



(12) 发明专利申请

(10) 申请公布号 CN 104133916 A

(43) 申请公布日 2014. 11. 05

(21) 申请号 201410400557. 4

(22) 申请日 2014. 08. 14

(71) 申请人 百度在线网络技术(北京)有限公司

地址 100085 北京市海淀区上地十街 10 号

百度大厦三层

(72) 发明人 吴先超

(74) 专利代理机构 北京品源专利代理有限公司

11332

代理人 胡彬 路凯

(51) Int. Cl.

G06F 17/30(2006. 01)

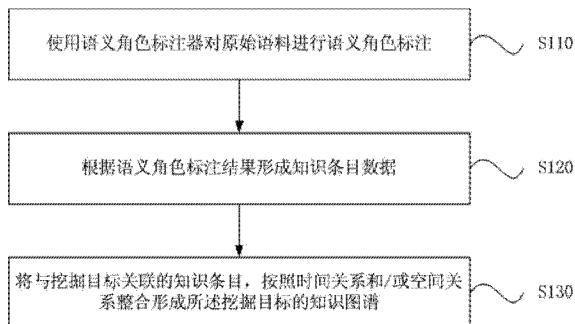
权利要求书2页 说明书9页 附图6页

(54) 发明名称

搜索结果信息组织方法和装置

(57) 摘要

本发明实施例公开了一种搜索结果信息组织方法和装置。所述搜索结果信息组织方法包括：使用语义角色标注器对原始语料进行语义角色标注；根据语义角色标注结果形成知识条目；将与挖掘目标关联的知识条目，按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。本发明实施例提供的搜索结果信息组织方法和装置能够方便用户从搜索引擎的搜索结果中获得有效信息。



1. 一种搜索结果信息组织方法,其特征在于,包括:
使用语义角色标注器对原始语料进行语义角色标注;
根据语义角色标注结果形成知识条目;
将与挖掘目标关联的知识条目,按照时间关系和 / 或空间关系整合形成所述挖掘目标的知识图谱。

2. 根据权利要求 1 所述的方法,其特征在于,使用语义角色标注器对原始语料进行语义角色标注包括:

利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词;

通过命名实体识别 NER 获取所述原始语料中的其他句子成分。

3. 根据权利要求 2 所述的方法,其特征在于,在利用预先训练的语料库对从互联网获取的所述原始语料进行语义分析之前,还包括:

使用预先生成的深层神经网络 DNN 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词;

将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料。

4. 根据权利要求 3 所述的方法,其特征在于,根据语义角色标注结果形成知识条目包括:

利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目。

5. 根据权利要求 2 所述的方法,其特征在于,在获取所述原始语料中的谓词之后,还包括:

对谓词标注语义分类,并将原始语料按照谓词的分类结果进行聚类整合。

6. 根据权利要求 1-5 任一所述的方法,其特征在于,所述挖掘目标为人名姓名或组织名称,则将与挖掘目标关联的知识条目,按照时间关系整合形成所述挖掘目标的知识图谱包括下述至少一项:

在所述知识条目中进行查询获取包括所述挖掘目标的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的履历图谱;

在所述知识条目中进行查询获取包括所述挖掘目标的关联人物或关联组织的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的关联人物图谱或关联组织图谱;

在所述知识条目中进行查询获取包括所述挖掘目标,且发生时间在设定时间范围内的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的新闻图谱。

7. 根据权利要求 1-4 任一所述的方法,其特征在于,在形成所述挖掘目标的知识图谱之后,还包括:

在用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。

8. 一种搜索结果信息组织装置,其特征在于,包括:

语义角色标注模块,用于使用语义角色标注器对原始语料进行语义角色标注;

知识条目形成模块,用于根据语义角色标注结果形成知识条目;

知识图谱形成模块,用于将与挖掘目标关联的知识条目,按照时间关系和 / 或空间关系整合形成所述挖掘目标的知识图谱。

9. 根据权利要求 8 所述的装置,其特征在于,所述语义角色标注模块包括:

语义分析单元,用于利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词;

