DOI: 10.3969/j.issn.1671-0673.2017.02.023

基于关系指示词库的开放式实体关系抽取算法

王 月[1,2],周 刚[1,2],南 煜[1,2],郑梓圣[1],田 菲[1]

(1. 信息工程大学,河南 郑州 450001; 2. 数学工程与先进计算国家重点实验室,河南 郑州 450001)

摘要: 为解决传统实体关系抽取方法适应性较差的问题,提出基于关系指示词库的开放式实体关系抽取方法(WCORE)。根据上下文环境自动构建与人工补充的方式构建关系指示词库,两种方式相互补充,主要论述 3 个问题: 关系三元组的确立; 关系指示词信息增益的计算与关系指示词库的构建; 基于关系指示词库的分类实体关系信息抽取。以真实微博文本作为语料,平均 F 值达到了 75.90%,实验结果表明该方法具有较好的可行性和适应性。

关键词: 关系三元组; 开放式实体关系抽取; 关系指示词; 指导库; 语义相似度

中图分类号: TN929.5 **文献标识码:** A **文章编号:** 1671-0673(2017)02-0242-06

Open Entity Relation Extraction Based on Library of Relation Word

WANG Yue[1,2], ZHOU Gang[1,2], NAN Yu[1,2], ZHENG Zisheng[1], TIAN Fei[1]

(1. Information Engineering University, Zhengzhou 450001, China;
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)

Abstract: Open Relation Extraction (ORE) can eschew hand-labeled training examples, tackle an unbounded number of entity relations, and scale linearly to handle large-scale web corpus. In this paper, we propose a model to extract entity relations on the basis of library of relation words (WCORE). After extracting relation triples, automatic relation word identification is one way to solve ORE problems. Thus, automatic construction and artificial supplement are combined to complement their results. This paper mainly discusses three issues: relation triples extraction; importance calculation of relation word and library building of relation word; automatic identification of relation word by using the library. Evaluations over Weibo text corpus give an average F-score of 75.90%, and the results indicate that this method could be open and adaptive.

Key words: relational triple; open entity relation extraction; relation word; library; semantic similarity

实体关系抽取的目的是发现实体并识别实体之间的语义关系[1],目前被广泛应用于本体学习、知识库构建等领域。文献[2]能够抽取 100 种关系和事件,但人工构建规则耗时耗力。文献[3]利用上下文信息对识别的实体进行聚类,抽取频率较高的语义标签作为实体集的关系,但抽取粒度还不够细。文献[4]采用领域独立的抽取模板,针对指定的关系进行抽取,能从大规模网页抽取大量关系。还有多种基于统计的机器学习方法,有效地减少了用户参与度并增强领域的适应性,但抽取关系仍然有限。

开放式关系抽取[5-6]的出现为应对大规模文本提供了新的研究范式: 利用语言自身完备性,抽取大量实体关系,从而避免了构建关系类型体系。文献[7]提出 TextRunner 系统,利用启发式规则从宾州树库中自动构建语料,然后训练模型识别关系

收稿日期: 2016-03-08; 修回日期: 2016-04-14

作者简介: 王 月(1988-),女,硕士生,主要研究方向为自然语言处理、知识图谱。

三元组。文献 [8] 提出 WOE 系统,使用 Wikipedia 信息框的内容来标注语料,有效提高了训练语料的数量与质量。文献 [9] 对 TextRunner 与 WOE 系统的抽取结果进行分析,提出了先识别关系指示词的 ReVerb 系统。但这些系统有两大缺陷: 仅抽取以动词为核心的关系和忽略上下文全局信息。文献 [10] 利用学习的开放式模板和依存分析很好地解决了以上问题。

与英文在关系抽取方面取得的进展相比,中文在这方面的研究还很少。文献 [11] 提出了无指导的开放式中文关系抽取方法,结合实体之间的距离限制和关系指示词的位置限制获取候选三元组,利用启发式规则过滤候选三元组,达到了 80% 以上的微观平均准确率,但其召回率相对较低。文献 [12] 提出基于语义的开放式中文关系抽取方法(ZORE: A Syntax-based System for Chinese Open Relation Extraction) ,通过增加双向扩展标注,将语义模式与关系抽取结合,取得了不错的效果。

