基于网络语义标签的多源知识库实体对齐算法

王雪鹏 刘 康 何世柱 刘树林 张元哲 赵 军

(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

摘 要 知识库是多种自然语言处理任务的重要数据资源,但单一知识库覆盖度低,不同知识库异构性强,不利于数据的共享和集成.因此,多源知识库融合技术的研究有着十分重要的意义.其中,多源知识库实体对齐是多源知识库融合技术中的重要组成部分.在语义万维网发展的推动下,国外开展了很多相关工作,大多适用于英文知识库,对于中文知识库的研究较少.出于对中文知识库融合的研究目的,该文提出了一种基于网络语义标签的多源知识库实体对齐算法.该算法综合利用属性标签、类别标签和非结构化文本关键词,对齐中文百科实体.经实验测试,该算法能够较好地解决多源知识库实体对齐问题,算法在近 95% 的准确率下,仍能保持近 55% 的较好的召回率,应用于实际系统中,满足了实际的多源知识库实体对齐应用需求.

关键词 语义标签;多源知识库;实体对齐;异构;实体歧义

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2017.00701

Multi-Source Knowledge Bases Entity Alignment by Leveraging Semantic Tags

WANG Xue-Peng LIU Kang HE Shi-Zhu LIU Shu-Lin
ZHANG Yuan-Zhe ZHAO Jun

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract Knowledge base is an essential data source in many natural language processing tasks. But the coverage of the uni-source knowledge base is so narrow. Moreover, the hierarchies of different knowledge bases are also different. So, there are much of difficulties in data sharing and integrating between different knowledge bases. Hence, the investigation on multi-source knowledge bases alignment turns to be much of significance. And multi-source knowledge bases entity alignment is an important component in multi-source knowledge bases aligning techniques. Driven by the development of the Semantic Web, there emerge numerous investigations on knowledge bases alignment among foreign researchers; most of them focus on the knowledge bases in English. But there are fewer similar works on the knowledge bases in Chinese. To explore the knowledge bases in Chinese, we proposed a kind of multi-source knowledge bases entity aligning method by leveraging the semantic tags. This method utilized attribute triples, category tags and key words from the unstructured text synthetically to align entities which are from Chinese encyclopedias. The experiments showed that our method makes an effective performance in solving the problem of knowledge bases entity alignment. It renders a 95% accuracy and a 55% recall at the same

收稿日期:2014-08-28;在线出版日期:2016-05-08.本课题得到国家自然科学基金项目(61533018)、国家"九七三"重点基础研究发展规划项目基金(2014CB340503)和"CCF-腾讯"犀牛鸟基金资助.**王雪鹏**,男,1988 年生,博士研究生,主要研究方向为本体映射、垃圾评论检测.E-mail:xpwang@nlpr.ia.ac.cn.**刘 康**(通信作者),男,1981 年生,博士,副研究员,主要研究方向为情感分析、问答系统、信息抽取.E-mail:kliu@nlpr.ia.ac.cn.**何世柱**,男,1987 年生,博士,助理研究员,主要研究方向为信息抽取与问答系统.**刘树林**,男,1989 年生,博士研究生,主要研究方向为事件抽取与事件预测.**张元哲**,男,1986 年生,博士研究生,主要研究方向为本体映射、异构数据问答.**赵 军**,男,1969 年生,博士,研究员,博士生导师,主要研究领域为信息检索、网络挖掘、问答系统、自然语言处理.

time. Our method works well in the utility system and satisfies the actual application requirements of the entity aliment in the multi-source knowledge bases.

Keywords semantic tags; multi-source knowledge bases; entity alignment; heterogeneous; entity ambiguity

1 引 言

从网络中挖掘有用的知识,构建结构化知识库,不仅对语义万维网的发展,而且对文本内容理解都具有重要的支撑作用,因此受到工业界和学术界的高度关注.目前互联网上已经出现了大量的知识资源和知识社区,例如 Wikipedia、百度百科、互动百科等.从这些知识资源中,已有研究已经挖掘出以实体、实体关系为核心的大规模知识库,例如 Yago[1-2]、DBpedia[3-4]、Freebase[5] 等.但是,仅仅使用单一知识库具有信息覆盖度低、信息描述不完全的缺点,随着这些知识库规模的不断扩大,不同知识库之间的异构问题日益显著,极大地阻碍了数据的共享和集成.因此,研究多源知识库融合技术,整合已有知识库资源便显得十分重要,对于下一代搜索技术、文本语义理解技术具有十分重要的意义.

知识库间的异构问题主要有两种:(1)体系差异.不同知识库的知识描述体系是不同的;(2)内容差异.在不同知识库中所填充的实体是不同的.例如,在互动百科中,有 10 个实体名称为"李娜"的条目,而在百度百科中,有 22 个实体名称为"李娜"的条目.这些条目中有些是可以对齐在一起的,有些则不能.已有工作多集中在第 1 点(体系差异)上,例如 OAEI[6-7]、Yago[1-2] 等,针对第 2 点中文知识库内容对齐方面,已有工作不多.现有的英文知识库一般都具有较完整的体系结构和丰富的类与实体信息.现有的技术[8-9]也大多依赖于知识库的体系结构来计算类的依存关系以及知识库中挂载的实体的相似度.相比之下中文知识库资源缺乏完整的体系结构,实体也缺乏丰富的描述信息.大多适用于英文知识库的融合技术方法,并不适合直接应用于中文知识库.因此,本文着重针对第 2 点(内容差异),即如何发现异构知识库中实体条目间的对应关系,给出了一种基于网络语义标签的中文知识库实体对齐算法.该算法充分利用描述实体条目的多种网络语义标签(类别标签、属性标签、非结构化文本关键词等),构建多个语义相似度计算模型,弥补中文知识库缺乏完整体系结构的不足,在此基础上挖掘异构知识库中条目间的"SameAs"关系,完成异构知识库间的实体对齐.通过对百度百科、互动百科上多个领域(人物、图书、电影、其他)中具有同一名称的实体条目的对齐实验,表明该方法的有效性.同时,我们也说明融合多种语义标签相对于单一语义标签能够有效地提高异构实体对齐的性能.

