(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105912625 B
(45)授权公告日 2019.05.14

(21)申请号 201610213411.8

(22)申请日 2016.04.07

(65)同一申请的已公布的文献号
申请公布号 CN 105912625 A

(43)申请公布日 2016.08.31

(73)专利权人 北京大学
地址 100871 北京市海淀区颐和园路5号

(72)发明人 葛涛 穗志方

(74)专利代理机构 北京万象新悦知识产权代理有限公司 11360
代理人 黄凤茹

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)

(56)对比文件
CN 101645064 A,2010.02.10,
CN 104484461 A,2015.04.01,
CN 104408148 A,2015.03.11,
CN 102436456 B,2016.03.30,
US 2016092476 A1,2016.03.31,
US 2015324454 A1,2015.11.12,
Lu Chunliang.“Entity modeling and search in Text”.《https://static.chunlianglyu.com/docs/thesis.pdf》.2016,

审查员 曾伟涛

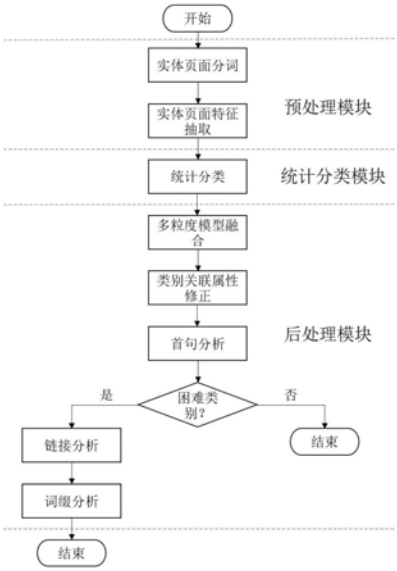
权利要求书2页 说明书8页 附图2页

(54)发明名称

一种面向链接数据的实体分类方法和系统

(57)摘要

本发明公布了一种面向链接数据的实体分类方法和系统，针对链接数据的实体分类问题，包括预处理、统计分类和后处理过程；其中，预处理通过对实体页面中的文本描述信息进行分词；由信息框的属性名和分词得到的词信息构成实体页面特征；统计分类过程采用多种切分粒度来训练统计分类模型对实体页面进行分类，得到实体类别的初步预测结果；后处理过程对实体统计分类结果进行修正，包括模型融合、语言知识、链接信息以及利用类别关联属性信息对融合后的实体类别进行修正等方法。本发明技术方案易实现、易调试、效率高、精度好，适合用来链接数据进行知识管理；能够实现对实体进行高精准分类。



CN 105912625 B

1.一种面向链接数据的实体分类方法，所述链接数据为多个实体页面，所述实体页面包含文本描述和信息框；所述实体分类方法包括预处理阶段、统计分类阶段和后处理阶段，具体包括如下步骤：

1)在预处理阶段过程，通过对实体页面中的文本描述信息进行分词，切分得到词信息；由信息框的属性名和所述词信息构成实体页面的特征；

2)在统计分类阶段，利用所述实体页面的特征，采用多种切分粒度来训练统计分类模型对实体页面进行分类，得到实体类别的初步预测结果；

3)在后处理阶段，对实体类别的初步预测结果进行修正，得到修正后的实体分类类别；所述修正包括如下步骤：

31)通过多粒度模型融合方法，将采用多个切分粒度训练的统计分类模型得到的实体类别的初步预测结果进行融合，得到融合后的实体类别结果；

32)构建类别属性数据库，利用类别属性数据库库中的类别关联属性信息，对融合后的实体类别进行修正，得到类别关联属性修正后的实体类别；

33)利用语法分析方法分析句子结构，通过对文本描述首句进行深度理解步骤32)所得到的类别关联属性修正后的实体类别，获取首句深度理解修正后的实体类别信息。

2.如权利要求1所述面向链接数据的实体分类方法，其特征是，步骤1)所述分词方法为前后最大匹配方法、后向最大匹配方法和基于统计序列标注方法中的一种。

3.如权利要求1所述面向链接数据的实体分类方法，其特征是，步骤2)采用两种切分粒度，分别为带有命名实体识别的切分粒度和不带有命名实体识别的切分粒度。

4.如权利要求1所述面向链接数据的实体分类方法，其特征是，所述统计分类模型为最大熵模型；步骤31)所述多粒度模型融合方法具体通过式1计算得到融合不同切分粒度分类器预测的概率分布，将多个切分粒度训练的最大熵分类模型对实体页面进行分类得到实体类别结果进行融合：

$$P_{multi}(y|x)=\lambda P_w(y|x)+(1-\lambda)P_n(y|x) \qquad (式1)$$

式1中，$P_{multi}(y|x)$为融合不同切分粒度分类器预测的概率分布；$P_w(y|x)$为只用词切分作为特征最大熵分类模型对于样本x预测的概率分布；y为样本类别，x为样本；$P_n(y|x)$为在词切分基础上加入命名实体标注作为特征的最大熵预测的概率分布；λ是调整线性插值权重的参数。