在研究开放式中文实体关系抽取时,遇到了以下问题: ①候选实体对方面,两个互不相关的命名实体往往根据规则组成候选实体对,因此任何关系指示词都无法正确表达候选实体对之间的关系,这种情况难以判断; ②关系指示词方面,实体对的候选关系指示词可能有多个,而正确的关系指示词可能只有一个,或者没有,又或者它的权重并不高,而关系指示词的权重计算有多种方法,因此合理的关系指示词权重设置十分重要,而且对关系指示词是否正确的自动识别也非常关键。

本文提出基于关系指示词库的开放式实体关系抽取算法(open entity relation extraction based on library of relation word, WCORE) ,该方法可有效解决关系指示词的权重计算问题,同时也能够避免因权重低而漏选的关系指示词,实验表明该方法具有较强的可行性和适应性。

1　基于关系指示词库的开放式关系抽取算法

如图 1 所示,WCORE 算法共包含 5 个模块: 预处理、候选三元组提取、GCORE(gravity model for chinese open entity relation extraction) 信息增益值计算、关系指示词库构建、关系指示词自动识别。

1.1　GCORE 原理

开放式中文实体关系抽取可以理解为是对候选三元组的抽取及它们之间是否存在可靠关系的一种判断,即判断候选实体对之间、候选实体对与关系指示词之间是否存在可靠的关系。GCORE 是判断候选三元组可靠性的一种方法,认为候选三元组的可靠性受 3 方面的影响: ①候选实体对的词频。上下文中候选实体对的词频越高,说明该候选实体对常常一起出现,越有可能存在某种关系; ②关系指示词的词频。上下文中某个关系指示词的词频越高,说明该词是一种常用词,越有可能作为某个实体对之间正确关系的表达; ③距离的影响指的是关系指示词与候选实体对之间在文本中间隔的字词数量,它们之间的距离越近,说明该关系指示词与候选实体对之间的关系越密切。因此,某个关系指示词能否作为正确的关系指示词,与实体对词频、关系指示词词频及它们之间的距离息息相关。

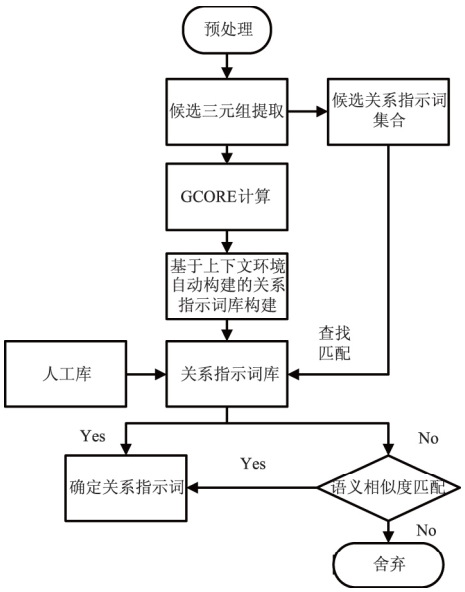

图 1　WCORE 算法框图

万有引力定律是解释物体间相互作用的引力的定律,GCORE 正是借鉴了该思想,认为候选三元组的可靠性可以用候选三元组间的相互作用力来表示,相互作用力的大小反映了可靠性的强弱,从而合理描述候选三元组的可靠性,具体是这样理解的: 候选三元组内部存在一种相互作用力,该力的大小正比于关系指示词词频与候选实体对词频的乘积,从而很好地反映出实体对词频与关系指示词词频对候选实体对可靠性的影响; 反比于关系指示词和候选实体对之间距离的平方,从而很好地反映关系指示词与实体对之间关系热度在距离上存在一种非线性关系。

