(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108415902 A
(43)申请公布日 2018.08.17

(21)申请号 201810138076.9

(22)申请日 2018.02.10

(71)申请人 合肥工业大学
地址 230009 安徽省合肥市包河区屯溪路193号

(72)发明人 吴共庆 何颖 胡学钢 胡东辉 李磊 吴信东

(74)专利代理机构 安徽合肥华信知识产权代理有限公司 34112
代理人 余成俊

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/30(2006.01)

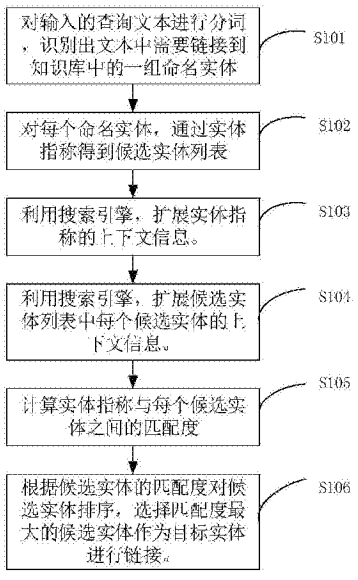
权利要求书2页 说明书10页 附图1页

(54)发明名称

一种基于搜索引擎的命名实体链接方法

(57)摘要

本发明公开了一种于搜索引擎的命名实体链接方法，包括下述步骤：对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体；对识别出的每个命名实体，通过实体指称在中文实体知识库中搜索得到候选实体列表；利用搜索引擎，扩展实体指称的上下文信息；利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息；计算实体指称与每个候选实体之间的匹配度；选择匹配度最大的实体进行链接。本方法通过在搜索引擎中检索实体指称和候选实体，从搜索结果中获取信息以扩展实体指称和候选实体的上下文信息，为提高实体链接的准确性提供附加信息。



CN 108415902 A

1.一种基于搜索引擎的命名实体链接方法，其特征在于：包括以下步骤：

步骤1，对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体；

步骤2，对每个命名实体，通过实体指称得到候选实体列表；

步骤3，利用搜索引擎，扩展实体指称的上下文信息；

步骤4，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息；

步骤5，计算实体指称与每个候选实体之间的匹配度；

步骤6，根据候选实体的匹配度对候选实体排序，选择匹配度最大的候选实体做为目标实体进行链接。

2.根据权利要求1所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤1中：

命名实体为人名、机构名、地名以及其他所有以名称为标识的实体，实体为现实生活中的对象或概念；

查询文本为用户输入的文本或文本集，包括实体指称和其上下文文本，其文本长度为固定长度；

知识库为若干实体条目的集合，包括人物、机构、地名以及其他所有以名称为标识的实体；

实体条目为每个实体在知识库中存储的具体信息，包括实体的上下文文本、实体属性和属性值，实体的上下文信息为对实体进行详细描述的文本，实体指称为查询文本中待链接实体的名词引用。

3.根据权利要求1所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤2中，通过实体指称得到候选实体的方法为：

依次取出知识库中的每条实体条目，遍历其名称属性值、别名属性值，计算查询文本中的实体指称与各个实体条目的名称属性值、别名属性值之间的字符串相似度；若查询文本的实体指称与实体条目的名称属性值相似度大于预先设定好的阈值，则把当前实体条目输出到候选实体列表中，若查询文本的实体指称与实体条目的别名属性值相似度大于预先设定好的阈值，则把所述别名属性所对应的实体条目输出到候选实体列表中；

通过上述方法得到初始候选实体列表，对初始候选实体列表进行筛选得到候选实体列表，方法为：

抽取知识库中的部分实体做为训练集，利用标注的<实体，类型>对集合训练实体-类型分类器，然后利用实体-类型分类器识别候选实体列表中实体的类型，把与查询文本中待链接实体属于同类型的实体保留，删除类型不同的实体，得到最终的候选实体列表；

所述实体的类型为实体的类别属性，包括人物、地点、机构、时间；所述<实体，类型>对为实体与其类型的映射，其中映射正确的<实体，类型>对为训练分类器的正向实例，反之为负向实例；所述实体-类型分类器为通过<实体，类型>对集合训练的多类分类模型，通过所述模型可以判断实体的类型。

4.根据权利要求1所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤3中，利用搜索引擎，扩展实体指称的上下文信息的方法为：

将步骤1中查询文本中一组命名实体的实体指称做为种子在搜索引擎中检索，得到与

所述实体指称相关的搜索结果网页，从搜索结果网页中抽取前m条搜索结果的标题和摘要信息，得到所述实体指称的扩展文本，查询文本与扩展文本一起构成所述实体指称的扩展的上下文信息。

5.根据权利要求1所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤4中，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息的方法为：

将候选实体列表中的每个候选实体在知识库中的上下文文本进行分词，对分词后的文本进行命名实体识别得到一组与所述候选实体相关的命名实体，然后把这一组命名实体做为一个种子在搜索引擎中检索，得到与所述候选实体相关的搜索结果网页，从搜索结果网页中抽取前n条搜索结果的标题和摘要信息，得到所述候选实体的扩展文本，候选实体在知识库中的上下文文本与扩展文本一起构成所述候选实体的扩展的上下文信息。