命名实体识别单元,用于通过命名实体识别 NER 获取所述原始语料中的其他句子成分。

10. 根据权利要求 9 所述的装置,其特征在于,所述语义角色标注模块还包括:

关键字泛化单元,用于在利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词之前,使用预先生成的深层神经网络 DNN 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词;

语料扩展搜索单元,用于将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料。

11. 根据权利要求 10 所述的装置,其特征在于,所述知识条目形成模块包括:

归一化单元,用于利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目。

12. 根据权利要求 9 所述的装置,其特征在于,所述语义角色标注模块还包括:

谓词语义分类单元,用于在获取所述原始语料中的谓词之后,对谓词标注语义分类,并将原始语料按照谓词的分类结果进行聚类整合。

13. 根据权利要求 8-12 任一所述的装置,其特征在于,所述知识图谱形成模块包括:

履历图谱获取单元,用于在所述知识条目中进行查询获取包括所述挖掘目标的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的履历图谱;

关联人物 / 组织图谱获取单元,用于在所述知识条目中进行查询获取包括所述挖掘目标的关联人物或关联组织的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的关联人物图谱或关联组织图谱;和 / 或

新闻图谱获取单元,用于在所述知识条目中进行查询获取包括所述挖掘目标,且发生在设定时间范围内的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的新闻图谱。

14. 根据权利要求 8-11 任一所述的装置,其特征在于,还包括:

图谱显示模块,用于在形成所述挖掘目标的知识图谱之后,在用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。

搜索结果信息组织方法和装置

技术领域

[0001] 本发明实施例涉及搜索引擎技术领域,尤其涉及一种搜索结果信息组织方法和装置。

背景技术

[0002] 近年来,由于网络和信息技术的快速发展,互联网上的信息呈爆炸式的增长。网络信息资源日益丰富。根据中国互联网信息中心 2010 年 7 月发布的《中国互联网络发展状况统计报告》第 26 次调查统计报告显示,2010 年互联网上的网页数量达到 336 亿个。用户面对如此海量的信息,利用搜索引擎进行信息检索已成为一个重要途径。

[0003] 现有的搜索引擎接收到用户的一个检索词以后,将与该检索词相关的网页链接和网页的主要内容以列表的方式一一列出。通过这种方式展现给用户的信息涵盖面比较广。但是,展现给用户的搜索结果之间没有任何的联系,用户需要对搜索结果进行进一步的人工筛选才能了解到关于检索词的有效信息。

发明内容

[0004] 有鉴于此,本发明实施例提出一种搜索结果信息组织方法和装置,以方便用户从搜索引擎的搜索结果中获得有效信息。

[0005] 第一方面,本发明实施例提供了一种搜索结果信息组织方法,所述方法包括:

[0006] 使用语义角色标注器对原始语料进行语义角色标注;

[0007] 根据语义角色标注结果形成知识条目;

[0008] 将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。

[0009] 第二方面,本发明实施例提供了一种搜索结果信息组织装置,所述装置包括:

[0010] 语义角色标注模块,用于使用语义角色标注器对原始语料进行语义角色标注;

[0011] 知识条目形成模块,用于根据语义角色标注结果形成知识条目;

[0012] 知识图谱形成模块,用于将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。

[0013] 本发明实施例提供的搜索结果信息组织方法和装置,通过使用语义角色标注器对原始语料进行语义角色标注,根据语义角色标注结果形成知识条目,将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱,从而将语义角色标注技术应用于对搜索引擎的搜索结果的组织,使得能够以知识图谱的形式向用户展示关于挖掘目标的信息,方便了用户从搜索引擎的搜索结果中获得有效信息。

附图说明

[0014] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述,本发明的其它特征、目的和优点将会变得更明显:

- [0015] 图 1 是本发明第一实施例提供的搜索结果信息组织方法的流程图；
- [0016] 图 2 是本发明第一实施例提供的知识条目的示意图；
- [0017] 图 3 是本发明第一实施例提供的搜索结果信息组织方法中知识图谱形成的流程图；
- [0018] 图 4a 是本发明第一实施例提供的履历图谱的示意图；
- [0019] 图 4b 是本发明第一实施例提供的关联人物图谱的示意图；
- [0020] 图 4c 是本发明第一实施例提供的新闻图谱的示意图；
- [0021] 图 5 是本发明第二实施例提供的搜索结果信息组织方法中知识图谱形成的流程图；
- [0022] 图 6 是本发明第三实施例提供的搜索结果信息组织方法中知识图谱形成的流程图；
- [0023] 图 7 是本发明第四实施例提供的搜索结果信息组织方法中知识图谱形成的流程图；
- [0024] 图 8 是本发明第五实施例提供的搜索结果信息组织方法的流程图；
- [0025] 图 9 是本发明第六实施例提供的搜索结果信息组织装置的结构图。

具体实施方式

[0026] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅用于解释本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部内容。

[0027] 图 1 至图 4 示出了本发明的第一实施例。

[0028] 图 1 是本发明第一实施例提供的搜索结果信息组织方法的流程图。所述搜索结果信息组织方法由搜索结果信息组织装置执行。所述搜索结果信息组织装置被部署在搜索引擎的服务器上。参见图 1,所述搜索结果信息组织方法包括:

[0029] S110,使用语义角色标注器对原始语料进行语义角色标注。

[0030] 所述原始语料是关于所述挖掘目标的原始语料。所述挖掘目标是进行信息组织的目标短语,也是在搜索引擎中经常被搜索,需要了解其历史或当前信息的目标,它通常是一个名词性的短语。所述挖掘目标可以是人名,比如,张亚勤,或者哈梅斯·罗德里格斯。所述挖掘目标还可以是地名,比如,班加西。所述挖掘目标还可以是品牌名称,比如,阿迪达斯。所述挖掘目标还可以是组织名称,比如,世界知识产权组织。

[0031] 优选的,所述原始语料可以通过搜索互联网而获得。所述原始语料通常是以完整的句子为单位进行获取和存储的。在获取和存储操作以后,所述原始语料也是以句子为单位进行分析处理的。在通过搜索互联网获得所述原始语料的过程中,对句子在网页中的起始位置和终止位置的判断是依赖于对网页文字中的标点符号的识别而进行的。

[0032] 所述语义角色标注器是根据语义角色标注技术构建的一个语料标注工具,它能够对输入的原始语料进行语义角色标注,并将标注好的语料数据输出。

[0033] 在对所述原始语料进行语义角色标注的过程中,将能分派语义角色的动词称为谓词,而将从谓词处接受语义角色的短语称为论元。在本实施例中,所述语义角色标注器采用管程(Pipeline)结构,即先对所述原始语料进行谓词识别,完成了谓词识别以后再根据谓

词识别的结果识别所述原始语料的论元。

[0034] 所述语义角色标注器对所述原始语料进行的语义角色标注是根据预先训练的语料库进行的。在本实施例中,预先建立了针对互联网上的中文文本的“中文命题库 (Chinese Propbank)”,并运用所述中文命题库对所述原始语料进行语义角色标注。相对于现有的语料库来说,所述中文命题库的特点在于,其语料数据的规模更大,并且对不同的话题领域更具有针对性。

[0035] 完成对原始语料的语义角色标注以后,所述原始语料中的各个句子成分就被区别开来,并且每个句子成分都被标注了其在句子中的具体的句子成分参数。例如,对原始语料“克里斯蒂娜昨天用棒球打了斯科特”的语义角色标注结果是:

[0036] 克里斯蒂娜 —— 主语

[0037] 昨天 —— 时间状语

[0038] 用棒球 —— 方式状语

[0039] 打了 —— 谓词

[0040] 斯科特 —— 宾语。

[0041] S120,根据语义角色标注结果形成知识条目。

[0042] 由于在对所述原始语料进行语义角色标注之后,所述原始语料已经被分成单独的句子成分,并且被标注了成分属性,所以可以按照不同的句子成分在原始语料中的成分属性根据所述原始语料的语义角色标注结果形成成为所述知识条目。

