$w_k=\frac{\dfrac{|SP_k|}{m\times(m-1)}}{\sum_{k=1}^{l}\dfrac{|SP_k|}{m\times(m-1)}}=\frac{|SP_k|}{\sum_{k=1}^{l}|SP_k|} \tag{2}$$

其中,$|SP_k|$代表元路径 P_k 连接的种子对数目,$m\times(m-1)$是总的种子对数,m 为种子数目,公式中的分母表示权重的归一化项.

3.2.2 PU learning 方法

PU learning 方法的主要思想是:利用少量的正例和没有标签的数据(包含潜在的正例和负例)来建立一个分类器进而用于没有标签的数据,判断其是否属于正例或者有多大概率属于正例[6].其主要特性是在训练模型时没有恰当的负例可以利用,而实体集扩展问题中有少量给定的种子正例和很多没有标签的数据,PU learning 可以用来对这样的数据进行学习.在本文中,我们采用文献[40]中的 PU learning 方法,它本质上是从这些非传统的输入数据中学习一个传统的分类器,并且基于这样一个假设,即,少量带标签的正例是从所有正例中随机选择的.这里,我们把种子对作为正例,把候选实体和种子组成的对作为没有标签的数据.另外,该 PU learning 方法可以调整训练的分类器,从而选择合适且效果好的分类器,因为实体集扩展问题中正例的数据是非常少的,诸如支

持向量机等方法并不适合,因此,这里我们选择 adaboost,它可以改变训练数据的分布,从而增加正例的重要性进而获得好的效果.

4 实 验

4.1 数据集

Yago 是一个大规模的知识图谱,它的数据主要来源于 Wikipedia、权威英文词典 WordNet 和著名数据库 GeoNames[13],以 RDF 数据结构描述.目前为止,它已经包含了超过 1 000 万的实体和超过 1.2 亿的事实记录.我们使用 Yago 中的“yagoFacts”“yagoSimpleTypes”和“yagoTaxonomy”这 3 部分,这些数据集中包含 35 种关系,超过 1 300 万的实体和 3 000 多种实体类型.表 2 是其数据描述.

Table 2 Description of the data

表 2 数据的描述

数据	三元组样式	#三元组
yagoFacts	⟨entity relation entity⟩	4 484 914
yagoSimpleTypes	⟨entity rdt:typewordnet_type⟩	5 437 149
yagoTaxonomy	⟨wordnet_typerdfs:subclassofwordnet_type_⟩	69 826

我们选择了 4 个具有代表性的实体集扩展任务来验证 FPMP_ESE 的性能.4 个扩展任务如下:配偶是演员且获得过艾美奖(Emmy award)的演员、在纽约的大学毕业的作家、获得过国家电影奖(national film award)奖项的导演导的电影、在位于马萨诸塞州剑桥(Cambridge of Massachusetts)的大学工作的科学家,分别记为 Actor*、Writer*、Movie*和 Scientist*,它们分别包含 193,60,653 和 202 个实例.

4.2 评价指标

实验中,我们采用 *percision*-at-*k*(*p@k*)和 Mean Average Precision(MAP)来评价算法的性能.*p@k* 是前 *k* 个结果中正确实例所占的比例,这里,我们使用 *p*@10、*p*@30 和 *p*@60.MAP 是 *p*@10、*p*@30 和 *p*@60 的平均准确度(average precision,简称 AP)的均值,这里, $AP=\frac{\sum_{i=1}^{k} p@i\times rel_i}{\#\text{of correct instances}}$,其中,若排在第 *i* 位的结果为正确实例,则 rel_i 为 1;否则为 0.

4.3 实验设置

本小节我们详细介绍实验的有关设置,将启发式和 PU learning 的权重方法相对应的实体集扩展方法分别记为 FPMP_ESE_He 和 FPMP_ESE_PU.因为已有的关于知识图谱中的实体集扩展问题的方法很少,因此我们设计了几种基本的方法 Link-Based、Neighbor、PCRW 和 MP_ESE.详细介绍如下.