6.根据权利要求1所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤5中，计算实体指称与每个候选实体之间的匹配度的方法为：

对候选实体列表中的每个候选实体，计算所述实体指称的扩展的上下文信息和所述候选实体的扩展的上下文信息之间的余弦相似度a，计算所述候选实体流行度b，计算实体指称与每个候选实体之间的匹配度$c=w_a \times a+w_b \times b$，其中：$w_a$、$w_b$为预设的权值，且$w_a+w_b=1$；所述候选实体的流行度为：实体指称做为超链接链接到所述候选实体页面的次数与实体指称做为超链接链接到所有候选实体的总次数的比值。

一种基于搜索引擎的命名实体链接方法

技术领域

[0001] 本发明涉及网络信息处理方法，具体是一种基于搜索引擎的命名实体链接方法。

背景技术

[0002] 随着信息化的不断提升，人们越来越多的通过互联网获取信息，传统的信息传播媒体如报纸、杂志、期刊逐渐被门户网站、电子图书馆和搜索引擎所超越。通过互联网获取的信息大多是通过新闻网站、微博、贴吧等平台传播的文本信息，这些文本中蕴含着大量的命名实体，造成信息时代在提供高效的阅读方式的同时，伴随着信息量爆炸式的增长。如何高效、准确地把这些大量的命名实体无歧义地链接到知识库中对应的实体中，成为信息融合、自然语言处理、信息检索等领域亟待解决的问题。

[0003] 命名实体链接主要分为两个阶段，第一阶段主要是根据实体指称在知识库中搜索得到潜在的候选实体集，第二阶段主要是根据实体指称和候选实体集进行实体消歧。目前，命名实体链接方法主要工作集中在优化实体消歧算法上，实体消歧的主要方法主要分为基于实体流行度排序的消歧方法、基于上下文相似度排序的消歧方法、基于分类的消歧方法、基于图的消歧方法等。

[0004] 基于实体流行度排序的消歧方法是最简单和最直接的实体消歧方法，这种方法根据实体流行度判断返回流行程度最高的实体。例如人物实体，知名度越高的人的被链接的可能性越大，这种知名程度通常通过与实体相关链接的多少来衡量，但这种方法不足之处在于：无论查询的实体流行与否都会返回同样的结果。

[0005] 基于上下文相似度排序的消歧方法利用实体指称的上下文与候选实体上下文之间的相似度来判断应该链接的实体。这种方法通常把实体指称的上下文与候选实体上下文表示成词袋向量或者把其中的一些关键词如文本中的命名实体表示成词空间向量，然后计算向量之间相似度，如余弦相似度。它的缺点在于相似度的计算依赖于词共现信息，需要大量的实体上下文信息，当上下文信息较少时判断结果的准确性不高。

[0006] 基于分类的消歧方法将链接过程看成一个分类过程，在训练阶段从知识库中抽取实体指称与知识库中实体建立链接的<实体指称，候选实体>对做为正向实例，未建立链接的<实体指称，候选实体>对做为负向实例。然后，使用这些正向实例和负向实例构成的训练集训练一个分类器，分类器使用的常见分类模型包括SVM、决策树、朴素贝叶斯等。在命名实体链接阶段，利用训练好的分类器对实体指称的候选实体进行分类。然而，这种方法容易得到两个或更多的链接实体，需要结合上下文相似度等方法再次进行排序。

[0007] 基于图的消歧方法将实体指称与候选实体做为图节点，将实体指称与候选实体之间的关系做为图的边建立图结构。通常在计算实体指称节点与其候选实体节点之间的关系时，多数方法利用了的两者之间的上下文相似度特征。因此，基于图的消歧方法在一定程度上也依赖于词共现信息，也存在当上下文信息较少时判断结果的准确性不高的问题。

[0008] 总结以上所述的实体消歧方法，命名实体链接的大部分方法都依赖于实体指称上下文与候选实体上下文相似度特征，然而这一特征的计算需要大量的实体指称和其候选实

体的上下文信息来克服计算相似度时过分依赖词共现信息的问题。为了得到更多的实体指称与候选实体的上下文信息，需要一种能扩展实体指称与其候选实体上下文信息的方法以提高计算结果的准确性。

发明内容

[0009] 本发明的目的是提供一种基于搜索引擎的命名实体链接方法，以解决现有技术存在的问题。

[0010] 为了达到上述目的，本发明所采用的技术方案为：

[0011] 一种基于搜索引擎的命名实体链接方法，其特征在于：包括以下步骤：

[0012] 步骤1，对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体；

[0013] 步骤2，对每个命名实体，通过实体指称得到候选实体列表；

[0014] 步骤3，利用搜索引擎，扩展实体指称的上下文信息；

[0015] 步骤4，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息；

[0016] 步骤5，计算实体指称与每个候选实体之间的匹配度；

[0017] 步骤6，根据候选实体的匹配度对候选实体排序，选择匹配度最大的候选实体做为目标实体进行链接。

[0018] 所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤1中：

[0019] 命名实体为人名、机构名、地名以及其他所有以名称为标识的实体，实体为现实生活中的对象或概念；

[0020] 查询文本为用户输入的文本或文本集，包括实体指称和其上下文文本，其文本长度为固定长度；

[0021] 知识库为若干实体条目的集合，包括人物、机构、地名以及其他所有以名称为标识的实体；

[0022] 实体条目为每个实体在知识库中存储的具体信息，包括实体的上下文文本、实体属性和属性值，实体的上下文信息为对实体进行详细描述的文本，实体指称为查询文本中待链接实体的名词引用。