本文主要贡献如下:

(1)融合多源知识库资源,建立了一个体系完整、数据丰富、易扩展的中文知识库资源;

(2)弥补了中文知识库体系结构不完整的不足,提出了一个基于网络语义标签的多源知识库实体对齐算法,且获得了较好的应用效果.

2 相关工作

在语义万维网发展的推动下,国外开展了很多关于多源知识库融合的工作,例如:

(1)从知识库体系差异角度出发的工作,除上文所述 OAEI[6-7]、Yago[1-2] 外,Parundekar 等人[10]和 Jain 等人[11]针对 Linked Open Data 数据也做了出色的工作;Dieng 和 Hug[12] 通过语言学技术与比较父类子类的方法来匹配知识库的体系结构;Maedche 和 Staab[13] 通过比较每个类的父类子类标签来计算两个知识库体系的差异度;Madhaven 等人[14]考虑了知识库中术语和数据类别信息来进行非循环结构的研究.

(2)从知识库内容差异角度出发,Raimond 等人[15]提出了一种基于网页相似度和相邻网页相似度的算法,来寻找条目间的"SameAs"关系,但适用于小规模的数据集;Nikolov 等人[16] 提出的 KnoFuss 体系结构含有数据对齐工作,不过需要所处理的数据都来自一致使用 OWL 表示的本体;Volz 等人[17]提出的 Silk 采用了索引的方式来管理数据资源,大大地降低了计算条目间"SameAs"关系的时间复杂度.FCA-Merge 系统[18] 使用形式概念分析法来融合拥有共同实体集的两个知识库,但忽略了类的属性信息;T-tree 系统[8] 研究了拥有共同实体集的不同

知识库中的类依存关系；Giunchiglia 和 Shvaiko[9] 通过给定初始类等价关系以及知识库体系关系分析来发现类的等价或蕴含关系.

根据现有的一些文献综述[19-23]的总结，现有的技术都是基于计算知识库体系的相似度，或是基于利用多类型信息（如实体名称、体系结构关系）来计算不同知识库实体间的相似度，或是两者的结合，从而达到知识库融合的目的. 现有的英文知识库一般都具有较完整的体系结构和丰富的类与实体信息. 现有的技术也大多依赖于知识库的体系结构来计算类的依存关系以及知识库中挂载的实体的相似度. 相比之下中文知识库资源缺乏完整的体系结构，实体也缺乏丰富的描述信息. 例如实体信息较丰富的百度百科缺乏明确的分类体系，互动百科虽有分类树，但相对于英文知识库，体系不够完整与成熟. 当我们将百度百科与互动百科的实体进行对齐时，没有可参照的体系结构信息来计算实体的相似度，只能依赖于实体本身的基本信息，如实体名称等，而已有众多工作证明此种方法不够有效. 因此，大多适用于英文知识库的融合技术方法，并不适合直接应用于中文知识库. 同时，对于中文知识库的研究工作较少. 关于中文知识库内容差异，陈珂锐等人[24]在所提出的 AVP 平台中，只利用了百科词条中的属性值对作为特征模板，辅助于属性值共现频率，利用扩展向量空间模型对词条进行歧义识别，未解决体系差异问题；Niu 等人[25]提出的 Zhishi.me 是首份关于中文 Linked Open Data 的工作，Zhishi.me 主要是利用原始网页中的页面重定位信息及实体名称归一化，对百科实体进行对齐，其对齐表现主要依赖于百科知识库原始网页的消歧信息，该工作只利用了实体名称方面的信息，未对实体对齐工作进一步深入研究.

3　问题描述

本文面临的任务是多源知识库的实体对齐. 在构建大规模知识库的任务中，需要处理大量来自多源知识库的实体数据. 在构建知识库之初，首先需建立一个知识描述体系，然后向体系中挂载实体数据. 由于不同知识库的信息来源不同，以及人工定义及校对的差异，语义上相同的实体在不同的知识库中会有不同的表现形式. 具有相同条目名称的实体也许表示着语义上的同一事物，也许表示着两种事物. 如具有以“李娜”作为条目名称的实体有多个，它们分别表示着网球运动员李娜、跳水运动员李娜、演员李娜等等. 当我们将新的实体挂载到体系后，需要和体系中已经存在的实体数据进行融合，即我们需要知道某一带有“李娜”条目名称的新的实体具体与体系中哪个实体对齐，此时依靠条目名称来区分和对齐实体已经不再有效，这时便遇到了多源知识库融合任务中的歧义问题，所以需要进行语义对齐处理.

同时，在网络百科中，描述一个实体（E）时，用户通常标记了不同的语义标签，例如：条目名称（TL）、条目 ID（ID）、类别标签（C）、属性标签三元组（TP）、非结构化文本关键词（S）来描述，形式化为 $E=\langle TL, ID, C, TP, S\rangle$. 值得说明的是，一个实体（$E$）往往会具有多个类别标签信息 $C=\{c_1, c_2, \cdots, c_n\}$，多个属性标签三元组信息 $TP=\{tp_1, tp_2, \cdots, tp_n\}$，这里 $tp_i=\{s, p, o\}$，s 表示实体条目名称，p 表示属性名，o 表示属性值以及非结构化文本关键词信息 $S=\{w_1, w_2, \cdots, w_n\}$.

那么，本文核心问题就是基于多种网络语义标签间的相似性，计算实体条目间的相似度，从而达到实体对齐的目的. 值得说明的是，本文的任务与传统实体消歧任务不同，传统的实体消歧任务是从非结构化文本中抽取实体指称项，并计算其与知识库中真实世界实体的相似度，建立语义连接关系；而本文任务是挖掘不同结构化知识库中实体间的连接关系，可以看做是面向知识库内部实体间的语义消歧问题.