5.如权利要求1所述面向链接数据的实体分类方法，其特征是，步骤33)所述利用语法分析方法分析句子结构，获取首句深度理解修正后的实体类别信息，具体包括如下步骤：

331)对实体描述的首句进行依存句法分析，识别首句的宾语是否属于判断句宾语；

332)在大规模未标注语料上训练汉语词向量，定义词汇语义相似度，计算词向量与判断句宾语的词汇语义相似度，得到词汇语义相似度最高的词向量；

333)通过余弦相似度计算方法，设定余弦相似度阈值，当判断句宾语与其最相似类别的词向量的余弦相似度大于余弦相似度阈值，将该实体的类别修正为最相似类别。

6.如权利要求1所述面向链接数据的实体分类方法，其特征是，在所述后处理阶段对实体类别的初步预测结果进行修正，得到修正后的实体分类类别之后，使用困惑矩阵识别出困难实体类别；针对识别出的困难实体类别，通过链接分析方法和词缀分析方法对实体类别结果进行验证；所述困惑矩阵识别方法具体是：在验证集上，当统计分类模型对于某一实

体类别y_i的预测精度未达到90％时，类别y_i被视为困难实体类别。

7.如权利要求6所述面向链接数据的实体分类方法，其特征是，所述链接分析方法具体是：设定分类器对实体页面e所做出的类别预测为y’，将实体页面e所链接的实体页面的集合记为N(e)，找出N(e)中有类别标注的页面，统计得到N(e)中有类别标注的页面最多的类别，记作y*；当类别y*与类别预测y’不一致时，利用y*来修正y’的结果，得到实体页面e的类别为y*。

8.如权利要求6所述面向链接数据的实体分类方法，其特征是，所述词缀分析方法具体是：针对实体名称以固定汉字结尾的实体类别，利用大规模无标注数据学习得到的实体类型相关联的词缀信息，通过分别对最相近词汇的词缀进行频次统计，得到困难实体类别相关联的词缀，通过分析词缀获得所述实体的类别。

9.利用权利要求1～8所述面向链接数据的实体分类方法实现的面向链接数据的实体分类系统，其特征是，包括预处理模块、统计分类模块和后处理模块；

所述预处理模块用于对实体页面中的文本描述信息进行分词，将信息框属性名和分词得到的词信息作为特征抽取出来，作为实体页面的特征表示；

所述统计分类模块通过采用最大熵分类算法来训练分类模型，利用实体页面中对实体的描述信息识别得到实体类别；

所述后处理模块用于采用多粒度模型融合、类别关联属性和首句深入理解对所述统计分类模块得到的实体类别进行修正，得到修正后的实体类别。

10.如权利要求9所述面向链接数据的实体分类系统，其特征是，所述分词工具为StanfordCoreNLP工具包；所述分类模型采用最大熵分类器软件包Maxent。

一种面向链接数据的实体分类方法和系统

技术领域

[0001] 本发明属于信息处理领域，涉及链接数据分类和搜索，尤其涉及一种面向链接数据中的实体页面进行高精准分类的方法和系统。

背景技术

[0002] 目前处在大数据时代，如何最大限度地利用数据来帮助计算机进行信息处理已经成为了当前信息处理领域最热门的研究课题。近年来，随着Web2.0时代的到来，链接数据(例如语义网、知识图谱等)因为其强大的关系描述能力，得到了人们的广泛关注。链接数据是指象百度百科、维基百科的数据组织形式，这种数据中，每个页面对应一个实体，实体间有相互的链接，因此被称为链接数据(linked data)。随着数据规模的不断增大，采用人工方法管理链接数据已经不现实，迫切需要能够对链接数据进行知识管理的高效方法和系统。

[0003] 链接数据的实体分类是链接数据知识管理领域的一个重要技术问题，针对链接数据进行实体分类，能够有效地组织链接数据中大量的实体页面，从而加强用户搜索和阅读的体验。

[0004] 目前，实体分类的常用方法是针对实体的描述文本进行分类。但是，这种简单的方法在很多情况下并不能够准确地分析出实体的类别，其不足主要表现在：

[0005] (一)对于人来说，尽管根据文本描述来判断实体类别是一件很容易的事情，但是对于目前基于特征的统计分类方法而言，想要高精准地通过文本描述判断实体类别并不现实；例如，文本“X是根据著名游戏改编的动画”与“A是根据著名动画制作的游戏”在词汇级别有着非常相似的表示，但是前者是对一个动画实体的描述而后者是对游戏实体的描述，其描述的实体类型完全不同。因此，单纯基于文本特征的统计分类方法识别精度不足，并不能精准地获得实体类别。

[0006] (二)很多实体页面并没有足够的文本描述信息，这种情况下，单纯利用文本描述信息来对实体进行分类，必然会导致分类错误，通过文本描述无法得到实体类别。

发明内容

[0007] 为了克服上述现有技术的不足，本发明提供一种面向链接数据的实体分类方法和系统，针对链接数据的实体分类问题，通过统计分类过程和后处理过程来达到高精准实体分类的目的；其中，统计分类过程通过针对文本信息建模来进行分类；后处理过程利用丰富资源(例如词缀信息、链接数据等信息)对实体统计分类的结果进行修正，包括模型融合、语言知识、链接信息以及利用类别关联属性信息对融合后的实体类别进行修正等方法。