[0023] 所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤2中，通过实体指称得到候选实体的方法为：

[0024] 依次取出知识库中的每条实体条目，遍历其名称属性值、别名属性值，计算查询文本中的实体指称与各个实体条目的名称属性值、别名属性值之间的字符串相似度；若查询文本的实体指称与实体条目的名称属性值相似度大于预先设定好的阈值，则把当前实体条目输出到候选实体列表中，若查询文本的实体指称与实体条目的别名属性值相似度大于预先设定好的阈值，则把所述别名属性所对应的实体条目输出到候选实体列表中；

[0025] 通过上述方法得到初始候选实体列表，对初始候选实体列表进行筛选得到候选实体列表，方法为：

[0026] 抽取知识库中的部分实体做为训练集，利用标注的<实体，类型>对集合训练实体-类型分类器，然后利用实体-类型分类器识别候选实体列表中实体的类型，把与查询文本中待链接实体属于同类型的实体保留，删除类型不同的实体，得到最终的候选实体列表；

[0027] 所述实体的类型为实体的类别属性，包括人物、地点、机构、时间；所述<实体，类型>对为实体与其类型的映射，其中映射正确的<实体，类型>对为训练分类器的正向实例，反之为负向实例；所述实体-类型分类器为通过<实体，类型>对集合训练的多类分类模型，通过所述模型可以判断实体的类型。

[0028] 所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤3中，利用搜索引擎，扩展实体指称的上下文信息的方法为：

[0029] 将步骤1中查询文本中一组命名实体的实体指称做为种子在搜索引擎中检索，得到与所述实体指称相关的搜索结果网页，从搜索结果网页中抽取前m条搜索结果的标题和摘要信息，得到所述实体指称的扩展文本，查询文本与扩展文本一起构成所述实体指称的扩展的上下文信息。

[0030] 所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤4中，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息的方法为：

[0031] 将候选实体列表中的每个候选实体在知识库中的上下文文本进行分词，对分词后的文本进行命名实体识别得到一组与所述候选实体相关的命名实体，然后把这一组命名实体做为一个种子在搜索引擎中检索，得到与所述候选实体相关的搜索结果网页，从搜索结果网页中抽取前n条搜索结果的标题和摘要信息，得到所述候选实体的扩展文本，候选实体在知识库中的上下文文本与扩展文本一起构成所述候选实体的扩展的上下文信息。

[0032] 所述的一种基于搜索引擎的命名实体链接方法，其特征在于：所述步骤5中，计算实体指称与每个候选实体之间的匹配度的方法为：

[0033] 对候选实体列表中的每个候选实体，计算所述实体指称的扩展的上下文信息和所述候选实体的扩展的上下文信息之间的余弦相似度a，计算所述候选实体流行度b，计算实体指称与每个候选实体之间的匹配度$c=w_a\times a+w_b\times b$，其中：$w_a$、$w_b$为预设的权值，且$w_a+w_b=1$；所述候选实体的流行度为：实体指称做为超链接链接到所述候选实体页面的次数与实体指称做为超链接链接到所有候选实体的总次数的比值。

[0034] 与已有技术相比，本发明的有益效果体现在：

[0035] 1、本发明采用了两阶段产生候选实体列表的方法。首先，计算字符串相似度得到初始候选实体列表。然后，根据对实体类型分类的方法，对初始候选实体列表进行筛选得到候选实体列表。两阶段产生候选实体列表的方法既可产生较多的与实体指称相关的候选实体，又可通过实体类型分类去除部分噪音候选实体以降低消歧的复杂度。

[0036] 2、本发明关键的贡献在于提供了对实体指称和候选实体的上下文文本信息进行扩展以解决命名实体链接问题的方法。实体指称上下文文本长度是有限的，知识库中包含的实体上下文信息也是有限的，从而导致依赖于词共现信息计算实体指称与候选实体上下文相似度的准确性不高。通过搜索搜索引擎检索实体指称和候选实体，得到有关实体指称和候选实体的所有网页，对网页进行整理后得到扩展的上下文信息，有助于我们提取更多的实体特征，从而为提高实体链接的准确性提供附加信息。