4　基于网络语义标签的多源知识库实体对齐算法

4.1　算法概述

本文核心任务是两个实体间语义相似度计算，对实体进行对齐. 实体含有一定量描述信息，除非结构化文本关键词，还有大量用户提供的语义标签可以利用. 算法流程如图 1.

在将实体进行对齐之前，我们首先对数据进行了预处理，统计并记录了在各自知识库中拥有相同条目名称的实体. 接下来输入挂载的新百科实体（NE）的条目名称，去体系中查找存在与否；若不存在，则作为新建实体保存到体系中；若存在，则将与之相同条目名称的实体取出，加入候选对齐集 $EC=\{Ec_1, Ec_2, \cdots, Ec_n\}$，进行对齐计算.

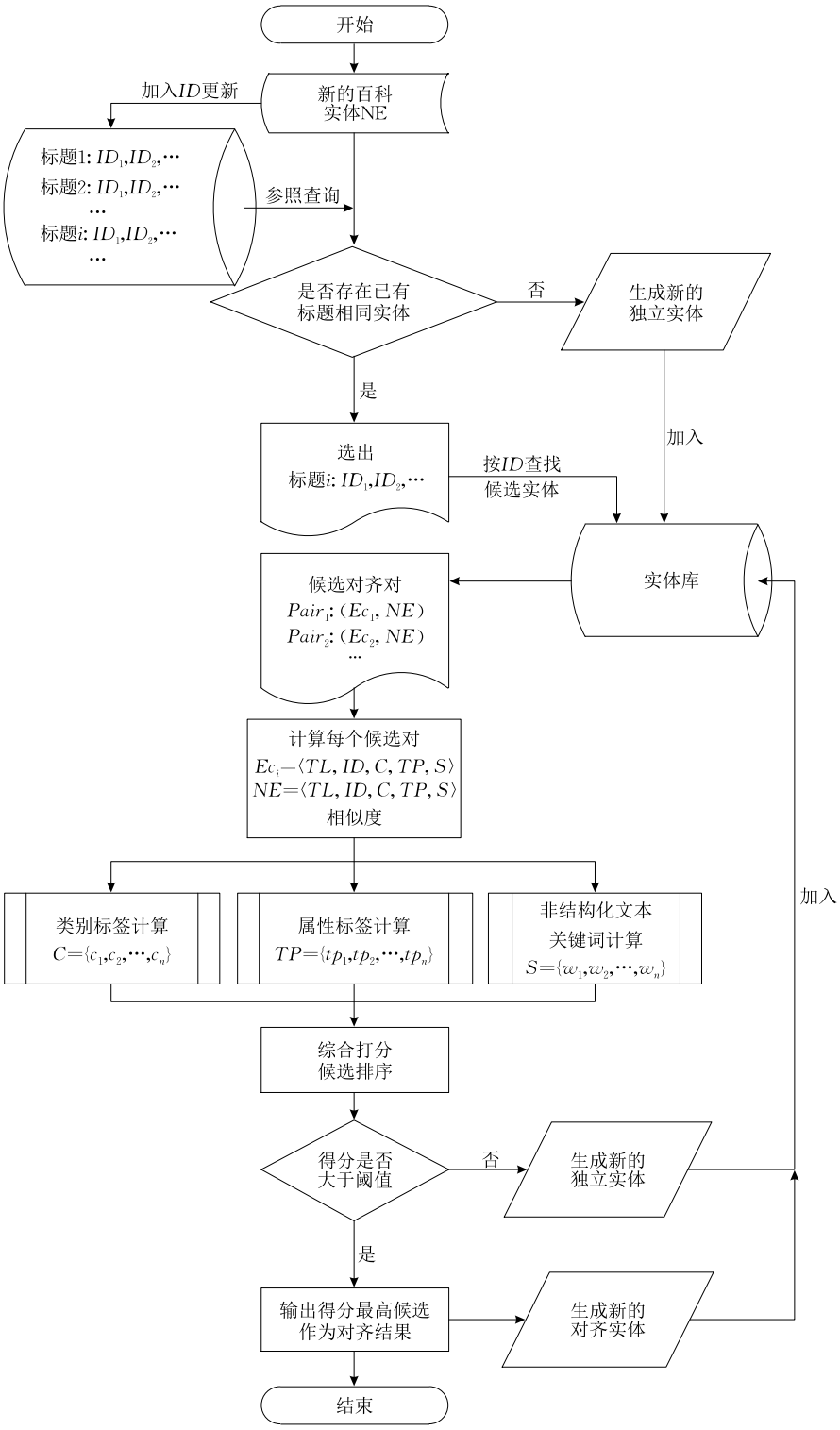


图 1 算法流程图

在对齐过程中,分别计算实体间类别标签、属性标签以及非结构化文本关键词三方面特征的相似度,综合计算 NE 与 Ec_i 的相似性,取相似性最大且大于一定阈值(在表 5 所示参数调优数据集上,使用网格搜索法,选取最优实验参数而定)的 Ec_i 作为对齐的结果输出,视为语义指向相同,并将 NE 作为与 Ec_i 对齐的实体保存到体系中;若相似性最大的候选实体不满足阈值,则视为 NE 在体系中并无语义指向相同的实体存在,将其作为新建实体保存到体系中,并加入到拥有相同条目名称的实体记录中.

为弥补中文知识库缺乏完整的描述体系结构的不足,我们参考了Schema.org(http://schema.org/),

并统计调研了互动百科、百度百科及豆瓣网的数据，依此建立了一个拥有 5 层分类，356 个类别和 278 个属性的知识描述体系. 接着依据实体类别标签与体系中类别进行相似度匹配计算，将互动百科的实体挂载到体系中，并用百度百科的实体数据对体系进行扩充. 我们总共向体系挂载了来自互动百科的实体 2777563 个，来自百度百科的实体 1844373 个.

4.2　基于属性标签匹配的语义相似度计算

属性标签是百科实体的重要特征，拥有很多个性化信息，例如两个同名电影实体，如果导演属性值或地区属性值不同，便具有了很大区分度. 属性特征匹配便显得十分必要了.