[0008] 链接数据中的实体页面通常包含文本描述和信息框(infobox)。本发明将文本描述进行切分以后，将信息框(infobox)属性名连同切分得到的词信息作为特征抽取出来，作为实体页面的特征表示；然后，对实体页面利用最大熵模型采用多种切分粒度进行分类，得到对实体类别的初步预测；再对所得到的实体类别进行后处理，以验证其分类结果是否可

靠;后处理具体包括对利用不同切分粒度的特征训练的分类器的分类结果进行融合;利用类别属性数据库库中的类别关联属性信息修正明显的预测错误;对文本描述首句进行深度理解,利用语法分析等方法分析句子结构,获取实体类别信息,以修正之前的预测结果;优选地,还可利用困惑矩阵识别难以正确分类的类别,针对难以正确分类的类别的预测进行进一步验证,包括使用实体页面所链接的相邻页面的类别对实体类别进行修正和使用实体页面的词缀信息对实体类别进行修正。

[0009] 本发明提供的技术方案是:

[0010] 一种面向链接数据的实体分类方法,所述链接数据为多个实体页面,所述实体页面包含文本描述和信息框;所述实体分类方法包括预处理阶段、统计分类阶段和后处理阶段,具体包括如下步骤:

[0011] 1)在预处理阶段过程,通过对实体页面中的文本描述信息进行分词,切分得到词信息;由信息框的属性名和所述词信息构成实体页面的特征;

[0012] 2)在统计分类阶段,利用所述实体页面的特征,采用多种切分粒度来训练统计分类模型对实体页面进行分类,得到实体类别的初步预测结果;

[0013] 3)在后处理阶段,对实体类别的初步预测结果进行修正,得到修正后的实体分类类别;所述修正包括如下步骤:

[0014] 31)通过多粒度模型融合方法,将采用不同切分粒度训练的统计分类模型得到的实体类别的初步预测结果进行融合,得到融合后的实体类别结果;

[0015] 32)构建类别属性数据库,利用类别属性数据库库中的类别关联属性信息,对融合后的实体类别进行修正,得到类别关联属性修正后的实体类别;

[0016] 33)利用语法分析方法分析句子结构,通过对文本描述首句进行深度理解步骤32)所得到的类别关联属性修正后的实体类别,获取首句深度理解修正后的实体类别信息。

[0017] 针对上述面向链接数据的实体分类方法,进一步地,步骤1)所述分词方法包括前后最大匹配方法、后向最大匹配方法和基于统计序列标注方法。

[0018] 针对上述面向链接数据的实体分类方法,进一步地,步骤2)采用两种切分粒度,分别为带有命名实体识别的切分粒度和不带有命名实体识别的切分粒度。

[0019] 针对上述面向链接数据的实体分类方法,进一步地,所述统计分类模型为最大熵模型;步骤31)所述多粒度模型融合方法具体通过式1计算得到融合不同切分粒度分类器预测的概率分布,将多个切分粒度训练的最大熵分类模型对实体页面进行分类得到实体类别结果进行融合:

[0020] $P_{multi}(y|x)=\lambda P_w(y|x)+(1-\lambda)P_n(y|x)$ (式1)

[0021] 式1中,$P_{multi}(y|x)$为融合不同切分粒度分类器预测的概率分布;$P_w(y|x)$为只用词切分作为特征最大熵分类模型对于样本x预测的概率分布;y为样本类别,x为样本;$P_n(y|x)$为在词切分基础上加入命名实体标注作为特征的最大熵预测的概率分布;λ是调整线性插值权重的参数。

[0022] 针对上述面向链接数据的实体分类方法,进一步地,步骤33)所述利用语法分析方法分析句子结构,获取首句深度理解修正后的实体类别信息,具体包括如下步骤:

[0023] 331)对实体描述的首句进行依存句法分析,识别首句的宾语是否属于判断句宾语;

[0024] 332)在大规模未标注语料上训练汉语词向量,定义词汇语义相似度,计算词向量与判断句宾语的词汇语义相似度,得到词汇语义相似度最高的词向量;

[0025] 333)采用余弦相似度计算方法,设定余弦相似度阈值,当判断句宾语与其最相似类别的词向量的余弦相似度大于余弦相似度阈值,将该实体的类别修正为最相似类别。

[0026] 针对上述面向链接数据的实体分类方法,进一步地,在所述后处理阶段对实体类别的初步预测结果进行修正,得到修正后的实体分类类别之后,使用困惑矩阵识别出困难实体类别;针对识别出的困难实体类别,通过链接分析方法和词缀分析方法对实体类别结果进行验证;所述困惑矩阵识别方法具体是:在验证集上,当统计分类模型对于某一实体类别y_i的预测精度未达到90%时,类别y_i被视为困难实体类别。

[0027] 进一步地,所述链接分析方法具体是:设定分类器对实体页面e所做出的类别预测为y',将实体页面e所链接的实体页面的集合记为N(e),找出N(e)中有类别标注的页面,统计得到N(e)中有类别标注的页面最多的类别,记作y*;当类别y*与类别预测y'不一致时,利用y*来修正y'的结果,得到实体页面e的类别为y*。