附图说明

[0037] 图1为本发明一种基于搜索引擎的命名实体链接方法的流程图。

具体实施方式

[0038] 一种基于搜索引擎的命名实体链接方法包括下述步骤：

[0039] 步骤1，对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体。

[0040] 步骤2，对每个命名实体，通过实体指称得到候选实体列表。

[0041] 步骤3，利用搜索引擎，扩展实体指称的上下文信息。

[0042] 步骤4，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息。

[0043] 步骤5，计算实体指称与每个候选实体之间的匹配度。

[0044] 步骤6，根据候选实体的匹配度对候选实体排序，选择匹配度最大的候选实体做为目标实体进行链接。

[0045] 步骤1中，命名实体为人名、机构名、地名以及其他所有以名称为标识的实体，实体为现实生活中的对象或概念；查询文本为用户输入的文本或文本集，包括实体指称和其上下文文本，其文本长度为固定长度，知识库为若干实体条目的集合，包括人物、机构、地名以及其他所有以名称为标识的实体；实体条目为每个实体在知识库中存储的具体信息，包括实体的上下文文本、实体属性和属性值，实体的上下文信息为对实体进行详细描述的文本，实体指称为查询文本中待链接实体的名词引用。

[0046] 步骤2中，通过实体指称得到候选实体的方法为：依次取出知识库中的每条实体条目，遍历其名称属性值、别名属性值，计算查询文本中的实体指称与各个实体条目的名称属性值、别名属性值之间的字符串相似度。若查询文本的实体指称与实体条目的名称属性值相似度大于预先设定好的阈值，则把当前实体条目输出到候选实体列表中，若查询文本的实体指称与实体条目的别名属性值相似度大于预先设定好的阈值，则把别名属性所对应的实体条目输出到候选实体列表中；通过上述方法得到初始候选实体列表。对初始候选实体列表进行筛选得到候选实体列表。方法为：抽取知识库中的部分实体做为训练集，利用标注的<实体，类型>对集合训练实体-类型分类器。然后利用实体-类型分类器识别候选实体列表中实体的类型，把与查询文本中待链接实体属于同类型的实体保留，删除类型不同的实体，得到最终的候选实体列表，实体的类型为实体的类别属性，包括人物、地点、机构、时间；<实体，类型>对为实体与其类型的映射，其中映射正确的<实体，类型>对为训练分类器的正向实例，反之为负向实例；实体-类型分类器为通过<实体，类型>对集合训练的多类分类模型，多类分类模型是把实体分成多个类别，一个实体属于且只属于多个类别中的一个，不同类之间是互斥的，通过多类分类模型可以判断实体的类型。

[0047] 步骤3中，利用搜索引擎，扩展实体指称的上下文信息的方法为：将步骤1中查询文本中一组命名实体的实体指称做为种子在搜索引擎中检索，得到与所述实体指称相关的搜索结果网页，从搜索结果网页中抽取前m条搜索结果的标题和摘要信息，得到所述实体指称的扩展文本，查询文本与扩展文本一起构成所述实体指称的扩展的上下文信息。

[0048] 步骤4中，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息的方法为：将候选实体列表中的每个候选实体在知识库中的上下文文本进行分词，对分词后的文本进行命名实体识别得到一组与所述候选实体相关的命名实体，然后把这一组命名实体做为一个种子在搜索引擎中检索，得到与所述候选实体相关的搜索结果网页，从搜索结果网

页中抽取前n条搜索结果的标题和摘要信息，得到所述候选实体的扩展文本，候选实体在知识库中的上下文文本与扩展文本一起构成所述候选实体的扩展的上下文信息。

[0049] 步骤5中，计算实体指称与每个候选实体之间的匹配度的方法为：对候选实体列表中的每个候选实体，计算实体指称的扩展的上下文信息和所述候选实体的扩展的上下文信息之间的余弦相似度a，计算所述候选实体流行度b，计算实体指称与每个候选实体之间的匹配度$c=w_a\times a+w_b\times b$。其中：$w_a$、$w_b$为预设的权值，且$w_a+w_b=1$。候选实体的流行度为：实体指称做为超链接链接到所述候选实体页面的次数与实体指称做为超链接链接到所有候选实体的总次数的比值。

[0050] 具体实施例：

[0051] 本实施例提供了一种基于搜索引擎的命名实体链接方法，下面结合图1说明本实施例中基于搜索引擎的命名实体链接方法的步骤：

[0052] (1)如图1的S101所示，对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体。一名用户输入一段查询文本，例如某新闻里的一段话“李娜退役仪式纳达尔献花，娜姐泪奔享全场欢呼。”，首先，对输入的查询文本进行分词，然后利用命名实体识别工具识别出文本中的一组命名实体。其中命名实体就是人名、机构名、地名以及其他所有以名称为标识的实体，更广泛的实体还包括数字、日期、货币、地址。

[0053] 对输入的查询文本进行分词是利用分词工具将查询文本按照中文词典进行词的划分，本实施例中分词所用的工具是由中国科学院计算技术研究所提供的NLPIR中文分词工具(网址：http://ictclas.nlpir.org/downloads)。分词得到的词和其类型标注的集合为：W＝{李娜/nr退役/vi仪式/n纳达尔/nr献花/vi，/wd娜姐/nr泪/n奔/v享/vg全场/n欢呼/v}，其中“/nr”、“/vi”、“/n”、“/wd”、“/v”、“/vg”是词的类型标注。