假设某候选三元组共提取出 M 个候选关系指示词,则该候选三元组的可靠性值 W 计算:

$$W_k = f(e_i, e_j) * f(r_k) / d_{ki}^2 + d_{kj}^2), k \in M \quad (1)$$

$$W_{\max} = \max(W_k) \quad (2)$$

其中，r_k 是命名实体对 (e_i, e_j) 的词频，e_j 是关系指示词 W_k 的词频，e_i，e_j 与 r_k 分别是关系指示词 r_k 到命名实体 e_i 和 e_j 的距离，W_k 是命名实体对 e_i，e_j 与关系指示词 r_k 组成的候选三元组的可靠性分值，而 $W_{\max}$ 是 M 个候选三元组中可靠性得分最高的一组，即可靠性得分为 $W_{\max}$ 的候选三元组是最有可能正确的表达候选命名实体对之间、候选命名实体对与关系指示词之间某种关系的最佳三元组。

GCORE 将实体对词频、关系指示词词频及两者间的距离有机地结合起来，能够简单直接的表达候选三元组的关系强度，也作为一种有效的计算关系指示词信息增益的方法。

1.2 关系指示词库构建原理

在关系抽取过程中，合理的关系指示词权重设置十分关键。然而经过权重设置并通过筛选的关系指示词也并不一定完全正确，经过多次研究发现，某些通过筛选的关系指示词往往并不具有关系指示作用，这就无法表示实体对的关系，因此，如何自动识别指示词是否具有关系指示作用成为一个值得研究的问题。

WCORE 利用文本上下文环境自动构建与人工自定义的方式构建关系指示词库，对候选关系指示词进行分析，以确定其是否具有关系指示作用。基于上下文环境自动构建的关系指示词库，是通过对全文中的所有关系指示词以 GCORE 信息增益的方式计算其权重值，权重值高于算法设定的某个阈值时，则认为这些关系指示词可以表达实体对的关系指示作用，将其中权重较高的提取出来作为标准关系指示词，它的作用是设立一个标准，认为只有这些标准关系指示词才具有关系指示作用，通过将实体对的候选关系指示词与标准关系指示词做对比分析，即可确定候选关系指示词是否具有关系指示作用。

此外，人工自定义的常用关系指示词库定义了各种类型的常用关系指示词，作为自动构建库的补充，可为没有选入自动构建库但却具有一定关系指示作用的关系指示词的有效补充。基于上下文环境的关系指示词自动构建库与人工自定义常用关系指示词库的结合，构成了一个既适用于各个领域又不失全面的关系指示词库。

1.3 关系指示词库的构建方法

将实体对关系类型划分 5 类：人名 - 人名命名实体的关系；人名-机构命名实体的关系；人名 - 地点命名实体的关系；机构 - 机构命名实体的关系；因为地名 - 地名和地名 - 机构两种实体关系的关系指示词很相近，都是和位置相关，而且数目相对较少，因此将两种实体关系合二为一，重新定义为位置关系类型。自动构建库在构建过程中，要按照以上 5 种关系类型对实体对关系分类，并在类别内分别择优收录候选关系指示词，这样可以避免因某种实体对关系类型的关系指示词整体数量较少而发生遗漏的情况，达到类别公平的目的。人工自定义常用关系指示词库同样也是按照 5 种类型收录关系指示词。

1.4 语义相似度计算

由关系指示词语义得到的特征向量可以作为关系指示词之间相关性的计算，如关系指示词“女儿”与“闺女”、“外婆”与“姥姥”等同义词，它们的语义相似度非常高，因此，基于语义特征的关系指示词匹配方法能很大程度上提供关系指示词匹配的准确率和泛化能力，适用于 WCORE 中的关系指示词语义匹配分析，可实现利用关系指示词库中有限个数的标准关系指示词来分析大量数目的候选关系指示词集合的目的。

本文采用文献 [13] 提出的方法，利用基于 HowNet 的相似度计算，认为一个词可由一个或多个概念组成，概念则是由义原描述。那么，词语间的相似度就可以用表示 2 个词语的概念之间的相似度来衡量，而概念的相似度则用义原之间的相似度来体现。通过 HowNet 中的“义原”，可以计算不同词之间的语义相似度。