[0043] 所述知识条目中记录了所述原始语料库中包括谓词和论元的句子主干信息。图2是本发明第一实施例提供的知识条目的示意图。参见图2,所述知识条目210包括时间状语211、谓词212以及所述挖掘目标213。所述谓词212表明了所述挖掘目标213在该时间状语的时间点上实施了何种动作;所述时间状语211表明了所述挖掘目标213实施该动作的时间。

[0044] 所述知识条目还可以包括地点状语214以及宾语。所述地点状语214表明了所述挖掘目标213实施所述动作的地点。所述宾语表明了所述挖掘目标213实施的动作的对象。

[0045] 优选的,根据所述语义角色标注结果形成知识条目的操作可以是将通过语义角色标注生成的标注有成分属性的句子成分按照其成分属性进行归类排列。

[0046] S130,将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。

[0047] 根据语义角色标注结果形成知识条目以后,按照所述知识条目之间的时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。优选的,可以按照所述知识条目中时间状语的时间顺序将所述知识条目整合成所述挖掘目标的知识图谱,还可以按照所述知识条目中地点状语的空间关系将所述知识条目整合成所述挖掘目标的知识图谱。

[0048] 优选的,将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱包括按照时间关系和/或空间关系整合形成所述挖掘目标的履历图谱、关联人物/组织图谱和/或新闻图谱。

[0049] 图3是本发明第一实施例提供的搜索结果信息组织方法中知识图谱形成的流程图。参见图3,优选的,所述挖掘目标为人名或组织名称,则将与挖掘目标关联的知识条目,按照时间关系整合形成所述挖掘目标的知识图谱包括下述至少一项:

[0050] S131,在所述知识条目中进行查询获取包括所述挖掘目标的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的履历图谱。

[0051] 所述履历图谱用于显示所述挖掘目标所经历的重大事件。以目标人物为例,所述挖掘目标的履历图谱就是显示该目标人物的生平履历的知识图谱。图 4a 是本发明第一实施例提供的履历图谱的示意图。参见图 4a,对目标人物“张亚勤”的履历图谱显示了目标人物“张亚勤”一生的主要经历,包括在哪些具体时间 411,进行了哪些具体的活动 412。

[0052] S132,在所述知识条目中进行查询获取包括所述挖掘目标的关联人物或关联组织的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的关联人物图谱或关联组织图谱。

[0053] 所述关联人物图谱用于显示与所述挖掘目标有关联的人物。对应的,所述关联组织图谱用于显示与所述挖掘目标有关联的组织。以目标人物的关联人物图谱为例,所述目标人物的关联人物图谱是用来显示所述目标人物的亲属、同学、同事等主要社会关系的知识图谱。图 4b 是本发明第一实施例提供的关联人物图谱的示意图。参见图 4b,对目标人物“张亚勤”的关联人物图谱包括与目标人物“张亚勤”相关联的人物的姓名 421,关联人物与目标人物之间的关系 422 以及关联人物与目标人物之间的这种关系建立的具体时间 423。

[0054] 优选的,可以选取对所述原始语料中包含除所述挖掘目标以外的其他人物或组织的语料的挖掘结果来生成所述关联人物图谱或关联组织图谱。

[0055] S133,在所述知识条目中进行查询获取包括所述挖掘目标,且发生时间在设定时间范围内的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的新闻图谱。

[0056] 所述新闻图谱用于显示距离当前时间较近的一段时间内与所述挖掘目标相关联的新闻事件。以目标人物为例,所述挖掘目标的新闻图谱是用来显示所述目标人物近期直接参与的新闻事件。图 4c 是本发明第一实施例提供的新闻图谱的示意图。参见图 4c,所述新闻图谱显示了所述目标人物“张亚勤”近期直接参与的新闻事件 431,以及这些新闻事件发生的具体时间 432。

[0057] 优选的,选取时间状语在设定时间范围内的知识条目,来生成所述新闻图谱。

[0058] 需要说明的是,图 4a 至图 4c 所示出的知识图谱都是按照时间顺序组织的知识图谱。但是,所述知识图谱也可以是按照空间顺序组织的知识图谱。例如,可以将具有相同地点状语的原始语料的语义角色标注结果整合成为知识图谱。