- Link-Based:受文本或者网页中的基于模式的方法的启发[41],给出基于实体一跳链路关系的方法;
- Neighbor:受文献[34,35]的启发,给出同时考虑一跳链路和一跳实体的最近邻方法;
- PCRW:一种基于路径受限随机游走的相似性度量方法[42],这里,我们采用其广度优先搜索的策略,并且用是否连接来度量,设置路径长度是 1,2,3,分别记为 PCRW1,PCRW2,PCRW3;
- MP_ESE:最近,文献[9]提出了一种知识图谱中的实体集扩展方法,元路径是单向自动产生,然后采用简单的启发式方法进行整合的.

在算法 FPMP_ESE 中,我们根据经验设置支持数阈值 σ 为 $m-1$,关系路径阈值 τ 为 $m\times(m-1)/2+1$,最大路径长度为 4.其他算法分别设置最优参数.

4.4 有效性实验

在这一小节,我们将 FPMP_ESE 和其他基本方法进行比较,验证其在以上 4 个任务上的有效性.对每个任务,我们随机选择 3 个种子进行实验,实验运行 20 次取平均值,如图 3 所示.

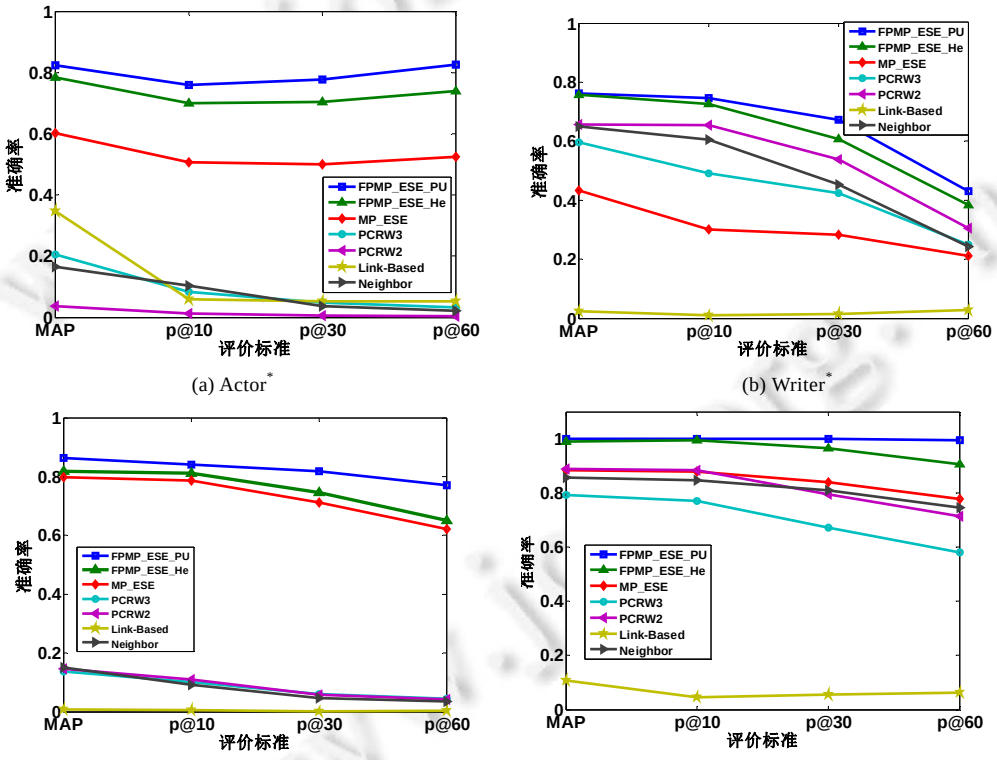
从图 3 中可以发现以下 3 种现象.

(1) 采用元路径的方法 MP_ESE 和 FPMP_ESE 较其他方法具有更好的性能.因为重要元路径可以捕捉种子实体之间潜在的共同特征,过滤掉一些噪音,从而进行更好的实体集扩展;