4.2.1　属性名的匹配

百科知识描述体系的建立结合百科页面特点，定义了 278 个属性. 为了更好地克服不同百科知识库属性间的异构性，我们围绕定义的 278 个属性，对互动百科与百度百科的实体数据进行了较全面的属性统计，经人工比对校验，编定了 524 条属性映射规则. 我们通过标签特征将百科实体挂载到体系中，接着依据编定的属性映射规则，将挂载到体系中的百科实体属性映射到定义的 278 个属性上. 映射覆盖率情况如表 1 所示.

表 1　百科知识库中属性三元组映射覆盖率

	挂载实体原属性三元组计数	被映射属性三元组计数	映射规则覆盖率/%
互动百科	7049496	6462515	91.67
百度百科	1258916	1116863	88.72

注：由于百度百科实体数目小于互动百科实体数目，故百度百科属性三元组数目较小.

对于两个百科实体间属性名的匹配，我们采用了这 524 条映射规则，对于两个属性标签三元组 $tp_1=\{s_1,p_1,o_1\}$ 与 $tp_2=\{s_2,p_2,o_2\}(s_1=s_2)$，计算两个属性名 p_1 与 p_2 是否映射到了同一属性上，如果是，则进一步计算两个属性值 o_1 和 o_2 的相似度.

4.2.2　属性值的匹配

考虑到很多属性值具有可归一化的特性，如密度、长度、重量等. 经过统计调研，我们制定了 7 类共计 37 条归一化规则(表 2)，如 1,45 厘米→1.45 米、145 cm→1.45 米. 对属性值进行归一化后，利用编辑距离[26]计算属性值的相似度.

将 $TP_1=\{tp_1,tp_2,\cdots,tp_n\}$ 与 $TP_2=\{tp_1,tp_2,\cdots,tp_n\}$ 间匹配属性的属性值相似度累加值作为两个百科实体的属性标签相似度.

表 2　属性值归一化规则分类计数

类别	规则数
密度	4
长度	10
重量	7
时间	5
面积	5
价格	5
人口	1
共计	37

$$SIM(TP_1,TP_2)=\sum_{tps_k\in Pair} sim(tps_k) \tag{1}$$

这里，$Pair$ 表示匹配的属性对集合，属性名匹配并且属性值相似度大于一定阈值(在表 5 所示参数调优数据集上，使用网格搜索法，选取最优实验参数而定)的属性对入选该集合；tps_k 表示匹配的属性对 (tp_i,tp_k)；$sim(tps_k)$ 表示匹配属性对 tps_k 的属性值相似度，依据属性值间的编辑距离计算得出. 此外，在实验过程中，我们发现两个实体匹配的属性名越多，两个实体越可能表示语义上的同一事物，故式(1)未做归一化处理，如果做归一化处理将很难体现此种特性.

4.3　基于类别标签匹配的语义相似度计算

在构建知识库的过程中，我们通过类别标签映射的方法，将实体挂载到知识库描述体系的相应类别下. 共挂载互动百科实体 2777563 个，百度百科实体 1844373 个. 在本文的实体对齐工作中，百科实体的类别标签特征是十分重要的信息(将在实验中说明).

4.3.1　基于编辑距离的类别标签语义相似度计算

在初期工作中，对于类别标签特征的匹配，采取了简单的对实体类别标签向量 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$ 计算编辑距离的方法，即计算两个类别标签向量间每个维度的编辑距离最小值的算术平均值，作为相似度输出. 但此种方法的实验效果并不理想.

统计发现，百度百科类别标签数为 849984 个，互动百科标签数为 31439 个，两者字面完全相同的标签数为 16676 个，仅占百度百科总数的 1.96%，占互动百科总数的 53.04%.

从同名实体类别标签重合度方面考虑，我们选取了百度百科和互动百科实体各 500000 个，得到了条目名称完全相同的实体对 59762 个，并按如下公式计算类别标签重合度.

$$overlap=\frac{2\times commCats}{|C_1|+|C_2|} \tag{2}$$

其中：$commCats$ 表示两个同名实体的相同类别标签数；$|C_1|$ 表示实体 1 类别标签数；$|C_2|$ 表示实体 2 的类别标签数.

从表 3 中可以看出，标签重合度大于 0.6 的实体对比例累加值只有 32.89%.

表 3 实体间的类别标签重合度

类别标签重合度	实体对数	比例/%
1	8360	13.99
0.8～1	5312	8.89
0.6～0.8	5983	10.01
0.4～0.6	7044	11.79
0.2～0.4	9971	16.69
0～0.2	21404	35.82

注：59762 个实体对中有 1688 对两个实体均无标签特征，比例为 2.82%.

综合两方面统计结果可知，百度百科与互动百科类别标签命名及标注存在着很大的差异. 故计算实体类别标签向量采用的方法不够有效，需要进一步寻找适宜的方法.

4.3.2 基于随机游走的类别标签语义相似度计算

为克服中文知识库缺乏完整的结构描述体系给中文知识库实体对齐带来的局限，我们从实体类别标签关于实体条目名称的共现信息入手，计算实体在潜在类别依赖关系中的相似度，即基于随机游走的类别标签语义相似度.