[0059] 本实施例通过使用语义角色标注器对原始语料进行语义角色标注,根据语义角色标注结果形成知识条目,并将与挖掘目标关联的知识条目,按照时间关系和 / 或空间关系整合形成所述挖掘目标的知识图谱,使得能够以知识图谱的形式向用户展示关于挖掘目标的信息,方便了用户从搜索引擎的搜索结果中获得有效信息。

[0060] 图 5 示出了本发明的第二实施例。

[0061] 图 5 是本发明第二实施例提供的搜索结果信息组织方法中知识图谱形成的流程图。所述搜索结果信息组织方法以本发明第一实施例为基础,进一步的,使用语义角色标注器对原始语料进行语义角色标注包括:利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词;通过命名实体识别(Named Entity Recognition, NER)获取所述原始语料中的其他句子成分。

[0062] 参见图 5,使用语义角色标注器对原始语料进行语义角色标注包括:

[0063] S111,利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词。

[0064] 在本实施例中,对所述原始语料进行的语义角色标注采用管程结构,即先获取所述原始语料中的谓词,再根据获取到的谓词对所述原始语料中的其他句子成分一一标注。

[0065] 在使用语义角色标注器对关于挖掘目标的原始语料进行语义角色标注的时候,对所述原始语料进行语义分析,并根据语义分析的结果从所述原始语料中获取谓词。所述语义分析是利用预先训练的语料库进行的语义分析。所述预先训练的语料库是预先训练的中文命题库。

[0066] S112,通过 NER 获取所述原始语料中的其他句子成分。

[0067] NER 是指识别文本中具有特定意义的实体。一般来说,命名实体识别识别的任务就是对于一篇待处理文本,识别出其中出现的人名 (person)、地名 (location)、机构名 (organization)、日期 (date)、时间 (time)、百分数 (percentage)、货币 (monetary value) 这七类命名实体。

[0068] 在通过语义分析获取所述原始语料中的谓词以后,对所述原始语料执行 NER,便可以获取所述原始语料中的其他句子成分。

[0069] NER 可以是基于隐马尔可夫模型的 NER、基于最大熵模型的 NER 或者基于 Agent 的 NER。

[0070] 采用管程结构及 NER 对所述原始语料进行语义角色标注的一个优点在于,能够自动的识别出所述原始语料中的各种句子成分,也就是句子中的谓词和各种论元。

[0071] 采用管程结构及 NER 对所述原始语料进行语义角色标注的另一个优点在于,语义角色标注的结果与所述原始语料的语序无关。比如,对于“警方正在详细调查事故原因”以及“警方正在对事故原因进行详细调查”这两个原始语料的语义角色标注结果是相同的。

[0072] 本实施例通过利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词,以及通过 NER 获取所述原始语料中的其他句子成分,使得搜索引擎能够自动识别原始语料中的各种句子成分,并且标注结果与原始语料的语序无关,提高了对原始语料的语义角色标注的标注精度。

[0073] 图 6 示出本发明的第三实施例。

[0074] 图 6 是本发明第三实施例提供的搜索结果信息组织方法中知识图谱形成的流程图。所述搜索结果信息的组织方法以本发明第二实施例为基础,进一步的,使用语义角色标注器对原始语料进行语义角色标注包括:在利用预先训练的语料库对从互联网获取的所述原始语料进行语义分析之前,使用预先生成的深层神经网络 (Deep neutral network, DNN) 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词;将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料。

[0075] 优选的,根据语义角色标注结果形成知识条目包括:利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目。

[0076] 参见图 6,使用语义角色标注器对原始语料进行语义角色标注包括:

[0077] S113,使用预先生成的 DNN 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词。

[0078] 所述原始语料中经常会出现形式不同的名词的指代对象相同的情况。对于目标人

物来说,同一个人物可能会有曾用名和现用名。对于目标组织来说,该组织的名称可能会被变更过。例如,“高通公司”、“高通股份有限公司”以及“夸尔柯姆股份有限公司”实际上指的是同一个组织。这样一来,如果仅仅使用所述挖掘目标的一个名字来获取所述原始语料将会造成获取的原始语料数据不全,也就是原始语料数据的数据稀疏。