[0028] 针对上述面向链接数据的实体分类方法,进一步地,所述词缀分析方法具体是:针对实体名称以固定汉字结尾的实体类别,利用大规模无标注数据学习得到的实体类型相关联的词缀信息,通过分别对最相近词汇的词缀进行频次统计,得到困难实体类别相关联的词缀,通过分析词缀获得所述实体的类别。

[0029] 本发明还提供利用上述面向链接数据的实体分类方法实现的面向链接数据的实体分类系统,包括预处理模块、统计分类模块和后处理模块;所述预处理模块用于对实体页面中的文本描述信息进行分词,将信息框属性名和分词得到的词信息作为特征抽取出来,作为实体页面的特征表示;所述统计分类模块通过采用最大熵分类算法来训练分类模型,利用实体页面中对实体的描述信息识别得到实体类别;所述后处理模块用于采用多粒度模型融合、类别关联属性和首句深入理解对所述统计分类模块得到的实体类别进行修正,得到修正后的实体类别。

[0030] 上述面向链接数据的实体分类系统中,所述分词工具为Stanford CoreNLP工具包;所述分类模型采用最大熵分类器软件包Maxent。

[0031] 与现有技术相比,本发明的有益效果是:

[0032] 本发明提供一种面向链接数据的实体分类方法和系统,针对链接数据的实体分类问题,通过统计分类过程和后处理过程来达到高精准实体分类的目的。其中,在对文本进行基本分类的基础上,对于实体描述文本分类的结果进行修正,采用方法包括:

[0033] (一)采用多粒度词语切分模型融合方法,用于克服单一切分粒度在文本特征抽取上的缺陷;

[0034] (二)利用类别关联属性信息对融合后的实体类别进行修正,以达到修正明显错误的目的;

[0035] (三)通过首句深入理解,达到降低文本噪音的效果;

[0036] (四)能够识别困难样本,并对识别结果使用链接分析和词缀等方法进行验证。

[0037] 与现有技术相比,目前现有的实体分类方法不再进行处理,对于实体识别分类可能错误的情况无法修正结果;而本发明通过后处理流程对基于文本统计分类模块可能错误的情况进行修正。本发明所提出的技术方案易实现、易调试、效率高、精度好,非常适合企业

用来链接数据进行知识管理;能够对实体进行高精准分类。在JIST2015实体分类评测比赛中,本发明的方案准确率为98.6%,为当次评测比赛准确率最高的分类方案。

附图说明

[0038] 图1是本发明提供的面向链接数据的实体分类方法的流程框图。

[0039] 图2是本发明实施例提供的面向链接数据的实体分类系统的结构框图。

[0040] 图3是本发明提供方法中首句深入理解步骤的流程框图。

具体实施方式

[0041] 下面结合附图,通过实施例进一步描述本发明,但不以任何方式限制本发明的范围。

[0042] 本发明提供一种面向链接数据的实体分类方法和系统,针对链接数据的实体分类问题,通过统计分类过程和后处理过程来达到高精准实体分类的目的;其中,统计分类过程通过针对文本信息建模来进行分类;后处理过程利用丰富资源(例如词缀信息、链接数据等信息)对实体统计分类的结果进行修正,图1是本发明提供的针对链接数据的实体分类方法的流程框图。如图1所示,本发明方法包括预处理过程、统计分类过程和后处理过程;首先对实体页面进行分词特征抽取,然后利用抽取得到的特征训练统计分类模型。对于分类所得到的结果,我们首先利用多粒度模型融合来修正单模型预测错误,然后利用类别关联属性信息对融合后的实体类别进行修正,来修正一些明显的错误预测,再对实体页面的首句描述进行深度分析,来确定其类别。对于一些难以正确分类的类别的样本,本发明可通过链接分析和词缀分析方法对其类别进行再次修正。具体步骤包括:

[0043] 1)对于实体页面进行预处理,包括汉语分词(典型的分词方法有前后最大匹配、后向最大匹配以及基于统计序列标注的方法)、特征抽取(抽取词特征以及实体信息框属性名特征对页面进行表示)等,得到实体页面特征;

[0044] 2)利用步骤1)中抽取得到的实体页面特征,对实体页面利用最大熵模型采用多种切分粒度进行分类,得到对实体类别的初步预测;

[0045] 在本发明实施例中,利用最大熵模型训练两个分类器;一个分类器的特征表示用的是带有命名实体识别粒度切分的词+infobox属性;另一个分类器用的是不带有命名实体识别所进行的切分产生的词和infobox属性。

[0046] 3)对步骤2)中所得到的实体类别进行后处理,验证其分类结果是否可靠;具体包括如下步骤:

[0047] 31)对利用不同切分粒度的特征训练的分类器的分类结果进行融合;

[0048] 在本发明实施例中,采用两种切分粒度,分别指带有命名实体识别的切分和不带有命名实体识别;

[0049] 32)预先构建类别属性数据库,利用类别属性数据库库中的类别关联属性信息修正明显的预测错误;