受 Han 等人[27]研究工作的启发，我们将实体的每个类别标签视为知识库中的概念，类别标签所标定的每个实体可视作概念的实例. 将每个百科实体用类别标签对应的概念向量来表示，$\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$. 利用式(3)来计算两个实体的类别标签语义相关度.

$$SIM(\boldsymbol{C}_1,\boldsymbol{C}_2)=\frac{1}{2}\times(SR(\boldsymbol{C}_1\rightarrow\boldsymbol{C}_2)+SR(\boldsymbol{C}_2\rightarrow\boldsymbol{C}_1)) \tag{3}$$

其中 $SR(\boldsymbol{C}_1\rightarrow\boldsymbol{C}_2)$ 表示类别标签向量 $\boldsymbol{C}_1$ 到类别标签向量 $\boldsymbol{C}_2$ 的语义相关度，则有

$$SR(\boldsymbol{C}_1\rightarrow\boldsymbol{C}_2)=\frac{\sum_{c_i\in\boldsymbol{C}_1}^{|\boldsymbol{C}_1|}w(c_i,\boldsymbol{C}_1)\times w(Align(c_i,\boldsymbol{C}_2),\boldsymbol{C}_2)\times sr(c_i,Align(c_i,\boldsymbol{C}_2))}{\sum_{c_i\in\boldsymbol{C}_1}^{|\boldsymbol{C}_1|}w(c_i,\boldsymbol{C}_1)\times w(Align(c_i,\boldsymbol{C}_2),\boldsymbol{C}_2)} \tag{4}$$

其中，$w(c_i,\boldsymbol{C}_1)$ 表示类别标签 c_i 在类别标签向量 $\boldsymbol{C}_1$ 中的权重.

$$w(c_i,\boldsymbol{C}_1)=|\boldsymbol{C}_1|^{-1}\Big(\sum_{c_k\in\boldsymbol{C}_1,c_k\neq c_i}sr(c_i,c_k)\Big) \tag{5}$$

其中，$Align(c_i,\boldsymbol{C}_2)$ 表示类别标签 c_i 在类别标签向量 $\boldsymbol{C}_2$ 中，与之语义相关性最大的类别标签.

$$Align(c_i,\boldsymbol{C}_2)=\arg\max_{c_p\in\boldsymbol{C}_2}sr(c_i,c_p) \tag{6}$$

式(6)中，$sr(c_i,c_k)$ 表示 c_i 和 c_k 的语义相关度.

以往有许多关于维基百科概念间的语义相关度计算工作，如 Strube 等人[28]，Gabrilovich 等人[29]，Witten 等人[30]. Han 等人[27]参考了 Witten 等人[30]衡量概念间语义相关度的方法. 受此启发并结合中文百科知识库特点，我们采用随机游走[31-32]方法，来衡量中文百科实体类别标签间的语义相关度.

考虑到实体的歧义性，我们选取了在百度百科中条目名称唯一的 5747561 个实体，在互动百科中条目名称唯一的 4431821 个实体，将其按条目名称合并类别标签信息，并统计记录了 834412 个类别标签及其频数. 部分统计情况列举如表 4.

表 4 类别标签频数排列

Top N	类别标签	频数
1	人物	803238
2	图书	440434
3	地理	223701
4	文化	193426
5	书籍	182093
6	生活	135040
7	词语	119198
8	历史	107650
9	美食	106901
10	文学	103420

出于算法的时间和空间代价考虑，我们无法用上所有类别标签进行随机游走的计算，必须选取合适的类别标签数目，关于类别标签数目的选取将在实验分析中说明.

选取好合适的类别标签数目后，我们建立类别标签间关于实体条目名称的共现矩阵，利用有重启的随机游走方法，计算类别标签间的语义相关度矩阵. 公式如下：

$$\boldsymbol{P}_i=(1-\lambda)\boldsymbol{M}_{\text{norm}}\cdot\boldsymbol{P}_{i-1}+\lambda\boldsymbol{P}_0 \tag{7}$$

这里，$\boldsymbol{P}_i$ 表示第 i 步随机游走得到的类别标签间语义相关度矩阵；$\boldsymbol{P}_0$ 表示初始单位矩阵；$\boldsymbol{M}_{\text{norm}}$ 表示归一化后的标签间关于实体条目名称的共现矩阵.

4.4　基于非结构化文本关键词的语义相似度计算

我们利用词向量相似度的方法，计算百科实体非结构化文本关键词之间的相似度. 我们通过 TF-IDF[33] 方法提取出非结构化文本（页面摘要）中的关键词，建立关键词的词向量 $\boldsymbol{S}=\{w_1,w_2,\cdots,w_n\}$，进而通过计算词向量的夹角余弦值作为相似度.

$$SIM(\boldsymbol{S}_1,\boldsymbol{S}_2)=\cos(\boldsymbol{S}_1,\boldsymbol{S}_2) \tag{8}$$

这里：$\boldsymbol{S}_1$ 表示实体 1 的非结构化文本关键词向量；$\boldsymbol{S}_2$ 表示实体 2 的非结构化文本关键词向量；$\cos(\boldsymbol{S}_1,\boldsymbol{S}_2)$ 表示实体 1 与 2 的非结构化文本关键词向量的夹角余弦值.

4.5　基于多语义标签匹配的语义相似度计算

综合类别标签、属性标签及非结构化文本关键词三方面特征信息，我们采用式(9)计算两个百科实体的相似度.

$$\begin{aligned}SIM(\boldsymbol{E}_1,\boldsymbol{E}_2)=&w_1\times SIM(\boldsymbol{TP}_1,\boldsymbol{TP}_2)+\\&w_2\times SIM(\boldsymbol{C}_1,\boldsymbol{C}_2)+\\&w_3\times SIM(\boldsymbol{S}_1,\boldsymbol{S}_2)\end{aligned} \tag{9}$$

得分最大且大于一定阈值（在表 5 所示参数调优数据集上，使用网格搜索法，选取最优实验参数而定）的候选百科实体 $Align(NE,EC)$ 作为 NE 的对齐结果输出.

$$Align(NE,EC)=\arg\max_{Ec_p\in EC} SIM(NE,Ec_p) \tag{10}$$

表 5　参数调优数据集中实体类别分布情况

类别	数目
人物	33
影视	33
图书	8
其他	9
总计	83

5　实验分析

5.1　数据集

出于算法时间和空间计算代价的考虑，为在随机游走中选定适宜的类别标签数量以及各优化算法参数，我们随机选取了 50 个实体条目名称进行调优实验. 这些名称在挂载到体系中的百度百科数据中，对应 83 个实体（均能找到语义指向相同的互动百科实体）. 该数据集为选定算法最优参数而设定，规模较小，方便人工标定与反复实验. 参数调优数据集中实体的类别分布情况如表 5.