(2) 本文提出的方法 FPMP_ESE_He 和 FPMP_ESE_PU 较其他方法有更好的性能,因为 FPMP_ESE 可以尽可能全面地找到种子实体之间的重要元路径,不会因为一些潜在的因素剪掉某些重要元路径.例如,在 Actor*任务中,方法 MP_ESE 中寻找元路径的方法是单向搜索的,在搜索到第 3 跳时,其中的一条路径 $\xrightarrow{isMarriedTo}\xrightarrow{hasWonPrize}\xrightarrow{hasWonPrize^{-1}}$ 连接了种子对,之后在剪枝的步骤中,因设计的剪枝条件剪掉了这条路径.那么在后续的搜索过程中,我们就不可能搜索到长度为 4 且表达其语义(配偶是演员且获得过艾美奖(Emmy award)的演员)的路径 $\xrightarrow{isMarriedTo}\xrightarrow{hasWonPrize}\xrightarrow{hasWonPrize^{-1}}\xrightarrow{isMarriedTo^{-1}}$,从而会导致不好的扩展结果.而 FPMPG 算法由于是从每个种子实体进行扩展,然后进行有效连接,所以很好地避免了这一问题,因此会有更好的性能.对于 Link-Based 方法,在所有的任务上都有很差的性能,原因是它只考虑了一跳链路,具有很少的语义信息.对于 Neighbor 和 PCRW3 方法,性能也较差,原因是它们都只考虑了一跳链路和一跳实体,包含的信息也较少;

(3) FPMP_ESE_PU 较 FPMP_ESE_He 有更好的性能,说明与启发式的方法相比,PU learning 方法可以更好地学习到不同元路径的重要性,从而为不同的元路径分配更恰当的权重.

总之,FPMP_ESE 方法有最好的性能,因为它可以尽可能全面地找到种子实体之间重要的元路径,从而更好地捕捉种子实体之间潜在的共同特征,并且,PU learning 的方法可以学习到更加恰当的元路径权重,从而建立更恰当的实体集扩展模型.



(a) Actor* (b) Writer*

(c) Movie* (d) Scientist*

Fig.3 Results of entity set expansion on four tasks

图 3 在 4 个任务上的实体集扩展结果

为了更加直观地观察元路径的有效性,表 3 列出了在 Movie*任务上发现的前 3 条重要的元路径,其中,第 2 列 Gini 重要性表示进行 PU learning 训练后元路径的重要性,第 3 列表示采用启发式方法得到的元路径的重要性.从表 3 可以看出:第 1 条元路径很好地描述了 Movie*任务中种子实体的最重要的语义,即,这些电影都是获得过同一奖项的人导演的.其他元路径也揭示了种子实体的部分隐含信息,第 2 条元路径表明,某些导演也出演了自己导的电影,这在实际情况下也是很合理的.总之,提出的方法可以自动地找到这些有意义的元路径,并且分配恰当的权重,以便很好地发掘种子实体之间的重要语义关系,从而更好地进行实体集扩展.

Table 3 Top 3 meta paths for Movie*

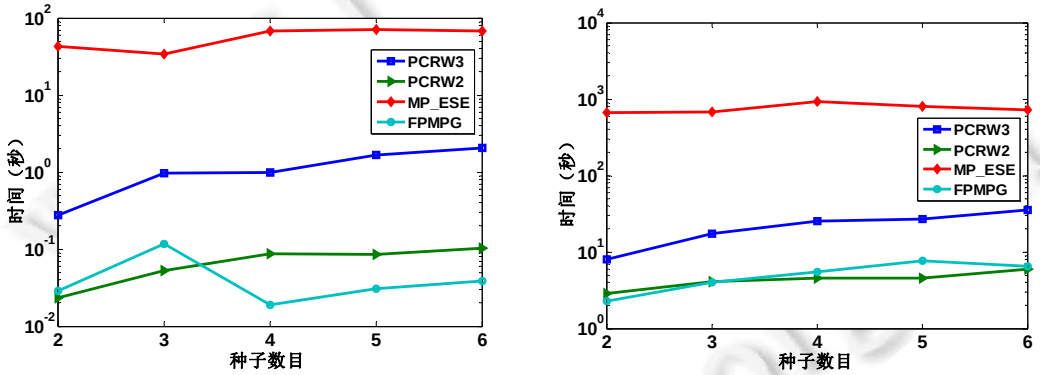
表 3 Movie*任务上最重要的前 3 条元路径

元路径	Gini 重要性	启发式权重
$Movie \xrightarrow{directed^{-1}} People \xrightarrow{hasWonPrize} Award \xrightarrow{hasWonPrize^{-1}} People \xrightarrow{directed} Movie$	0.120 77	0.026 77
$Movie \xrightarrow{actedIn^{-1}} People \xrightarrow{hasWonPrize} Award \xrightarrow{hasWonPrize^{-1}} People \xrightarrow{directed} Movie$	0.119 74	0.098 88
$Movie \xrightarrow{directed^{-1}} People \xrightarrow{hasWonPrize} Award \xrightarrow{hasWonPrize^{-1}} People \xrightarrow{actedIn} Movie$	0.112 56	0.098 88