对于 W_1 和 W_2，如果 W_1 有 n 个概念：S_{11}，S_{12}，…，S_{1n}，W_2 有 m 个概念：S_{21}，S_{22}，…，S_{2m}，规定 W_1 和 W_2 的相似度为各个概念的相似度最大值：

$$Sim(W_1, W_2) = \max_{i=1,\cdots,n, j=1,\cdots,n} Sim(S_{1i}, S_{2j}) \quad (3)$$

式中，Sim 表示相似度，W_1 和 W_2 分别表示两个汉语词语；S_1、S_2 表示两个不同的义原。

这样词语语义相似度的问题转移到了概念相似度的问题上，又由于所有的概念都是由义原描述的，所以概念相似度最终转移到了计算义原相似度上。在对候选关系指示词进行基于关系指示词库的自动识别时，使用词语语义相似度作为关系指示词之间的相似度。

1.5 WCORE 模型

在抽取实体关系过程中，候选三元组按照 2.3 节所述 5 种关系类型进行分类，权重最高的候选关系指示词往往是正确的关系指示词，因此，将候选关系指示词按权重值降序排列，并与关系指示词库

中的同类别的标准关系指示词依次比较分析,以判断该候选关系指示词是否具有关系指示作用。

①预处理。采用 NLPIR 汉语分词系统[14] 对文本进行断句、分词、词性标注、命名实体识别等。

②候选三元组提取。候选三元组提取范围是在一个自然句中,即不跨句号、问号、分号和叹号。

提取候选命名实体对　将文本中两两相邻命名实体构成一对候选命名实体对,候选命名实体对中间不包含其它命名实体或命名实体对。

提取候选关系指示词　关系指示词是表达两个实体之间关系的一个或一组词语,该词的词性往往是名词或动词。本算法抽取关系指示词的规则如下: 关系指示词词性为名词或动词; 关系指示词位于候选命名实体对中间和左右两侧,左右两侧边界范围是以本候选实体对至另外候选命名实体边界为界,以及句首或者句末; 本算法对停用词做了处理,一种是由中科院收集的停用词表,共 1208 个停用词; 另外一种是用户自定义的停用词(如"给"、"据说"、"还有"等词),以上这些停用词都符合关系指示词抽取的条件,但很明显这些词不适合做关系指示词;

③GCORE 信息增益计算。候选三元组提取后,每一个候选实体对都会对应一个候选关系指示词集合,该集合可能会有零到多个关系指示词,实体对会和每个关系指示词组成一个候选三元组。GCORE 计算就是要计算该集合中每个关系指示词的权重值,即关系指示词词频统计; 实体对词频统计; 根据(1)式和(2)式,以每个候选三元组为单位,分别计算候选关系指示词集合中每个关系指示词的权重值,并按分值大小降序排列;

④关系指示词库构建。包括基于上下文环境自动构建的关系指示词库和人工库,其中前者选取每类实体关系类型中 GCORE 值较高的关系指示词自动构建;

⑤基于关系指示词库的关系指示词自动识别。该模块分为两步: 首先将候选关系指示词按 GCORE 值由高到低依次在关系指示词库中快速查找匹配,如果查找到,则认为将该候选关系指示词确定为正确的关系指示词并结束后续查找; 如果未找到,则进行语义相似度匹配,当语义相似度大于系统设定的阈值后,就认定该候选关系指示词为正确的关系指示词并结束后续计算。如果最终计算结果没有符合条件的候选关系指示词,则返回空。

其关系指示词自动识别算法伪代码如下:

①Begin

②Read 候选关系指示词集合 ListforCandidateRelationWords、关系指示词库 List for Relation Word Data Base

③按照权重大小降序排序 ListforCandidateRelationWords

④初始化变量 $i \leftarrow 0$

⑤初始化变量 Flag ← false

⑥初始化变量 $R \leftarrow$ NULL

⑦初始化关系指示词变量 RelationWord ← NULL

⑧初始化语义相似度计算值 SimilarityCalculationValue ← 0

⑨初始化语义相似度阈值 ThresholdValue 为系统设定的默认值

⑩While RelationWord = = NULL

⑪$R \leftarrow$ ListforCandidateRelationWords [i]