选择好类别标签数目与各调优算法参数后，我们将实验扩展到了更大的数据集上. 随机选取了 800 个已挂载的百度百科实体，经人工认定其中有 622 个实体能在互动百科中找到与之对齐的实体.

5.2　实验设定

在表 5 所示的集合上我们分别用频数前 1200 个类别标签和频数前 8431 个类别标签的随机游走方法进行了对比实验，并人工标定实验结果.

在实体对齐的以往工作中，很多都依赖于属性标签的匹配，我们分别利用属性标签、类别标签、非结构化文本关键词进行了实验，并与综合利用属性、类别标签的实验及综合利用属性、类别、非结构化文本关键词的实验进行了对比. 对于算法中所需阈值选定以及式(9)中的权重参数，我们通过在表 5 所示参数调优数据集上，使用网格搜索法，选取最优实验参数而定.

5.3　实验结果

在进行类别标签数目选取实验之前，我们统计了类别标签关于挂载到知识库描述体系下互动百科和百度百科实体的覆盖率. 我们首先选取了综合覆盖率在 99.5% 附近的前 1200 个类别标签，实验效果如表 6 所示，并不理想；便又选取了频数在 100 以上的前 8431 个类别标签进行实验，取得了准确率为 90.36% 的较好实验效果.

表 6　类别标签数 1200 与 8431 的实验结果

类别标签数	准确率/%
1200	63.86
8431	90.36

从表 7 可以看出，占总体数量 1.01%（8431/834412）的类别标签覆盖了 99.996% 的百科实体条目. 选取大于 9000 的类别标签数对于覆盖率几乎没有提升，反而增加了计算的时间和空间代价；选取小于 8000 的类别标签数达不到对于实验有效的覆盖率.

实体对齐实验结果的 PR 曲线对比情况如图 2 所示. 单独利用属性标签进行匹配时的 PR 曲线，无法达到较高召回率，且在召回率 18% 左右开始急剧下降，主要原因是很多百度百科实体没有属性标签信息，匹配计算相似度为 0.

表 7 前 N 个类别标签的覆盖率

Top N	实体条目覆盖数(互动)	覆盖率(互动)/%	实体条目覆盖数(百度)	覆盖率(百度)/%	覆盖率(综合)/%
500	2720421	0.979427	1818532	0.985989	0.982046
1000	2753324	0.991273	1839413	0.997311	0.993683
1200	2759042	0.993332	1841031	0.998188	0.995270
2000	2772172	0.998059	1843096	0.999308	0.998557
3000	2775052	0.999096	1844061	0.999831	0.999389
5000	2776841	0.999740	1844322	0.999972	0.999833
8000	2777279	0.999898	1844354	0.999989	0.999934
8431	2777398	0.999941	1844355	0.999990	0.999960
9000	2777400	0.999941	1844356	0.999991	0.999961

注:共挂载互动百科实体 2777563 个,百度百科实体 1844373 个.

表 8 的统计显示,在我们所收集的所有百科实体数据中,含有属性标签的实体比例很低(互动百科为 20.43%,百度百科为 4.38%),但含有类别标签的实体均占了 65%以上(互动百科为 70.19%,百度百科为 65.31%). 表 9 的统计结果显示,在映射到我们知识库体系的百科实体中,含有属性标签的实体所占的比例仍然很低(互动百科为 32.08%,百度百科为 8.22%),而全部的实体含有类别标签信息(由于我们依据类别标签信息,将百科实体挂载到体系中,故含有类别标签信息的实体比例为 100%). 所以,根据互动百科和百度百科这两个中文知识库数据的特点,类别标签的匹配便显得十分必要.

表 8 百科知识库中含属性标签或类别标签的实体比例

来源	总数	含属性标签数	含属性标签比例/%	含类别标签数	含类别标签比例/%
互动百科	2556621	522324	20.43	1794397	70.19
百度百科	6058005	265628	4.38	3956660	65.31

表 9 体系中已挂载的含属性标签或类别标签的实体比例

来源	挂载总数	含属性标签数	含属性标签比例/%	含类别标签数	含类别标签比例/%
互动百科	2777563	890917	32.08	2777563	100
百度百科	1844373	151672	8.22	1844373	100

如图 2 所示,单独利用类别标签特征进行匹配时的 PR 曲线有着较好的特性,在召回率上升时,准确率没有大幅降低. 但由于一些百科实体的类别标签较泛化,单单利用类别标签还是无法区分两个百科实体,如两个同名实体都只标了人物的类别标签,就不能判断这两个实体是否指的是同一人. 另外也存在这样的情形,利用类别标签可以区分同名的两个实体,一个是电影,另一个是小说;但是无法区分同名的两个实体,一个是美国版电影,一个是韩国版电影. 这时就需要进行属性标签的匹配作为补充,进一步计算同名实体间的相似性. 如图 2 所示,并且类别标签和属性标签的实验结果 PR 曲线均优于其他两个标签单用,在召回率上升时,准确率也没有大幅降低.

如图 2 所示,单独利用非结构化文本关键词特征进行匹配时,PR 曲线在召回率上升时保持着较高的准确率,但在召回率接近 30% 时开始急剧下降.

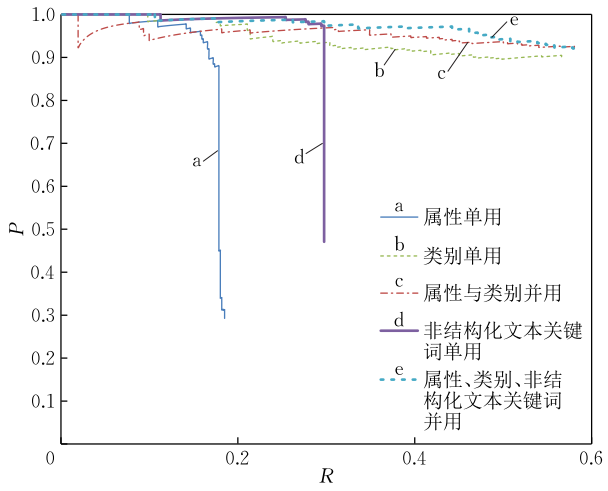


图 2 对比实验结果的 PR 曲线

主要原因如表 10,体系中已挂载的百度百科实体含非结构化文本关键词的比例很低(17.91%),不宜单独用作匹配特征.