[0079] 针对这种情况,使用 DNN 技术生成了 DNN 泛化词典。所述 DNN 泛化词典能够根据输入的名词输出语义与输入名词相同的其他名词。在获取所述原始语料之前,使用预先生成的 DNN 泛化词典将所述挖掘目标的名字泛化为关于所述挖掘目标的名字的至少一个泛化词,并将所述至少一个泛化词也作为所述挖掘目标的名字,以提高获取所述原始语料时的检全率。

[0080] S114,将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料。

[0081] 需要说明的是,在经过对所述词泛化处理以后,所述挖掘目标的关键字的数量为至少一个。利用数量大于一个的关键字来获取所述原始语料,能够增加获取到的原始语料的数量,从而克服原始语料数据稀疏的问题。

[0082] S115,利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词。

[0083] S116,通过 NER 获取所述原始语料中的其他句子成分。

[0084] 优选的,根据语义角色标注结果形成知识条目包括:

[0085] S121,利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目。

[0086] 由于在获取所述原始语料之前,对所述挖掘目标的关键字进行了泛化,并且在所述原始语料中对同一个事物会包括多种不同的表述方式,所以如果对形成的原始语料中的词语进行归一化处理,最终形成的知识图谱数据的对同一事物的表述会不统一,造成知识图谱数据的一致性较差。

[0087] 对于同一个事物的不同表达,可以采用所述 DNN 泛化词典进行归一化。利用所述 DNN 泛化词典对所述原始语料进行归一化以后,所述原始语料中对同一个事物会采用统一的表述方式。

[0088] 所述原始语料中还会出现一些指示代词,比如“他”、“他们”等等。对于这些指示代词,其所指代的对象一般会在上下文中明确的出现。因此,可以通过对原始语料的上下文进行语义分析,获取到这些指示代词的指代对象,以完成对所述指示代词的归一化。

[0089] 本实施例通过使用预先生成的 DNN 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词,再将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料,并在获取所述原始语料以后,利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目,克服了原始语料中数据稀疏的问题,并且进一步的保证了知识图谱的数据一致性。

[0090] 图 7 示出了本发明的第四实施例。

[0091] 图 7 是本发明第四实施例提供的搜索结果信息组织方法中知识图谱形成的流程图。所述搜索结果信息组织方法以本发明第二实施例为基础,进一步的,在获取所述原始语料中的谓词之后,使用语义角色标注器对原始语料进行语义角色标注还包括:对谓词标注

语义分类,并将原始语料按照谓词的分类结果进行聚类整合。

[0092] 参见图 7,使用语义角色标注器对原始语料进行语义角色标注包括:

[0093] S117,利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词。

[0094] S118,对谓词标注语义分类,并将原始语料按照谓词的分类结果进行聚类整合。

[0095] 从所述原始语料中获取到谓词以后,根据上下文对获取到的谓词进行语义分类。进行语义分类的目的在于明确在具体的原始语料中,所述谓词的真实含义。比如,谓词“打”有时有“购买”的含义,比如“打车票”,“打酒”。而在另外一些情况下,谓词“打”有“举”的含义,比如“打伞”,“打旗子”。为了区分谓词在不同语境下的不同含义,将同一个谓词的不同含义分别用不同的语义分类标记表示,并在从所述原始语料中获取到所述谓词以后,有不同的语义分类标记对所述谓词的语义进行分类。

[0096] 对所述谓词进行的语义分类就是使用不同的语义分类标记对所述谓词进行标注。比如,谓词“打”在所述原始语料中的含义是“购买”,而“购买”的语义对应的语义分类标记是①,则采用语义分类标记①对所述谓词“打”进行标注。

[0097] 优选的,可以参照预先建立的词对信息对所述谓词进行语义分类。在人们的日常用语中,谓词会与一些其他句子成分,比如宾语、主语或者补语形成一些固定的搭配,比如“打岔”,“打主意”。将这些包含谓词的固定搭配进行收集,并标明其中谓词的语义分类,形成词对信息。在对所述谓词进行语义分类时,检查所述谓词的上下文中是否出现所述词对信息中除所述谓词的其他成分,如果出现了所述词对信息中除所述谓词的其他成分,参考预先收集的词对信息对所述谓词的语义进行分类。