[0050] 33)通过句法分析器对文本描述首句进行深度理解,利用语法分析等方法分析句子结构,从而获取实体类别信息,以修正之前的预测结果;

[0051] 34)利用困惑矩阵识别难以正确分类的类别,对该类别的预测进行进一步验证,包

括：

[0052] 341)使用实体页面所链接的相邻页面的类别对实体类别进行修正；

[0053] 342)使用实体页面的词缀信息对实体类别进行修正。

[0054] 图2是本发明实施例提供的面向链接数据的实体分类系统的结构框图。链接数据的实体分类系统包括预处理模块、统计分类模块和后处理模块；针对各模块进一步叙述如下：

[0055] 预处理模块

[0056] 链接数据中的实体页面通常包含文本描述和信息框(infobox)。

[0057] 在预处理模块中，我们利用了Stanford CoreNLP工具包对实体页面中的文本描述信息进行分词。本实施例中，我们采取了两种不同切分粒度：有命名实体识别和无命名实体识别。例如，在有命名实体识别的切分下，“纽约时代广场”将被视为一个词汇，而在无命名实体识别的切分下，该词将被切分为“纽约”、“时代”、“广场”三个词。

[0058] 在对于汉语文本进行切分以后，我们将信息框(infobox)属性名连同切分得到的词信息作为特征抽取出来，作为实体页面的特征表示。

[0059] 统计分类模块

[0060] 本发明主要利用实体页面中对实体的描述信息来作为判断实体类别的依据。本发明采用了自然语言处理领域常用的对数线性模型——最大熵分类算法来训练分类模型。如预处理模块所提到，统计分类模块所用到的特征包括词特征和信息框属性特征；词特征是经典的词袋模型特征表示；信息框属性特征对于识别实体的类别有着非常重要的作用，例如，“出生日期”也可能与人物类型的实体相关联。

[0061] 在文本分类模块，我们采用了不同粒度的词切分来训练文本分类模型，这是因为在有些情况下，一种切分粒度并不能满足对于分类的要求。例如，“纽约时代广场”如果作为一个命名实体来看待的话，对于分类的作用并不如将其切分成“纽约”“时代”和“广场”，因为“广场”一词对于类别有着至关重要的影响。另一方面，如果我们不进行命名实体识别，那么像“张一山”就会被切分成“张”“一”“山”，那么这也会对分类结果造成影响。因此，在统计分类模块中，本发明实施例通过最大熵分类器软件包Maxent(可由以下链接网站下载最大熵分类器软件包：http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html)训练了两种分类模型，一种是带有命名实体识别的细粒度切分、一种是单纯的粗粒度词切分。

[0062] 后处理模块

[0063] 基于文本统计分类模块可能错误的情况，本发明利用后处理模块来进行修正。后处理模块可执行以下过程：

[0064] 31)多粒度模型融合过程

[0065] 尽量模型融合在机器学习领域被广泛应用，但大多模型融合的方法都是针对不同种机器学习模型的融合。对于自然语言(尤其是中文)来说，切分粒度的不同对于整个模型的效果会产生影响。针对不同切分粒度的各自优劣性，本发明提出了利用模型融合的方法来对各种切分粒度所得到的分类模型进行“取长补短”。

[0066] 我们定义$P_w(y|x)$为只用词切分作为特征、最大熵分类模型对于样本x预测的类别y概率分布，$P_n(y|x)$为在词切分基础上加入命名实体标注作为特征的最大熵预测的概率分布。我们将这两种分类器的结果用以下方法进行融合：

[0067] $P_{multi}(y|x)=\lambda P_w(y|x)+(1-\lambda)P_n(y|x)$ (式1)

[0068] 式1中，$P_{multi}(y|x)$为融合不同切分粒度分类器预测的概率分布；$P_w(y|x)$为只用词切分作为特征最大熵分类模型对于样本x预测的概率分布；y为样本类别，x为样本；$P_n(y|x)$为在词切分基础上加入命名实体标注作为特征的最大熵预测的概率分布；λ是调整线性插值权重的参数，本实施例中，设λ＝0.5。

[0069] 32)类别关联属性修正预测

[0070] 该模块利用类别关联属性修正一些明显错误的类别预测。该模块所利用的主要是信息框属性的类别特异性。如表1所示，对于某些属性而言，它们不可能与有些特定的类别相关联。例如，“游戏平台”不可能与城市实体相关联。因此，利用这些属性的特异性，可以修正分类器明显的预测错误。本发明针对预定义好的实体类型人工建立了类别属性数据库，用来进行对预测的修正。

[0071] 表1类别关联属性示例

[0072]

属性	可能类别

[0073]

游戏平台	游戏
出版时间	小说、漫画
出生日期	歌手、政治家
面积	城市、景点

[0074] 33)通过依存句法分析器深入理解实体描述的首句，进一步精准识别实体类别；

[0075] 链接数据(例如:维基百科、百度百科等)中的实体页面描述的第一句话通常是对实体的定性描述(例如:砸六家是一种流行于天津的扑克牌游戏)。如果能够深入理解实体描述的首句，那么将会对精准识别实体类别有着非常大的帮助。