表 10　体系中已挂载的含非结构化文本（页面摘要）的实体比例

来源	挂载总数	含非结构化文本(页面摘要)	比例/%
互动百科	2777563	1926614	69.36
百度百科	1844373	330417	17.91

从图 2 还可以看出，在综合利用属性标签与类别标签进行匹配的基础上，加入非结构化文本关键词特征，实验结果的 *PR* 曲线为最优，在较高召回率时仍保持了较高的准确率.

此外我们将本文算法与陈珂锐等人[24]和 Niu 等人[25]的工作进行了性能对比. 在本文测试集（随机选取了 800 个已挂载的百度百科实体，经人工认定其中有 622 个实体能在互动百科中找到与之对齐的实体）上做了对比实验，实验结果如表 11 所示.

表 11　对比实验结果

算法	*P*/%	*R*/%	*F*1
陈珂锐等人[24]	85.2	52.0	0.6458
Niu 等人[25]	80.1	51.3	0.6254
本文算法	94.8	55.9	0.7033

如表 11 所示，本文算法明显优于其他两种算法. 陈珂锐等人[24]在所提出的 AVP 平台中，只利用了百科词条中的属性值对作为特征模板，辅助属性值共现频率，利用扩展向量空间模型对词条进行歧义识别，未解决体系差异问题；Niu 等人[25]提出的 Zhishi.me 是首份关于中文 Linked Open Data 的工作，Zhishi.me 主要是利用原始网页中的页面重定位信息及实体名称归一化，对百科实体进行对齐，其对齐表现主要依赖于百科知识库原始网页的消歧信息，该工作只利用了实体名称方面的信息，未对实体对齐工作进一步深入研究. 在缺乏完整的知识库描述体系结构的条件下，本文提出的算法综合利用实体多种语义标签，并运用随机游走算法计算实体类别标签的关联关系，来对齐多源知识库中的实体，一定程度上弥补了中文知识库结构不完整的不足，同时也建立了一个较为合理的中文知识库体系结构，进一步扩展了已有的中文知识库融合工作.

6　总结与展望

近年来，随着互联网规模和用户的快速增长，网络上越来越多地富集起大量的知识信息，这些知识信息为人们的学习和生活提供了很大的便利. 知识库作为其中的一种知识信息富集的载体在研究与应用中起到了很大的作用. 知识库是多种自然语言处理任务的重要数据资源，但是单一知识库覆盖度低，不同知识库异构性强，不利于数据的共享和集成. 因此，多源知识库融合技术的研究有着十分重要的意义. 其中，多源知识库实体对齐是多源知识库融合技术中的重要组成部分. 本文给出了一种基于网络语义标签的多源知识库实体对齐算法，通过数据统计与实验对比发现，该模型能够较好地解决多源知识库实体对齐问题，算法在近 95% 的准确率下，仍能保持近 55% 的较好的召回率，达到了实际的应用标准. 在以后工作中，我们将考虑如何准确地挖掘各种网络语义标签之间的关联关系，同时利用诸如主题模型（PLSA、LDA）、主分量分析（PCA）、词向量模型（Word Embeddings）等方法提高语义相似度的计算精度，从而不断地完善与提高多源知识库实体对齐系统的性能.

致　谢　感谢各位评审老师给出的宝贵意见！

参　考　文　献

[1] Suchanek F M, Kasneci G, Weikum G. Yago: A core of semantic knowledge//Proceedings of the 16th International Conference on the World Wide Web. Banff, Canada, 2007: 697-706

[2] Suchanek F M, Kasneci G, Weikum G. Yago: A large ontology from Wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217

[3] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia—A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165

[4] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of Open Data//Aberer K, Choi K S, Noy N, et al, eds. The Semantic Web. Berlin Heidelberg, Germany: Springer, 2007: 722-735

[5] Bollacker K, Cook R, Tufts P. Freebase: A shared database of structured general human knowledge//Proceedings of the 22nd Conference on Artificial Intelligence. Vancouver, Canada, 2007: 1962-1963

[6] Grau B C, Dragisic Z, Eckert K, et al. Results of the ontology alignment evaluation initiative 2013//Proceedings of the 8th International Semantic Web Conference Workshop on Ontology Matching (OM). Washington, USA, 2013: 61-100

[7] Meilicke C, García-Castro R, Freitas F, et al. MultiFarm: A benchmark for multilingual ontology matching. Web Semantics: Science, Services and Agents on the World Wide Web, 2012, 15: 62-68

[8] Euzenat J. Brief overview of T-tree: The Tropes taxonomy building tool. Advances in Classification Research Online, 1993, 4(1): 69-88

[9] Giunchiglia F, Shvaiko P. Semantic matching. The Knowledge Engineering Review, 2003, 18(3): 265-280

[10] Parundekar R, Knoblock C A, Ambite J L. Linking and building ontologies of linked data//Patel-Schneider P F, Pan Y, Hitzler P, et al, eds. The Semantic Web—ISWC 2010. Berlin Heidelberg, Germany: Springer, 2010: 598-614

[11] Jain P, Hitzler P, Sheth A P, et al. Ontology alignment for linked open data//Patel-Schneider P F, Pan Y, Hitzler P, et al, eds. The Semantic Web—ISWC 2010. Berlin Heidelberg, Germany: Springer, 2010: 402-417

[12] Dieng R, Hug S. Comparison of personal ontologies represented through conceptual graphs//Proceedings of the 13th European Conference on Artificial Intelligence. Brighton, UK, 1998: 341-345

[13] Maedche A, Staab S. Measuring similarity between ontologies //Gómez-Pérez A, Benjamins V R eds. Richard Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Berlin Heidelberg, Germany: Springer, 2002: 251-263