⑫Flag ←查找 R 是否存在于 ListforRelationWordDataBase

⑬If Flag = = true then

⑭RelationWord ← R

⑮Break

⑯Else

⑰For j from 0 to ListforRelationWordDataBase. length-1 step 1

⑱SimilarityCalculationValue ←语义相似度计算 R 与 ListforRelationWordDataBase [j]

⑲If SimilarityCalculationValue > = ThresholdValue then

⑳ RelationWord ← R

㉑ Break

㉒ End if

㉓End for

㉔ End if

㉕ i + +

㉖ End while

㉗ Return RelationWord

㉘ End

设候选关系指示词个数为 n,同类别的标准关系指示词个数为 m,排序方法实现了 Java 类的 Comparator 接口,重载了 compare 函数,是一种归并排序,其时间复杂度为 $O(n\log_2 n)$; 查找与语义相似度计算的时间复杂度都为 $O(nm)$,因为 $n < m$,则算法的时间复杂度为 $O(n(m+\log_2 n))$。

2　实验结果与分析

2.1　数据及实验设计

本次实验分为两组,第 1 组数据采用 ZORE 系统[12] 中的数据集进行对比实验,目的是验证 WCORE 算法的可行性和有效性; 第 2 组数据使用新浪微博文本,目的是验证 WCORE 算法在开放式

网络文本中的适应性。

2.2 实验结果及评价

在第 1 组数据中，新闻文本中实体关系共 69 组，被抽取出 163 组，关系指示词 2139 个；维基百科文本中实体关系共 53 组，被抽取出 140 组，关系指示词 1849 个。

图 2 和图 3 展示了 WCORE 与 ZORE 的性能对比情况，WCORE 总体性能要优于 ZORE，相同准确率下，WCORE 的召回率都要优于 ZORE。图 2 中，当准确率为 80% 时，ZORE 的召回率为 25%，而 WCORE 的召回率为 35%；图 3 中，当准确率为 80% 时，ZORE 的召回率为 20%，而 WCORE 的召回率为 37.5%。同样，在相同召回率的情况下，WCORE 的准确率也是优于 ZORE。结果初步表明，WCORE 具有一定的可行性和有效性。

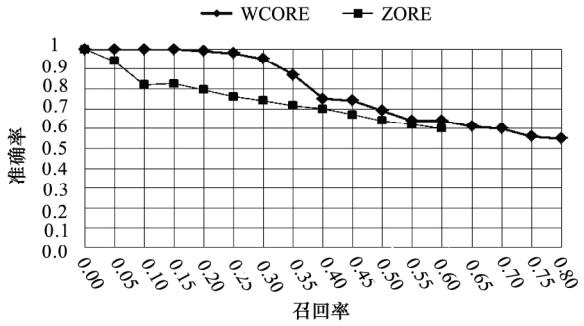

图 2 性能对比图(新闻文本)

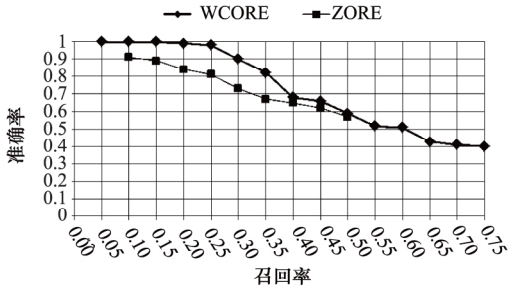

图 3 性能对比图(维基百科)

在第 2 组实验中，从新浪微博中爬取 100 万条数据，为了便于准确统计实验结果，随机抽取包含人名 - 人名、人名 - 机构、人名 - 地名、机构 - 机构、位置关系等 5 类关系类型的数据 1 万条，共抽取实体对 14878 对。