[0054] 分词得到一组词的集合后，利用命名实体识别技术识别出分词后的文本所包含的实体集合：E＝{李娜、纳达尔、娜姐}，本实施例采用的命名实体识别方法为根据分词得到的词和其类型标注，提取类型标注为nr、ns、nt、nz的词组成命名实体集合。这里的“李娜”、“纳达尔”、“娜姐”就是识别的命名实体“李娜”、“纳达尔”、“娜姐”的实体指称。本实施例以实体指称“李娜”的链接过程为例介绍命名实体链接过程。

[0055] (2)如图1的S102所示，对每个命名实体，通过实体指称得到候选实体列表。依次取出知识库中的每条实体条目，遍历其名称属性值、别名属性值，计算查询文本中的实体指称与各个实体条目的名称属性值、别名属性值之间的字符串相似度。若查询文本的实体指称与实体条目的名称属性值相似度大于预先设定好的阈值，则把当前实体条目输出到候选实体列表中，若查询文本的实体指称与实体条目的别名属性值相似度大于预先设定好的阈值，则把别名属性所对应的实体条目输出到候选实体列表中。通过上述方法得到初始候选实体列表。本实施例中所使用的知识库是清华大学发布的中文知识库(网址：http://keg.cs.tsinghua.edu.cn/project/ChineseKB)，其中包含800,000不同的实体及其属性，名称属性值包含在<rdf:Description rdf:about＝""></rdf:Description>标签的rdf:about属性中，别名属性包含在<ont:别名></ont:别名>标签中。本实施例预先设定阈值为0.7。例如，搜索李娜这个实体指称得到的满足阈值的初始候选实体条目如表1所示：

[0056] 表1实体指称“李娜”的初始候选实体列表

[0057]

实体名称	实体上下文	其他属性
李娜（网球运动员）	李娜（1982 年 2 月 26 日—），中国女子网球选手。2011 年 6 月 4 日，在法国的巴黎西部蒙特高地的罗兰·加洛斯体育场内，李娜获得法网女单冠军。成为有史以来第一个获得大满贯网球赛事冠军的亚洲人。	……
李娜（歌手）	2005 超级女声广州赛区的季军李娜，嘉年华里的双料明星。出演了由嘻哈包袱铺掌柜高晓攀编剧和执导的音乐剧《爱·疯了》，并在剧中深情演唱多首经典歌曲。	……
李娜（北大教授）	李娜，女，博士，北京大学化学与分子工程学院副教授、硕士生导师。	……
李娜效应	李娜效应是指 2011 年李娜法网捧杯而引发的一系列蝴蝶效应，波及的远不止体育，还有地理、文化、外交等等。	……
李娜英	李娜英（1979 年 2 月 22 日——）女，韩国影视演员。毕业于新丘大学经济系，1998 年在 CF 中初露头角后气质绝佳的她一直都是广告界的宠儿。2011 年 3 月，有媒体报道称李娜英将与裴勇俊 4 月完婚，当事人双方皆否认此消息。	……

[0058] 表1中，实体名称为提取的知识库中实体条目的名称属性值，其包含在<rdf:Description rdf:about=""></rdf:Description>标签的rdf:about属性中，实体上下文为提取的知识库中实体条目的上下文属性值，其包含在<ont:ABSTRACT></ont:ABSTRACT>标签中，其他属性为实体在知识库中实体条目的其他属性值，包含在除<rdf:Description rdf:about=""></rdf:Description>、<ont:ABSTRACT></ont:ABSTRACT>标签的其他标签中。

[0059] 然后，对初始候选实体列表进行筛选得到候选实体列表，方法为：抽取知识库中的部分实体做为训练集，利用标注的“<实体，类型>对”集合训练“实体-类型”分类器。本实施例抽取知识库中部分实体做为训练集的方法为：从每个类型的实体条目集合中随机抽取200条实体条目做为训练集，所采用的分类器模型为SVM(Support Vector Machine)。然后利用“实体-类型”分类器识别候选实体列表中实体的类型，把与查询文本中待链接实体属于同类型的实体保留，删除类型不同的实体，得到最终的候选实体列表。实体类型为实体的类别属性，包括人物、地点、机构、时间。筛选后的候选实体列表如表2所示：

[0060] 表2实体指称“李娜”的候选实体列表

[0061]

实体名称	实体上下文	其他属性
李娜（网球运动员）	李娜（1982 年 2 月 26 日—），中国女子网球选手。2011 年 6 月 4 日，在法国的巴黎西部蒙特高地的罗兰·加洛斯体育场内，李娜获得法网女单冠军。成为有史以来第一个获得大满贯网球赛事冠军的亚洲人。	……
李娜（歌手）	2005 超级女声广州赛区的季军李娜，嘉年华里的双料明星。出演了由嘻哈包袱铺掌柜高晓攀编剧和执导的音乐剧《爱·疯了》，并在剧中深情演唱多首经典歌曲。	……
李娜（北大教授）	李娜，女，博士，北京大学化学与分子工程学院副教授、硕士生导师。	……
李娜英	李娜英（1979 年 2 月 22 日——）女，韩国影视演员。毕业于新丘大学经济系，1998 年在 CF 中初露头角后气质绝佳的她一直都是广告界的宠儿。2011 年 3 月，有媒体报道称李娜英将与裴勇俊 4 月完婚，当事人双方皆否认此消息。	……