4.5 效率实验

本小节我们比较采用不同方法寻找元路径的时间,主要从两个角度来研究,即,种子数目和不同的种子组合对寻找路径的效率的影响.

在种子数目对寻找路径时间的影响上,我们分别在 Movie*和 Scientist*任务上选取 2~6 个种子进行实验,对不同种子数目,我们分别在相应的任务上随机选取同等规模的种子进行实验,重复 20 次取平均值,结果如图 4 所示.



(a) Movie* (b) Scientist*

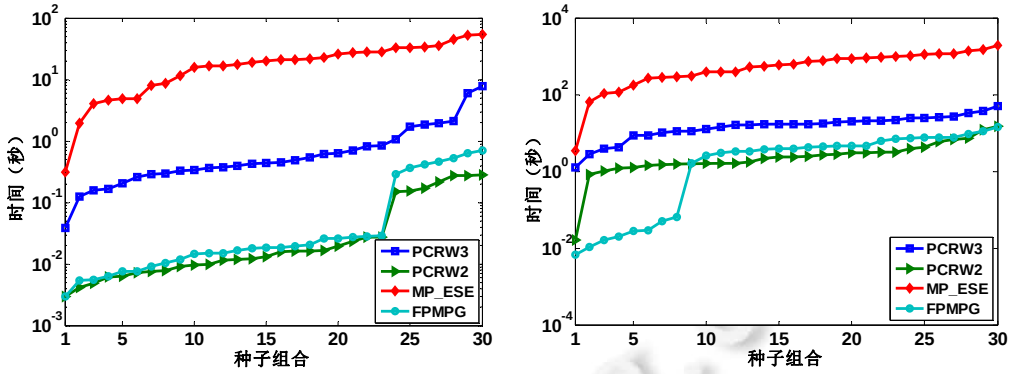
Fig.4 Running time of finding path of different methods with different seed size

图 4 不同的种子数目下采用不同方法寻找元路径的时间

从图中可以看出:随着种子数目的增加,寻找路径的时间整体上有增长的趋势,我们的方法 FPMPG 寻找路径的时间是最短的,因为 FPMPG 是基于种子实体进行路径扩展,然后进行路径连接,比其他单向的扩展方法要节省很多时间.PCRW2 也有较短的运行时间,这是因为它只找了最大长度为 2 的路径,这也导致了其不好的扩展性能.MP_ESE 方法寻找路径的时间是比较慢的,原因是它不仅采用的是单向搜索方式,而且在搜索过程中需要进行各种设定条件的判断,还需要进行剪枝等操作.

对于不同的种子组合对寻找路径时间的影响,我们也分别在 Movie*和 Scientist*任务上选取 3 个种子情况下不同的种子组合进行了 20 次实验取均值,结果如图 5 所示.从图中可以看出:在同样的种子数目下,不同的种子组合,其寻找路径的时间是不同的.可见,种子对寻找路径是有影响的.我们的方法 FPMPG 在不同的种子组合

下寻找路径的时间比 PCRW3 和 MP_ESE 方法都快.采用 PCRW2 方法寻找路径的时间比较快是因为其只找了最大长度为 2 的路径,但有效性很差.