[0076] 图3是本发明提供方法中首句深入理解步骤的流程框图。本发明首先利用依存句法分析器来找出实体页面文本描述首句中的判断句宾语，然后利用该判断句宾语分析实体页面的类别；具体包括如下步骤：

[0077] 331)判断句宾语识别

[0078] 本发明利用了斯坦福大学依存句法分析器，对实体描述的首句进行依存句法分析，分析出首句中的主语、谓语和宾语。如果依存句法所得到的首句的宾语与“是”有直接的依存关系，那么该宾语被称为“判断句宾语”；否则，该宾语被称为“非判断句宾语”。

[0079] 如果实体文本描述的首句的宾语为判断句宾语，我们可以利用该宾语为线索确定实体的类别，从而验证分类器预测的结果是否准确。如果分类器预测的结果与断句宾语所得出的结论矛盾，则利用该结果修正分类器的预测。如果首句中不存在判断句宾语，则跳过该步骤，进入34)。

[0080] 例如，在“砸六家是一种流行于天津的扑克牌游戏”句中，依存句法分析结果分析得到“游戏”为该句宾语，并且“游戏”与“是”有直接依存关系，那么“游戏”即为该句的判断句宾语。如果“游戏”是实体分类体系中预定义的实体类别，那么我们用它来作为该实体的类别。

[0081] 332)利用判断句宾语修正类别预测

[0082] 在一些情况下,即使我们找出了判断句宾语,也不能随意用来对预测进行修正,因为这样可能会引入一些不必要的错误。同时,在很多情况下,判断句宾语并不完全匹配类别名称。例如:“野泽雅子是日本著名声优”,尽管依存句法分析可以得到“声优”是这句话的判断宾语,然而预定义的实体类别中有可能并没有“声优”这个类别。对此,本发明定义了修正条件,利用词汇语义相似度,即词向量间的余弦相似度,从大规模未标注语料中来寻找判断句宾语最相似的类别,来可靠地进行类别修正。

[0083] 在自然语言处理领域,余弦相似度通常被当作词汇的语义相似度。具体来说,本发明实施例首先利用了使用word2vec工具包(https://word2vec.googlecode.com/svn/trunk/)在Gigaword中文语料(汉语Gigaword是公开的数据集)上训练汉语词向量,利用训练得到的词向量来寻找与判断句宾语语义最相似的类别名称。如果判断句宾语与其最相似类别的词向量的余弦相似度大于预设定的阈值(本发明实施例中,通过计算余弦相似度的方法,余弦相似度阈值设定为0.9),才将该实体的类别修正为最相似类别。

[0084] 为此,我们定义实体页面首句文本描述的判断句宾语为w_0,类别词为$y \in Y$(Y为实体类别集合),$sim(w_1,w_2)$为词语w_1、w_2的词向量的余弦相似度。那么修正条件为式2如示:

[0085] $$y^* = \text{argmax}_{y \in Y} sim(w_0,y) \wedge sim(w_0,y^*) > 0.9 \quad (式2)$$

[0086] 式2中,^表示并且(与)关系;^前部分的内容(左边项)表明y*是语义相似度最高的类别,^后部分的内容(右边项)表示y*与w0的相似度需要高于0.9;修正条件(式2)满足才进行修正,即只有当y*是语义相似度最高的类别并且y*与w0的相似度需要高于0.9时,用y*来修正原有的类别预测。

[0087] 在上面例子(“野泽雅子是日本著名声优”)中,我们可以找出与“声优”最相似的类别是“演员”(如表2所示,表2是利用从汉语gigaword上训练的词向量计算出的与类别最相似的一些词汇,其中粗体词表示这些词汇与类别的相似度在0.9以上),并发现“演员”与“声优”的语义相似度在0.9以上,因此,将“野泽雅子”这个实体页面的类别修正为“演员”。

[0088] 表2类别最相似词汇

[0089]

insect	university	game	politician	city	song	novel	scene	cartoon	actor
昆虫	**大学**	**游戏**	**政治家**	**城市**	**歌曲**	**小说**	**景点**	**漫画**	**演员**
真菌	**师范大学**	MMORPG	外交家	省会	**歌词**	小说集	**景区**	**动画**	**导演**
软体动物	外国语	电脑游戏	哲学家	中小城市	**演唱**	自传体	**名胜**	**动画片**	编剧
哺乳动物	复旦	Playstation	知识分子	大都市	主题曲	毛姆	**旅游**	音乐剧	歌手
哺乳类	学院	电子游戏	政界	C B D	歌名	传记	旅游点	电影	知名演员
孢子	南开	j5	知识界	城区	曲目	短篇小说	风景点	电视电影	傅晶
地衣	中山大学	FLASH	活动家	商业区	诗朗诵	科幻	名胜古迹	动画制作	主演
植物	理工学院	玩	军事家	区向东	M V	寓言	风景区	舞台剧	王志飞
寄主	工学院	传统网络	历史学家	大中城市	主题歌	剧作	旅游区	电视动画	剧中
藻类	分校	J5	汉学家	远郊区	词谱	莎士比亚	游览	艺术片	王丽坤
害虫	商学院	WebGame	思想家	繁华	郭易	散文	新景点	纪录片	出演
浮游	同济	widget	远见卓识	核心区	录制	武侠小说	景观	歌舞片	林兆华
微生物	法学院	角色扮演	神学家	郊区	曲子	章回体	风景	系列片	李范秀