[0062] (3)如图1的S103所示，利用搜索引擎，扩展实体指称的上下文信息。以实体指称“李娜”为例，该实体指称的查询文本为“李娜退役仪式纳达尔献花，娜姐泪奔享全场欢呼。”，将从查询文本中识别的一组命名实体的实体指称“李娜、纳达尔、娜姐”做为搜索引擎的种子进行搜索，得到与实体指称“李娜”相关的搜索结果网页，从搜索结果网页中抽取前m条搜索结果的标题和摘要信息，得到如表3所示的实体指称“李娜”的扩展文本，查询文本与扩展文本一起构成实体指称“李娜”的扩展的上下文信息。本实施例中搜索引擎为必应搜索引擎，使用CSS选择器抽取搜索结果的标题和摘要信息，m取值为8。

[0063] 表3实体指称“李娜”的扩展文本

[0064]

序号	实体指称“李娜”的扩展文本	
1	标题	李娜退役仪式纳达尔献花 娜姐泪奔享全场欢呼_网易体育
	摘要	李娜的退役仪式在中网现场举行。WTA 主席阿拉斯特女士和李娜好朋友科维托娃先后发言并都盛赞李娜对中国、亚洲乃至整个女子网坛的影响和贡献，而西班牙天王纳达尔与 ...

[0065]

2	标题	李娜宣布正式退役 李娜法网夺冠赛程-知音女性网
	摘要	李娜退役仪式纳达尔献花 娜姐泪奔享全场欢呼 北京时间周二晚上，李娜的退役仪式在中网现场举行。WTA 主席阿拉斯特女士和李娜好朋友科维托娃先后发言并都 ...
3	标题	李娜正式宣布退役 - 网易体育 有态度的体育门户
	摘要	(2014-09-30) 神秘嘉宾纳达尔献花 姜山深情拥抱娜姐赢欢呼 (2014-09-30) 李娜退役仪式纳达尔献花 娜姐泪奔享全场 欢呼 (2014-09-30) 李娜给粉丝送视频:感谢球迷 你们为我 ...
4	标题	李娜 - 国搜百科
	摘要	在李娜的退役仪式结束后，将会是两场大满贯冠军分别出战的 WTA 和 ATP ... 网易体育：李娜退役仪式纳达尔献花 娜姐泪奔享全场 欢呼 [2]. 荆楚网：李娜简介 [3]. ...
5	标题	李娜宣布正式退役-搜狐体育
	摘要	李娜中网退役仪式纳达尔献花 群星祝福 直击-钻石球场李娜"独自退场" 球迷含泪告别传奇落幕 感言-李娜：很高兴在中网说再见 感谢球迷一直的陪伴 ...
6	标题	中网举行李娜退役仪式 自制 MV 感谢球迷_体育_腾讯网
	摘要	中网举行李娜退役仪式 自制 MV 感谢球迷 昨天晚上，李娜退役仪式在国家网球中心钻石球场举行。娜姐也特意亲自制作了 MV 感谢多年来一直支持她的娜离子们。
7	标题	李娜传记电影 2016 上映 曝陈可辛敲定赵薇主演- Micro …
	摘要	李娜致意到场人士现场泪崩李娜哽咽纳达尔献花姜山拥抱李娜李娜抹泪李娜抹泪李娜鼓掌李娜鼓掌热泪盈眶李娜热泪盈眶李娜热泪盈眶李娜发言李娜致意到场观众 ...
8	标题	李娜退役仪式昨晚举行 点点荧光照亮整个球场-搜狐体育
	摘要	李娜退役仪式昨晚举行本报讯（记者钱晞）9 月 30 日晚，中国网球金花李娜的退役仪式在中国网球公开赛钻石球场举行。仪式中李娜深情望向全场高呼自己名字的 ...