[0098] 对谓词标注语义分类以后,可以将原始语料按照谓词的分类结果进行聚类整合,以统一在所述知识图谱中谓词的表达。

[0099] S119,通过 NER 获取所述原始语料中的其他句子成分。

[0100] 本实施例通过在获取所述原始语料中的谓词之后,对谓词标注语义分类,并将原始语料按照谓词的分类结果进行聚类整合,使得最终生成的知识图谱中的数据一致性更好。

[0101] 图 8 示出了本发明的第五实施例。

[0102] 图 8 是本发明第一实施例提供的搜索结果信息组织方法的流程图。所述搜索结果信息组织方法以本发明第一实施例为基础,进一步的,在形成所述挖掘目标的知识图谱之后,还包括:在用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。

[0103] 参见图 8,所述搜索结果信息组织方法包括:

[0104] S810,使用语义角色标注器对原始语料进行语义角色标注。

[0105] S820,根据语义角色标注结果形成知识条目。

[0106] S830,将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。

[0107] S840,在用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。

[0108] 在形成了关于所述挖掘目标的知识图谱以后,当用户输入的搜索式与挖掘目标匹

配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。优选的,可以通过在搜索结果页面上添加所述知识图谱,来向用户显示所述挖掘目标的知识图谱。

[0109] 本实施例通过在形成知识图谱以后,当用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示,从而向用户提供关于挖掘目标的更为有序的搜索结果信息。

[0110] 图 9 示出了本发明的第五实施例。

[0111] 图 9 是本发明第五实施例提供的搜索结果信息组织装置的结构图。参见图 9,所述搜索结果信息组织装置包括:语义角色标注模块 910、知识条目形成模块 920 以及知识图谱形成模块 930。

[0112] 所述语义角色标注模块 910 用于使用语义角色标注器对原始语料进行语义角色标注。

[0113] 所述知识条目形成模块 920 用于根据语义角色标注结果形成知识条目。

[0114] 所述知识图谱形成模块 930 用于将与挖掘目标关联的知识条目,按照时间关系和/或空间关系整合形成所述挖掘目标的知识图谱。

[0115] 优选的,所述语义角色标注模块 910 包括:语义分析单元 913 以及命名实体识别单元 915。

[0116] 所述语义分析单元 913 用于利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词。

[0117] 所述命名实体识别单元 915 用于通过命名实体识别 NER 获取所述原始语料中的其他句子成分。

[0118] 优选的,所述语义角色标注模块 910 还包括:关键字泛化单元 911 以及语料扩展搜索单元 912。

[0119] 所述关键字泛化单元 911 用于在利用预先训练的语料库对所述原始语料进行语义分析,以获取所述原始语料中的谓词之前,使用预先生成的深层神经网络 DNN 泛化词典将所述挖掘目标的关键字进行泛化,以获取至少一个泛化词。

[0120] 所述语料扩展搜索单元 912 用于将所述泛化词添加为所述挖掘目标的关键字,采用所述关键字在互联网中搜索获取所述原始语料。

[0121] 优选的,所述知识条目形成模块 920 包括:归一化单元 921。

[0122] 所述归一化单元 921 用于利用所述 DNN 泛化词典对所述语义角色标注结果中的词进行归一化,以形成所述知识条目。

[0123] 优选的,所述语义角色标注模块 910 还包括:谓词语义分类单元 914。

[0124] 所述谓词语义分类单元 914 用于在获取所述原始语料中的谓词之后,对谓词标注语义分类,并将原始语料按照谓词的分类结果进行聚类整合。

[0125] 所述知识图谱形成模块 930 包括:履历图谱获取单元 931、关联人物/组织图谱获取单元 932 和/或新闻图谱获取单元 933。

[0126] 所述履历图谱获取单元 931 用于在所述知识条目中进行查询获取包括所述挖掘目标的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的履历图谱。