关系指示词库分为人工库和自动库，人工库示例如表 1，自动库示例如表 2(以 GCORE 值前 10% 择优构建)。

表 1 人工库构建的关系指示词词表部分样例

实体对类型	关系指示词词表
人名 - 人名	妻子、同学、男友、室友、饰演
人名 - 机构	主任、校花、主持人、导演、董事长
人名 - 地名	市长、首富、访问、生于、定居
机构 - 机构	合并、投资、签约
位置关系	首都、吞并、省会、位于、保护区

表 2 自动库构建的关系指示词词表部分样例

实体对类型	关系指示词词表
人名 - 人名	合作、求婚、夫妇、男友、女儿
人名 - 机构	加盟、秘书长、解约、CEO、亮相
人名 - 地名	总统、主席、书记、访问、歌手
机构 - 机构	合并、合作、联手、打造
位置关系	首都、冲突、合作、支持、参加

表 3 是实验结果中部分关系三元组样例及其结果分析。其中，关系三元组“郭晓冬 夫妻 程莉莎”中，关系指示词“夫妻”因为 GCORE 值低而未能被收录到自动库中，而该词是一种常用词，可以表示人名 - 人名的实体关系，因此可以被收录到人工库中，对自动库做有效的补充。人工库中收录的大部分关系指示词都是一些常见但在抽取文本的上下文环境中，词频却不一定高的词汇。

在表 3 中出现的关系指示词，如果在“候选关系指示词集合及 GCORE 值”一列中也同样出现而且 GCORE 值也相对较高，一般是择优被收录到自动库中，如“董事长”、“联手”等，否则就是被提前收录到人工库中，如“搭档”、“定居”等词。比较“库中关系指示词”和“抽取结果”两列的内容可以

表 3 部分关系三元组抽取结果样例

实体类型	实体对	候选关系指示词集合及 GCORE 值	库中关系指示词	抽取结果
人名 - 人名	郭晓冬 程莉莎	夫妻(0.461)，情侣(0.048)，想想(0.011)	夫妻	夫妻
	汪涵 马可	合作(8.117)，搭档(0.800)，加盟(0.453)	搭档	合作
人名 - 机构	马化腾 腾讯	董事长(7526.400)，CEO(4166.400)，参与(18.823)	董事长	董事长
	南京大学 鹿化煜	教授(1.000)，选派(0.200)，IODP(0.153)	教授	教授
人名 - 地名	乐仁波 西雅图	旅居(0.117)	定居	旅居
	徐国平 长兴	调研(0.653)，厅长(0.153)，省水利厅(0.034)	访问	调研
机构 - 机构	腾讯 富士康	联手(15924.670)，打造(1911.323)，合作(289.800)	联手	联手
	苹果 三星	联合(2.250)，希望(0.730)，对抗(0.100)	联手	联合
位置关系	俄罗斯 莫斯科	首都(4.200)，举行(0.577)，红场(0.482)	首都	首都
	阿里巴巴 以色列	投资(2.300)，JVP5(0.017)	投资	投资

发现,两列中关系指示词相同的情况是因为后者已经被收录到关系指示词库中,通过在库中快速查找被正确识别出来,而两列中关系指示词不相同的情况,如“定居”与“旅居”等,是因为后者没有被收录到关系指示词库中,但是关系指示词库中有语义相似的关系指示词,通过语义匹配分析被正确识别出来,这几对关系指示词的语义相似度都非常高,实际计算结果为 1.0。

除了表 3 中数据之外,还有更多的关系指示词如“老师”与“师傅”、“主演”与“饰演”、等,都可以通过语义相似度计算将它们准确的识别出来,凭借语义相似度计算的可行性,就可以通过语义分析去重复的方法来达到优化关系指示词库结构的目的,使关系指示词词库精简而不臃肿,在提高关系指示词匹配的泛化能力和准确率的同时,增强了算法的可行性、有效性和计算效率。