[0090] 34)使用困惑矩阵识别困难样本

[0091] 在实际应用中,我们经常会遇到某些类别的样本难以区分,这类样本称为困难样本。例如对于“城市”和“景点”两个类别的实体,分类器往往会做出错误的预测,因为这两类实体的描述和信息框属性都很相似。为了提高分类的精准度,本发明使用困惑矩阵来找出

分类词容易出错的样本类别。具体来说，如果在验证集上，统计分类模型对于某一实体类别y_i的预测精度未达到90％，则类别y_i被视为困难样本类别。例如，在验证集上，统计分类模型对18个实体页面预测为“城市”类别，但其中只有15个页面确实为“城市”类别，因此统计分类模型在“城市”类别的预测精度仅为83.33％(15/18)，“城市”类别被认定为困难样本类别。对于那些被统计分类模型预测为困难样本类别的样本，我们称之为困难样本。

[0092] 对于识别出的困难样本，我们利用了以下两种方法来对结果进行验证。

[0093] 341)链接分析

[0094] 对于困难样本，单靠实体页面上的内容可能不足以做出正确的判断，因此，本发明采用了链接分析方法来对困难样本进行分类结果验证。

[0095] 在链接数据中，一个实体页面通常会链接到与其相关的其它的实体页面。通常来说，其链接到的其它实体页面的类别非常有可能与其本身的类别的相同的。因此，利用一个实体页面链接到的其它实体页面的类别，可以帮助系统更好的判断该实体的类别。

[0096] 具体来说，对于某实体页面e，我们分析e所链接的实体页面，其集合记为N(e)。N(e)中会有一部分页面有类别标注信息。本发明找出N(e)中有类别标注的页面，并统计出这些页面最多的类别y*，判断该类别是否与分类器对e所做出的类别预测y’一致。如结果不一致，利用y*来修正y’的结果。

[0097] 342)词缀分析

[0098] 对于某些难以区分的样本，本发明还利用了词缀分析法来验证其分类结果。对于某些类别，其实体名称通常以固定汉字结尾。例如，“城市”实体通常以“市、县”结尾，“景点”实体通常会以“湖、山”等结尾。表3列出了类别常见实体词缀的实例。

[0099] 表3常见实体的类别词缀

[0100]

景点	城市
山、湖、城、道、景、河、岭、洞	区、市、县

[0101] 本发明首先提出利用大规模无标注数据学习实体类型相关联的词缀信息，具体来说，我们利用词向量工具包word2vec在中文Gigaword数据集上训练词向量，然后通过计算余弦相似度的方法，找出每个类别语义最相近的词(词向量余弦相似度0.7以上的词)。然后，通过分别对这两个景点的最相近词汇的词缀进行频次统计，就可以得到困难样本类别相关联的词缀，从而通过分析词缀，来确定其所属类别。具体来说，如果某一实体页面词缀s在某一类别y_1中的频率显著高于(2倍以上)另一类别y_2中的出现频率，则我们将y_1作为该实体类别修正原有预测结果。举例来说，对于“庐山仙人洞”实体页面，其词缀“洞”出现在“景点”类别的频率明显高于出现在“城市”类别的频率，因此将该实体的预测类别修正为“景点”。

[0102] 需要注意的是，公布实施例的目的在于帮助进一步理解本发明，但是本领域的技术人员可以理解：在不脱离本发明及所附权利要求的精神和范围内，各种替换和修改都是可能的。因此，本发明不应局限于实施例所公开的内容，本发明要求保护的范围以权利要求书界定的范围为准。