[14] Madhaven J, Bernstein P, Rahm E. Generic schema matching using Cupid//Proceedings of the 27th International Conference on Very Large Data Bases. Rome, Italy, 2001: 49-58

[15] Raimond Y, Sutton C, Sandler M B. Automatic interlinking of music datasets on the semantic web//Proceedings of the Linked Data on the Web. Beijing, China, 2008: 269-276

[16] Nikolov A, Uren V, Motta E, et al. Integration of semantically annotated data by the KnoFuss architecture//Proceedings of the Knowledge Engineering: Practice and Patterns. Berlin Heidelberg, Germany: Springer, 2008: 265-274

[17] Volz J, Bizer C, Gaedke M, et al. Discovering and maintaining links on the web of data//Bernstein A, Karger D R, Heath T, et al, eds. The Semantic Web-ISWC 2009. Berlin Heidelberg, Germany: Springer, 2009: 650-665

[18] Stumme G, Maedche A. FCA-Merge: Bottom-up merging of ontologies//Proceedings of the IJCAI. Seattle, USA, 2001: 225-230

[19] Kalfoglou Y, Schorlemmer M. Ontology mapping: The state of the art. The Knowledge Engineering Review, 2003, 18(1): 1-31

[20] Noy N F. Semantic integration: A survey of ontology-based approaches. ACM SIGMOD Record, 2004, 33(4): 65-70

[21] Wache H, Voegele T, Visser U, et al. Ontology-based integration of information—A survey of existing approaches //Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, USA, 2001: 108-117

[22] Shvaiko P, Euzenat J. A survey of schema-based matching approaches//Spaccapietra S ed. Journal on Data Semantics IV. Berlin Heidelberg, Germany: Springer, 2005: 146-171

[23] Euzenat J, Shvaiko P. Ontology Matching. Berlin, Heidelberg, Germany: Springer, 2007

[24] Chen Ke-Rui, Pan Jun. Multi-source data fusion based on the expand vector space model. Journal of Shandong University (Natural Science), 2013, 48(11): 87-92(in Chinese)
(陈珂锐, 潘君. 基于扩展特征向量空间模型的多源数据融合. 山东大学学报(理学版), 2013, 48(11): 87-92)

[25] Niu X, Sun X, Wang H, et al. Zhishi.me-weaving Chinese linking open data//Aroyo L, Welty C, Alani H, et al, eds. The Semantic Web—ISWC 2011. Berlin Heidelberg, Germany: Springer, 2011: 205-220

[26] Ristad E S, Yianilos P N. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532

[27] Han X, Zhao J. Named entity disambiguation by leveraging Wikipedia semantic knowledge//Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). Hong Kong, China, 2009: 215-224

[28] Strube M, Ponzetto S P. WikiRelate! Computing semantic relatedness using Wikipedia//Proceedings of the 21st Conference on Artificial Intelligence. Boston, USA, 2006: 1419-1424

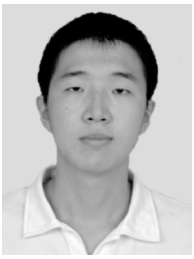
[29] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis// Proceedings of the International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 1606-1611

[30] Witten I, Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links//Proceeding of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy. Chicago, USA, 2008: 25-30

[31] Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications//Proceedings of the 6th International Conference on Data Mining (ICDM'06). Hong Kong, China, 2006: 613-622

[32] Tong H, Faloutsos C, Pan J Y. Random walk with restart: Fast solutions and applications. Knowledge and Information Systems, 2008, 14(3): 327-346

[33] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620



WANG Xue-Peng, born in 1988, Ph. D. candidate. His research interests include ontology alignment and opinion spam detection.

LIU Kang, born in 1981, Ph. D., associate professor. His research interests include sentiment analysis, question answering and information extraction.

HE Shi-Zhu, born in 1987, Ph. D., assistant professor. His research interests include information extraction and question answering.

LIU Shu-Lin, born in 1989, Ph. D. candidate. His research interests include event extraction and event prediction.

ZHANG Yuan-Zhe, born in 1986, Ph. D. candidate. His research interests include ontology alignment and question answering over heterogeneous linked data.

ZHAO Jun, born in 1966, Ph. D., professor, Ph. D. supervisor. His research interests include information retrieval and web mining, question answering and natural language processing.

Background

Knowledge base is an essential data source in many natural language processing tasks. But the coverage of uni-source knowledge base is so narrow. The hierarchies of different knowledge bases are also different. So, there are much of difficulties in data sharing and integrating between knowledge bases. Hence, the investigation on multi-source knowledge bases alignment turns to be much of significant. And multi-source knowledge bases entity alignment is an important component in multi-source knowledge bases aligning techniques. With the development of Semantic Web, there emerge numerous investigations on knowledge bases alignment; most of them focus on English knowledge bases. As summarized in recent surveys, the existing techniques are mostly based on calculating similarities between entities of two knowledge bases by utilizing various types of information in knowledge bases, e. g., entity names, taxonomy structures, constraints, and entities' instances. These methods can be classified into two categories: using a single strategy versus combining multiple strategies. In the former, all available information are defined as features in a single similarity function; while in the latter, different similarity functions are defined based on different types of information, and a composite method is used to combine the results of different similarities. But there are less similar works on Chinese knowledge base. Most of the existing methods rely on the taxonomy structures of the knowledge bases. But we realize that Chinese knowledge bases (e. g. Baidu Baike) are usually lack of complete taxonomy structures, and the existing methods are not so effect in aligning Chinese knowledge bases. To explore Chinese knowledge base, we tried a multi-source knowledge bases entity aligning method by leveraging semantic tags. This method utilized attribute triples, category tags and key words of unstructured text synthetically to align entities which are from Chinese encyclopedias. The experiments showed that our method is effective in solving the problem of knowledge bases entity alignment. It renders a 95% accuracy and a 55% recall at the same time. Our method meets the actual application requirements. Our work is supported by the National Natural Science Foundation of China (No. 61533018).