表 4 是对本次实验结果的统计与评价分析。从评价指标上看,实体对类型为人名 - 人名与人名 - 机构两种类型的提取效果比较好,精度、召回率及 F 值都处在一个较高的水平; 而其它 3 种实体对类型的提取效果却不太理想,3 个评价指标的值相对较低; 总体上看本次实验的效果达到了一个较为不错的水平,分析发现人名 - 人名命名实体与人名 - 机构命名实体两种类型的关系指示词种类多数量也大,而另外 3 种实体对类型的关系指示词,尤其是机构 - 机构、位置关系的两种类型的关系指示词种类和数目都相对较少,造成提取困难。

表 4　在微博文本上的关系三元组抽取结果

实体对类型	抽取三元组数量	评价指标%		
		Precision	Recall	F-value
人名 - 人名	4047	85.40	88.20	86.78
人名 - 机构	3146	89.06	85.91	87.46
人名 - 地名	3925	67.71	85.52	75.58
机构 - 机构	1046	61.80	63.83	62.80
位置关系	2714	63.02	87.36	66.87
均值		73.40	82.16	75.90

2.3　错误分析

本文算法不依赖于句法结构,对于关系抽取的限制条件也比较宽松,能够挖掘出文本中的大部分实体关系,得到较高的召回率,同时也抽取了一些错误的实体关系,影响本次实验效果的原因:

①微博口语化严重。如“百度李彦宏、腾讯马化滕分列 2、3 名”一句中,候选命名实体对“百度 李彦宏”没有关系指示词,而候选三元组“腾讯 分列 马化腾”中关系指示词明显是错误的;

②多组实体关系存在交叉或重叠的情况,如“中国富豪榜前十位的还有小米董事长雷军和京东董事长刘强东”,可以抽取出“小米 董事长 雷军”、“京东 董事长 刘强东”与“雷军 董事长 京东”,最后一组明显是错误的,这种情况难以仅通过关系指示词对实体对是否存在正确的关系做出准确的判断;

③机构 - 机构、位置关系两种实体类型的关系指示词种类和数目相对较少,造成提取困难,错误较多。

3　结束语

本文提出的 WCORE 方法,借鉴空间万有引力思想,将候选命名实体对词频、关系指示词词频以及候选命名实体对与关系指示词之间的距离三者有机的结合在一起,表达候选三元组的关系强度,作为一种有效的计算关系指示词信息增益的方法,并建立基于上下文环境自动构建的关系指示词库,最后结合人工常用关系指示词库,利用语义相似度计算方法实现实体关系抽取的目的,为开放式中文关系抽取方法提供了一种新的思路。

WCORE 结合具有挑战性的微博文本进行了实验,验证了方法的可行性及适应性,同时也发现一些问题: ①高词频的候选实体对不一定存在实体关系,如多组实体关系存在交叉或重叠的情况; 同样,低词频的候选实体对也不一定不存在实体关系,人工常用关系指示词库可以对后者做一定的改善,但对于前者目前还没有具体可行的办法; ②算法无法解决所有问题,词库也非常重要,微博文本的特点更加突出了词库的重要性,一个全面的、精准的词典在一定程度上影响了实体关系抽取的效果。

参考文献:

[1] Cutts B B, White D D, Kinzig A P. Open Information Extraction: The Second Generation. [C] // Twenty-second International Joint Conference on Artificial Intelligence. 2011: 3-10.

[2] Aone C, Ramos-Santacruz M. REES: a large-scale relation and event extraction system [C] //6th Conference on Applied Natural Language Processing. 2000: 18-22.

[3] Hasegawa T, Sekine S, Grishman R. Discovering Relations among Named Entities from Large Corpora [C] // Proc. Annual Meeting of Association of Computational Linguistics (ACL 04) . 2004: 415-422.

(下转第 252 页)

4 结论

本文通过分析在大数据背景下对大规模复杂网络进行可视化布局的需求与难点，针对性地提出了一种通过层次划分确定网络整体结构，对需展示部分进行逐级可视化的逐进式层次可视化方法。通过引入社团发现的方法对大规模复杂网络进行可视化布局可以大幅减少布局时间，同时可以清晰明确地展示出大规模复杂网络内部的层次关系。通过引入 *N*-body 模型实现最终的三维空间的可视化结构，一方面比采用传统力导引布局方法减少了布局时间，另一方面通过三维空间网络结构，使用者可以自由改变观察的角度和层次范围，提高了使用者的交互方式和认知效率，并通过实验证明了本文方法的有效性。

参考文献:

[1] 程学旗，靳小龙，王元卓. 大数据系统和分析技术综述 [J]. Journal of Software，2014，25(9)：1889-1908.