(a) Movie* (b) Scientist*

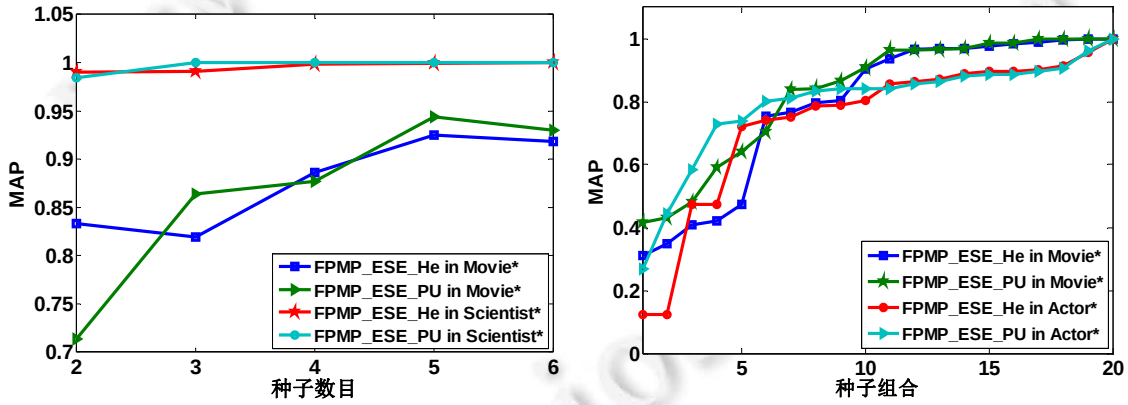
Fig.5 Running time of finding path of different methods with different seed combination

图 5 不同的种子组合下采用不同方法寻找元路径的时间

总之,种子数目和不同的种子组合对寻找路径的时间都是有影响的,因此在下一步工作中,我们可以进一步研究如何选择恰当数目的种子和最优的种子组合,进而得到最佳的寻找路径时间和最优的扩展性能.

4.6 种子个数和不同的种子组合对性能的影响

在这一小节,我们主要研究种子个数和不同的种子组合对实体集扩展性能的影响.为了研究种子个数对实体集扩展性能的影响,我们分别在 Movie*和 Scientist*任务上进行实验,从 2~6 变化种子数目,对不同种子数目,随机选择相同规模的种子进行实验 20 次取 MAP 的平均值,结果如图 6(a)所示.



(a) 种子数目对性能的影响 (b) 种子组合对性能的影响

Fig.6 Influence of seed size and different seed combinationon expansion performance

图 6 种子个数和不同的种子组合对实体集扩展性能的影响

从图中可以看出:在 Movie*任务上,随着种子数目的增加,FPMP_ESE_PU 的性能有稳定提升,FPMP_ESE_He 有一些震荡但整体性能是提升的,说明太少的种子数(如两个)包含较少的语义信息,其性能是较差的;当种子数增多时,语义信息比较丰富,性能就比较好.在 Scientist*任务上,FPMP_ESE_PU 和 FPMP_ESE_He 性能都比较好,原因可能是其语义类比较单一、明确.总之,越多的种子包含更多的信息,可以更好地表达潜在的语义,对算法

找到重要的元路径有更大的帮助;当种子数目增加到一定值时,性能趋于稳定.

为了研究不同的种子组合对性能的影响,我们分别在 Actor*和 Movie*任务上随机选择 3 个种子进行实验 20 次,取 MAP 的平均值,结果如图 6(b)所示.从图中可以看出:在两个任务上,最好和最差的性能之间有一个较大的差别.对 Actor*任务来说,最差的性能甚至不到 0.2,最好的性能接近 1.0.因此我们可以看出,不同的种子组合对结果有一个比较大的影响.可见:选择较好的种子对扩展性能是很重要的,那些劣势的种子应该被淘汰.接下来,我们将进一步研究如何选择最优的种子组合得到最佳的结果.

5 总 结

本文主要研究知识图谱中的实体集扩展问题,即:给定几个种子实体,利用知识图谱来得到更多的同类别的实体.具体地,我们把知识图谱建模成一个异质信息网络,采用元路径来探测种子实体之间潜在的共同特征.为了找到种子实体之间的重要的元路径,我们采用频繁模式挖掘技术,提出了一种新的自动寻找元路径的方法 FPMPG.FPMPG 把每个种子实体映射为一个实体事务,首先找到种子实体的频繁模式,然后连接频繁模式得到重要元路径.为了更好地组合元路径,我们设计了两种权重学习方法:一种是启发式方法,另一种是 PU learning 方法.最后,在 Yago 数据集上的实验,验证了所提方法较其他基本方法有更好的有效性以及更高的效率,并且研究了种子个数和不同的种子组合对实体集扩展性能的影响.在未来的工作中,我们将进一步研究实体集扩展问题中如何确定恰当的种子数目以及如何选取最优的种子.