[0066] (4)如图1的S104所示，利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息。首先，将候选实体列表中的每个候选实体在知识库中的上下文文本进行分词，本实施例中，候选实体的上下文信息包含在<ont:ABSTRACT></ont:ABSTRACT>标签中。以候选实体“李娜(网球运动员)”为例，对其上下文分词的结果为W^*={李娜/nr(/wkz 1982年/t 2月/t 26日/t—/wp)/wky,/wd中国/nr 女子/n 网球/n 选手/n。/wj 2011年/t 6月/t 4日/t,/wd在/p法国/nr的/ude1 巴黎/nr 西部/f 蒙/vi 特/d 高/a 地/n 的/ude1罗兰/nrf·加/b 洛/b 斯/b 体育场/n 内/f,/wd李娜/nr 获得/v 法网/n 女单/n 冠军/n。/wj成为/v 有史以来/d1第一/m 个/q 获得/v 大/a 满贯/n 网球赛/n 事/n 冠军/n 的/ude1亚洲/nr 人/n。/wj}，对分词后的文本进行命名实体识别得到一组与候选实体相关的命名实体。本实施例中采用的命名实体识别方法为根据分词得到的词和其类型标注，提取类型标注为nr、ns、nt、nz的词组成命名实体集合。命名实体识别的结果为E^*={李娜，中国，法国，巴黎，亚洲}。然后，将这一组命名实体“李娜、中国、法国、巴黎、亚洲”做为一个种子在搜索引擎中检索，得到一组与候选实体“李娜(网球运动员)”相关的搜索结果网页，从搜索结果网页中抽取前n条搜索结果的标题和摘要信息，得到如表4所示的候选实体“李娜(网球运动员)”的扩展文本，候选实体“李娜(网球运动员)”在知识库中的上下文文本与扩展文本一起构成候选实体“李娜(网球运动员)”的扩展的上下文信息。本实施例中搜索引擎为必应搜索引擎，使用CSS选择器抽取搜索结果的标题和摘要信息，n取值为10。

[0067] 表4候选实体“李娜(网球运动员)”的扩展文本

[0068]

序号	候选实体的扩展文本	
1	标题	李娜（中国女子网球名将）_百度百科
	摘要	李娜，1982 年 2 月 26 日出生于湖北省武汉市，中国女子网球运动员。2008 年北京奥运会女子单打第四名，2011 年法国网球公开赛、2014 年澳大利亚网球公开赛女子单打冠军 ...
2	标题	李娜将历史踩脚下 法网折桂亚洲人首捧单打大满贯 ...-新 …
	摘要	新浪体育讯 北京时间 6 月 4 日(巴黎当地时间 4 日)消息，在备受国人关注的本年度法国网球公开赛的女单决赛中，中国金花李娜(李娜新浪微博(@李娜))在拿下首盘 ...
3	标题	中国首位大满贯得主 亚洲网坛第一人李娜_体育中国_中国网
	摘要	中国首位大满贯得主 亚洲网坛第一人李娜 中国网 china.com.cn 时间： 2012-07-23 19:25 责任编辑: 蔚刚强 ... 5. 2011 年 6 月 4 日罗兰加洛斯，巴黎，法国 红土 弗朗切斯卡 ...
4	标题	李娜法网夺冠 创亚洲历史书写多项新纪录【图】(2)_体 …
	摘要	北京时间 6 月 4 日(巴黎当地时间 4 日)消息，在备受国人关注的本年度法国网球公开赛的女单决赛中，中国金花李娜在拿下首盘后，又在第二盘末段成功顶住了卫冕冠军 ...
5	标题	李娜获 2011 年法国网球公开赛冠军 中国、亚洲选手第一 …
	摘要	2011-6-6 • 李娜获 2011 年法国网球公开赛冠军 中国、亚洲选手第一人 收 1138 万巨额支票,li-han163 的网易博客,策马淮海，指点江山！，
6	标题	李娜_百度百科
	摘要	夺得了中国乃至亚洲的第一座大满贯单打冠军奖杯。6 月 24。李娜在温网比赛中晋级至 32 ... 李娜在巴黎加入劳伦斯世界体育学会。成为劳伦斯世界体育学会第 60 ...
7	标题	梦想成真! 亚洲中国女选手李娜首度荣登法网大满贯女单 …
	摘要	2011-6-4 • 梦想成真! 亚洲中国女选手李娜首度荣登法网大满贯女单冠军！(图文) 李娜 法网 大满贯 女单冠军 网坛 体育 roland garros 当五星红旗冉冉升起,中华人民共和国国歌奏 ...
8	标题	李娜入选劳伦斯学会:全世界仅 60 人 她是中国第 5 个_新浪 …
	摘要	李娜入选劳伦斯学会:全世界仅 60 人 她是中国第 5 个,昨天，法国巴黎，劳伦斯世界体育学会官方宣布，已经退役的中国网球名将李娜正式加入该学会，成为第 60 名 ...
9	标题	李娜夺冠 2011 赛季法国网球公开赛 亚洲首个大满贯赛 …
	摘要	李娜夺冠 2011 赛季法国网球公开赛 亚洲首个大满贯赛事是体育类高清视频，画面清晰，播放流畅，发布时间：2013-10-02。视频

[0069]

		简介：李娜 中国网球。
10	标题	李娜还有梦_李娜还有梦的故事-读者
	摘要	繁体网提供读者之李娜还有梦的故事，6 月 4 日北京时间 21 时，1 亿多中国人在电视机前，屏息注视着从法国巴黎传来的画面。一个矫健的身影在画面中跑动跳跃，人们的心情 ...