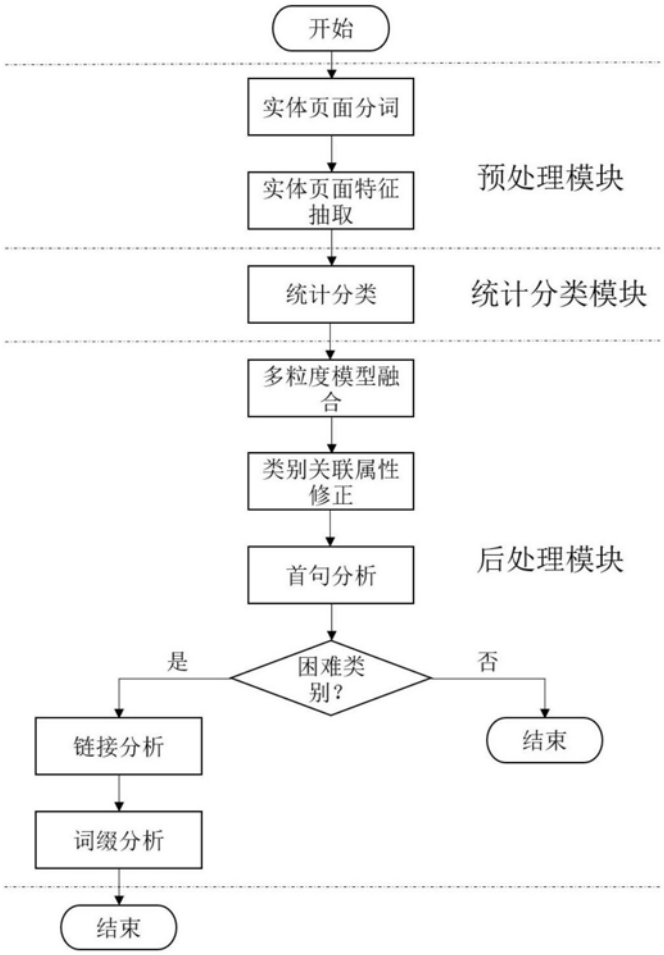

开始
实体页面分词
实体页面特征抽取
预处理模块
统计分类
统计分类模块
多粒度模型融合
类别关联属性修正
首句分析
后处理模块
是
困难类别?
否
链接分析
结束
词缀分析
结束

图1

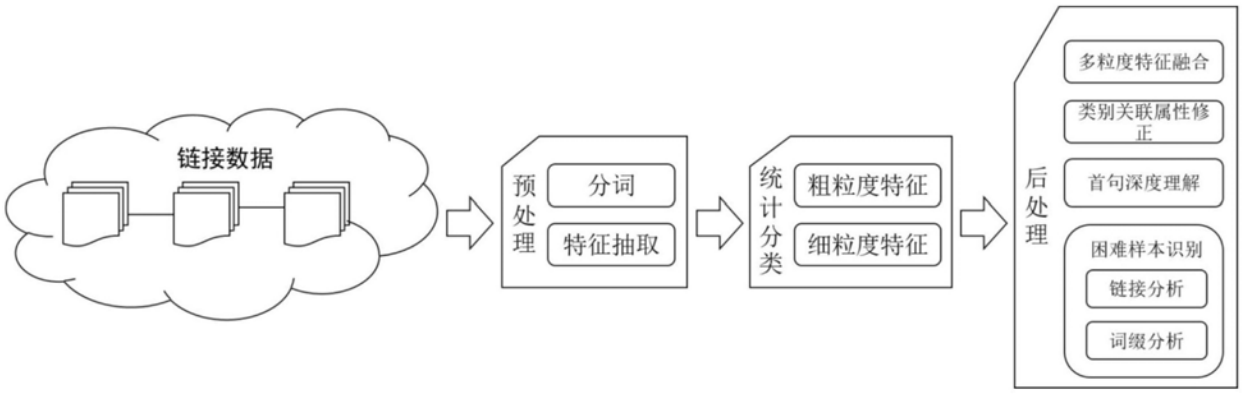

链接数据
预处理
分词
特征抽取
统计分类
粗粒度特征
细粒度特征
后处理
多粒度特征融合
类别关联属性修正
首句深度理解
困难样本识别
链接分析
词缀分析

图2

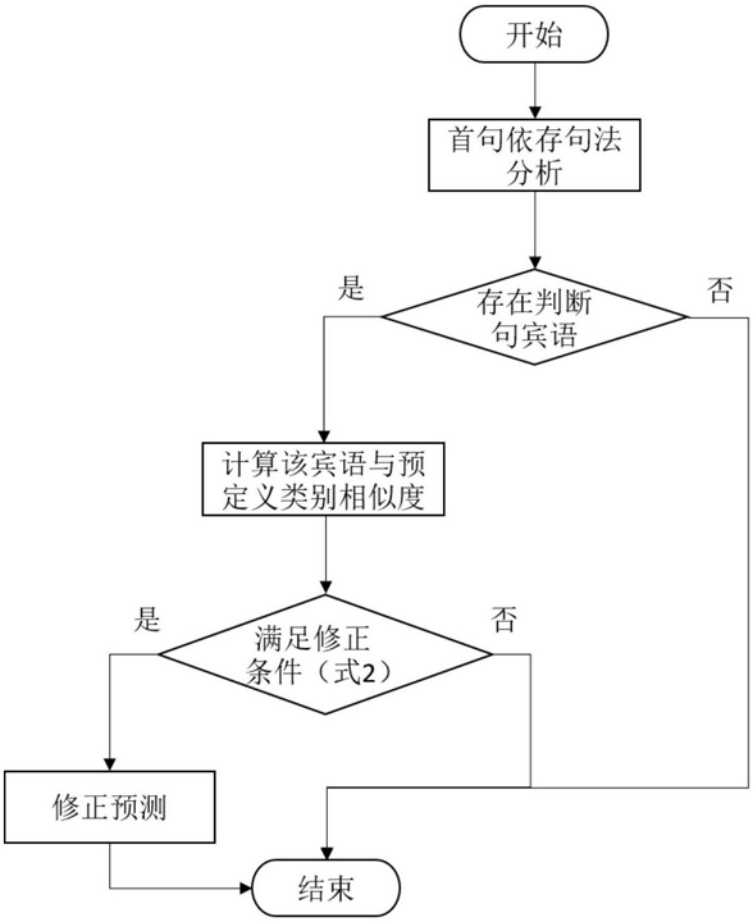


图3