[2] 程学旗，沈华伟. 复杂网络的社区结构 [J]. 复杂系统与复杂性科学，2011，8(1)：57-70.

[3] Newman M E J，Girvan M. Finding and evaluating community structure in networks [J]. Physical review E，2004，69(2)：1-15.

[4] 汪小帆，刘亚冰. 复杂网络中的社团结构算法综述 [J]. 电子科技大学学报，2009，38(5)：537-543.

[5] Blondel V D，Guillaume J L，Lambiotte R，et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics：Theory and Experiment，2008，2008(10)：1-12.

[6] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical review E，2004，69(6)：1-5.

[7] Fruchterman T M J，Reingold E M. Graph drawing by force-directed placement [J]. Softw.，Pract. Exper.，1991，21(11)：1129-1164.

[8] Kamada T，Kawai S. An algorithm for drawing general undirected graphs [J]. Information processing letters，1989，31(1)：7-15.

[9] 王伟，曾栩鸿，王福焕. 并行时空处理模型下的快速 *N*-body 算法 [J]. simulation，2011，5(11)：1006-1013.

[10] Di Battista G，Eades P，Tamassia R，et al. Algorithms for drawing graphs：an annotated bibliography [J]. Computational Geometry，1994，4(5)：235-282.

[11] 汪小帆. 复杂网络中的社团结构分析算法研究综述 [J]. 复杂系统与复杂性科学，2008(3)：1-12.

[12] Walshaw C. A multilevel algorithm for force-directed graph-drawing [J]. J. Graph Algorithms Appl.，2003，7(3)：253-285.

[13] 任磊，杜一，马帅. 大数据可视分析综述 [J]. Journal of Software，2014，25(9)：1909-1936.

[14] Fischer F，Fuchs J，Mansmann F，et al. Banksafe：Visual analytics for big data in large-scale computer networks [J]. Information Visualization，2015，14(1)：51-61.

(上接第 247 页)

[4] Tseng Y H，Lee L H，Lin S Y，et al. Chinese Open Relation Extraction for Knowledge Acquisition [C] // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014：12-16.

[5] Zhu J，Nie Z，Liu X，et al. StatSnowball：a statistical approach to extracting entity relationships [C] // Proceedings of the 18th international conference on World wide web. 2009：101-110.

[6] Riedel S，Yao L，Mccallum A. Modeling Relations and Their Mentions without Labeled Text [C] // Machine Learning and Knowledge Discovery in Databases，European Conference. 2010：148-163.

[7] Etzioni O，Banko M，Soderland S，et al. Open information extraction from the web [J]. Advanced Pharmaceutical Bulletin，2010，5(4)：2670-2676.

[8] Wu F，Weld D S. Open information extraction using Wikipedia [C] //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010：118-127.

[9] Anthony Fader，Stephen Soderland，Oren Etzioni. Identifying relations for open information extraction [C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing，Association for Computational Linguistics. 2011：1535-1545.

[10] Etzioni O，Bart R E，Mausam N，et al. Open Language Learning for Information Extraction，US20140156264 [P]. 2014.

[11] 秦兵，刘安安，刘挺. 无指导的中文开放式实体关系抽取 [J]. 计算机研究与发展，2015，52(5)：1029-1035.

[12] Qiu Likun，Zhang Yue. A Syntax-based System for Chinese Open Relation Extraction [C] //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014：1870-1880.

[13] 刘群，李素建. 基于《知网》的词汇语义相似度的计算 [C] //第 3 届汉语词汇语义学研讨会. 2002：59-76.

[14] 张华平. NLPIR 汉语分词系统 [EB/OL]. [2016-04-04]. http：//ictclas. nlpir. org/.