References:

[1] Cohen WW, Sarawagi S. Exploiting dictionaries in namedentity extraction: Combining semi-Markov extraction processesand data integration methods. In: Proc. of the KDD. ACM Press, 2004. 89−98.

[2] Pantel P, Lin D. Discovering word senses from text. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2002. 613−619.

[3] He Y, Xin D. Seisa: Set expansion by iterative similarity aggregation. In: Proc. of the WWW. ACM Press, 2011. 427−436.

[4] Wang RC, Cohen WW. Language-Independent set expansionof named entities using the Web. In: Proc. of the ICDM. IEEE, 2007. 342−350.

[5] Wang RC, Cohen WW. Iterative set expansion of named entities using the Web. In: Proc. of the ICDM. IEEE, 2008. 1091−1096.

[6] Li XL, Zhang L, Liu B, Ng SK. Distributional similarityvs. PU learning for entity set expansion. In: Proc. of the ACL. ACL Press, 2010. 359−364.

[7] Qi ZY, Liu K, Zhao J. A novel entity set expansion method leveraging entity semantic knowledge. Journal of Chinese Informantion Processing, 2013,27(2):1−10 (in Chinese with English abstract).

[8] Sun Y, Han J, Yan X, Yu PS, Wu T. Pathsim: Meta path-based top-*k* similarity search in heterogeneous information networks. Proc. of the VLDB Endowment, 2011,4(11):992−1003.

[9] Zheng Y, Shi C, Cao X, Li X, Wu B. Entity set expansion with meta path in knowledge graph. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Cham: Springer-Verlag, 2017. 317−329.

[10] Singhal A. Introducing the knowledge graph: Things, not strings. In: Proc. of the Official Google Blog. 2012.

[11] Lenat DB. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 1995,38(11):33−38.

[12] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively createdgraph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. New York. ACM Press, 2008. 1247−1250.

[13] Suchanek FM, Kasneci G, Weikum G. YAGO: A core of semantic knowledge unifying word netand wikipedia. In: Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2007. 697−706.

[14] Dong XL, Murphy K, Gabrilovich E, Heitz G, Horn W, Lao N, Strohmann T, Sun SH, Zhang W. Knowledge vault: A Web-scale approach to probabilisticknowledge fusion. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2014. 601−610.

[15] Paulheim H, Bizer C. Type inference on noisy RDF data. In: Proc. of the Semantic Web (ISWC 2013). LNCS 8218, Berlin, Heidelberg: Springer-Verlag, 2013. 510−525.

[16] Socher R, Chen DQ, Manning CD, Ng A. Reasoning with neural tensor networks for knowledge base completion. In: Proc. of the Advances in Neural Information Processing Systems 26 (NIPS 2013). Curran Associates, Inc., 2013. 926−934.

[17] Zhao Y, Gao S, Gallinari P, Guo J. Knowledgebase completion by learning pairwise-interaction differentiate dembeddings. Data Mining and Knowledge Discovery, 2015,29(5):1486−1504.

[18] Bryl V, Bizer C. Learning conflict resolutionstrategies for cross-language wikipedia data fusion. In: Proc. of the Companion Publication of the 23rd Int'l Conf. on World Wide Web Companion. Geneva: Int'l World Wide Web Conf. Steering Committee, 2014. 1129−1134.

[19] Paulheim H, Bizer C. Improving the qualityof linked data using statistical distributions. Int'l Journal on Semantic Web and Information Systems (IJSWIS), 2014,10(2):63−86.

[20] Zou L, Huang R, Wang H, Yu JX, He W, Zhao D. Natural language question answering over RDF: A graph datadriven approach. In: Proc. of the SIGMOD. ACM Press, 2014. 313−324.

[21] Cao X, Zheng Y, Shi C, Li J, Wu B. Link prediction in schema-rich heterogeneous information network. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Springer Int'l Publishing, 2016. 449−460.

[22] Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. Proc. of the IEEE, 2016,104(1):11−33.

[23] Sun Y, Yu Y, Han J. Ranking-Based clustering of heterogeneous information networks with star network schema. In: Proc. of the KDD. 2009. 797−806.