[0070] (5)如图1的S105所示，计算实体指称与每个候选实体之间的匹配度。首先对候选实体列表中的每个候选实体，计算实体指称的扩展的上下文信息和候选实体的扩展的上下文信息之间的余弦相似度a，计算候选实体流行度b，计算实体指称与每个候选实体之间的匹配度$c=w_a\times a+w_b\times b$。其中：$w_a$、$w_b$为预设的权值，且$w_a+w_b=1$。候选实体的流行度为：实体指称做为超链接链接到所述候选实体页面的次数与实体指称做为超链接链接到所有候选

实体的总次数的比值。本实施例中预设w_a=0.6,w_b=0.4。以步骤(3)中计算实体指称“李娜”与候选实体“李娜(网球运动员)”的匹配度为例,实体指称“李娜”的扩展的上下文信息与候选实体“李娜(网球运动员)”的扩展的上下文信息之间的余弦相似度a=0.53,候选实体“李娜(网球运动员)”的流行度b=0.88,因此,实体指称“李娜”与候选实体“李娜(网球运动员)”的匹配度c=0.6*0.53+0.4*0.88=0.67。表5列出了实体指称“李娜”与每个候选实体之间的匹配度。

[0071] 表5实体指称“李娜”与每个候选实体之间的匹配度

[0072]

候选实体	余弦相似度	候选实体的流行度	匹配度
李娜(网球运动员)	0.53	0.88	0.67
李娜(歌手)	0.43	0.056	0.28
李娜(北大教授)	0.39	0	0.23
李娜英	0.19	0.31	0.24

[0073] (6)如图1的S106所示,根据候选实体的匹配度对候选实体排序,选择匹配度最大的候选实体做为目标实体进行链接。本实施例根据实体指称与候选实体之间的匹配度对表5进行自大至小的排序得到表6,表6第1条中的候选实体即为匹配度最大的候选实体,由此可得实体指称“李娜”链接的目标实体为“李娜(网球运动员)”。

[0074] 表6按匹配度自大至小对实体指称“李娜”的候选实体的排序结果

[0075]

候选实体	余弦相似度	候选实体的流行度	匹配度
李娜(网球运动员)	0.53	0.88	0.67
李娜(歌手)	0.43	0.056	0.28
李娜英	0.19	0.31	0.24
李娜(北大教授)	0.39	0	0.23

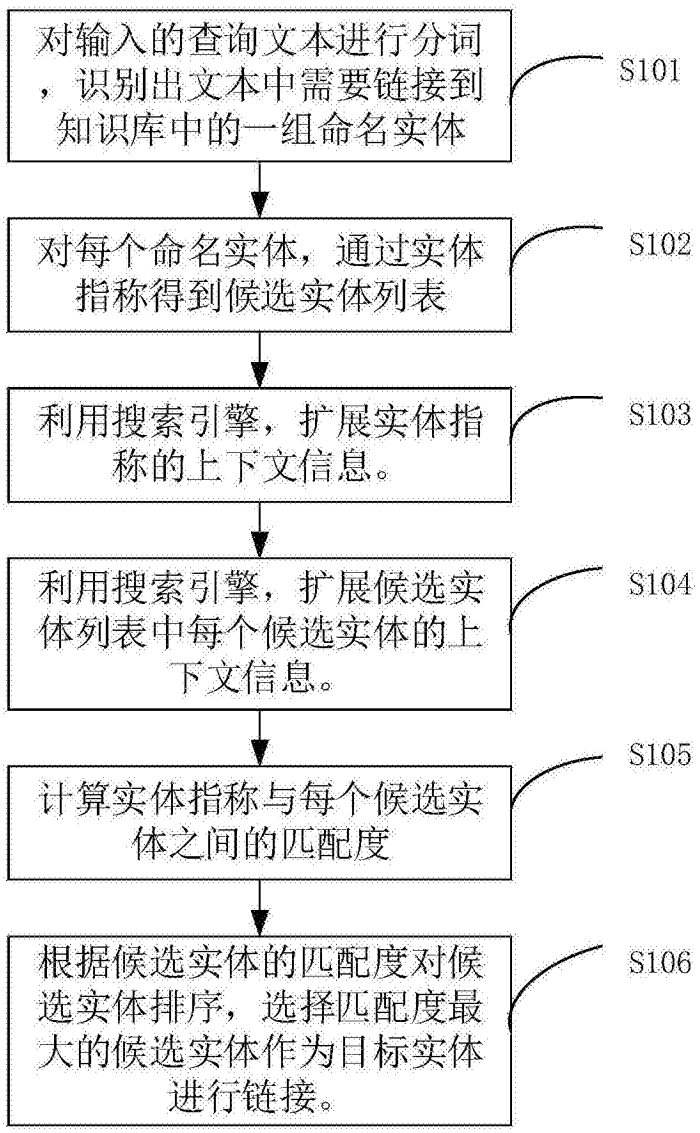

对输入的查询文本进行分词，识别出文本中需要链接到知识库中的一组命名实体
S101
对每个命名实体，通过实体指称得到候选实体列表
S102
利用搜索引擎，扩展实体指称的上下文信息。
S103
利用搜索引擎，扩展候选实体列表中每个候选实体的上下文信息。
S104
计算实体指称与每个候选实体之间的匹配度
S105
根据候选实体的匹配度对候选实体排序，选择匹配度最大的候选实体作为目标实体进行链接。
S106

图1