[24] Shi C, Li Y, Zhang J, Sun Y, Yu PS. A survey on heterogeneous information network analysis. IEEE Trans. on Knowledge and Data Engineering, 2017,29(1):17−37.

[25] Shi C, Kong X, Huang Y, Philip SY, Wu B. HeteSim: A general framework for relevance measure in heterogeneous networks. IEEE Trans. on Knowledge & Data Engineering, 2014,26(10):2479−2492.

[26] Agrawal R, Srikant R, *et al*. Fast algorithms for mining associationrules. In: Proc. of the 20th Int'l Conf. Very Large Data Bases, Vol.1215. VLDB, 1994. 487−499.

[27] Han J, Pei J, Yin Y. Mining frequent patterns withoutcandidate generation. ACM SIGMOD Record, 2000,29(2):1−12.

[28] Rakesh A, Srikant R. Mining sequential patterns. In: Proc. of the 11th Int'l Conf. on Data Engineering. IEEE, 1995.

[29] Abedjan Z, Naumann F. Improving RDF data through associationrule mining. Datenbank-Spektrum, 2013,13(2):111−120.

[30] Jiang T, Tan AH. Mining RDF metadata for generalized association rules. In: Proc. of the Int'l Conf. on Database and Expert Systems Applications. Springer-Verlag, 2006. 223−233.

[31] Pasca M. Weakly-Supervised discovery of named entities using Web search queries. In: Proc. of the CIKM. ACM Press, 2007. 683−690.

[32] Jindal P, Roth D. Learning from negative examples in setexpansion. In: Proc. of the ICDM. IEEE, 2011. 1110−1115.

[33] Yu X, Sun Y, Norick B, Mao T, Han J. User guided entitysimilarity search using meta-path selection in heterogeneous information networks. In: Proc. of the CIKM. ACM Press, 2012. 2025−2029.

[34] Metzger S, Schenkel R, Sydow M. Qbees: Query by entityexamples. In: Proc. of the CIKM. ACM Press, 2013. 1829−1832.

[35] Metzger S, Schenkel R, Sydow M. Aspect-Based similar entity search in semantic knowledge graphs with diversity-awareness and relaxation. In: Proc. of the CWI and IAT. IEEE Computer Society, 2014. 60−69.

[36] Chen J, Chen Y, Du X, Zhang X, Zhou X. Seed: A systemfor entity exploration and debugging in large-scale knowledgegraphs. In: Proc. of the ICDM. IEEE, 2016. 1350−1353.

[37] Zhang J, Tang J. Focus of the next generation search engineer: Knowledge graph. Chinese Computer Society Communication, 2013, 9(4):64−68 (in Chinese with English abstract).

[38] Zou L, Chen YG. Massive RDF data management. Chinese Computer Society Communication, 2012,8(11):32−43 (in Chinese with English abstract).

[39] Aggarwal CC, Han J. Frequent Pattern Mining. Springer-Verlag, 2014.

[40] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2008. 213−220.

[41] Shi B, Zhang ZZ, Sun L, Han XP. A probabilistic co-bootstrapping method for entity set expansion. In: Proc. of the 25th Int'l Conf. on Computational Linguistics (COLING 2014), Proc. of the Conf.: Technical Papers. Dublin, 2014. 2280−2290.

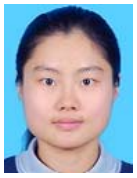
[42] Lao N, Cohen WW. Relational retrieval using a combination of path-constrained random walks. Machine Learning, 2010,81(1): 53−67.

附中文参考文献:

[7] 齐振宇,刘康,赵军.一种融合实体语义知识的实体集合扩展方法.中文信息学报,2013,27(2):1−10.

[37] 张静,唐杰.下一代搜索引擎的焦点:知识图谱.中国计算机学会通讯,2013,9(4):64−68.

[38] 邹磊,陈跃国.海量 RDF 数据管理.中国计算机学会通讯,2012,8(11):32−43.



郑玉艳(1992－),女,山东诸城人,博士生,主要研究领域为异质信息网络数据挖掘.

田莹(1996－),女,本科生,主要研究领域为数据挖掘.

石川(1978－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,机器学习,演化计算.