[0127] 所述关联人物/组织图谱获取单元 932 用于在所述知识条目中进行查询获取包括所述挖掘目标的关联人物或关联组织的知识条目,并按照时间顺序进行排序,以获取所述

挖掘目标的关联人物图谱或关联组织图谱。

[0128] 所述新闻图谱获取单元 933 用于在所述知识条目中进行查询获取包括所述挖掘目标,且发生时间在设定时间范围内的知识条目,并按照时间顺序进行排序,以获取所述挖掘目标的新闻图谱。

[0129] 优选的,所述搜索结果信息组织装置还包括:图谱显示模块 940。

[0130] 所述图谱显示模块 940 用于在形成所述挖掘目标的知识图谱之后,在用户输入的搜索式与挖掘目标匹配时,将所述挖掘目标的知识图谱作为搜索结果向用户显示。

[0131] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0132] 本领域普通技术人员应该明白,上述的本发明的各模块或各步骤可以用通用的计算装置来实现,它们可以集中在单个计算装置上,或者分布在多个计算装置所组成的网络上,可选地,他们可以用计算机装置可执行的程序代码来实现,从而可以将它们存储在存储装置中由计算装置来执行,或者将它们分别制作成各个集成电路模块,或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样,本发明不限制于任何特定的硬件和软件的结合。

[0133] 本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间的相同或相似的部分互相参见即可。

[0134] 以上所述仅为本发明的优选实施例,并不用于限制本发明,对于本领域技术人员而言,本发明可以有各种改动和变化。凡在本发明的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

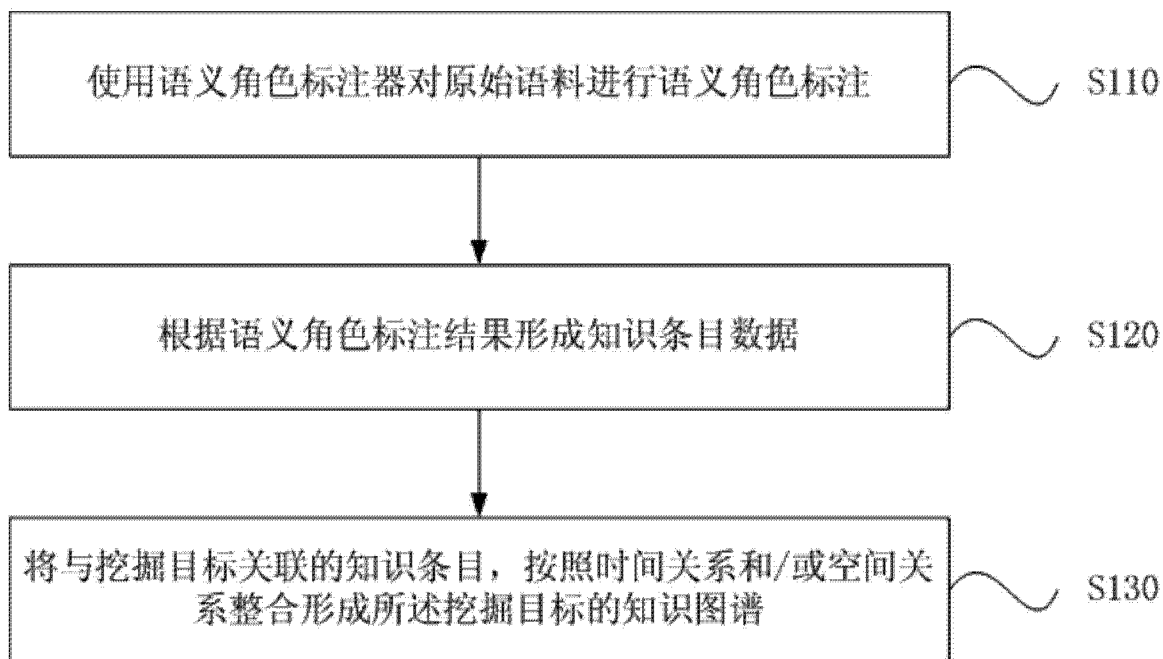


图 1

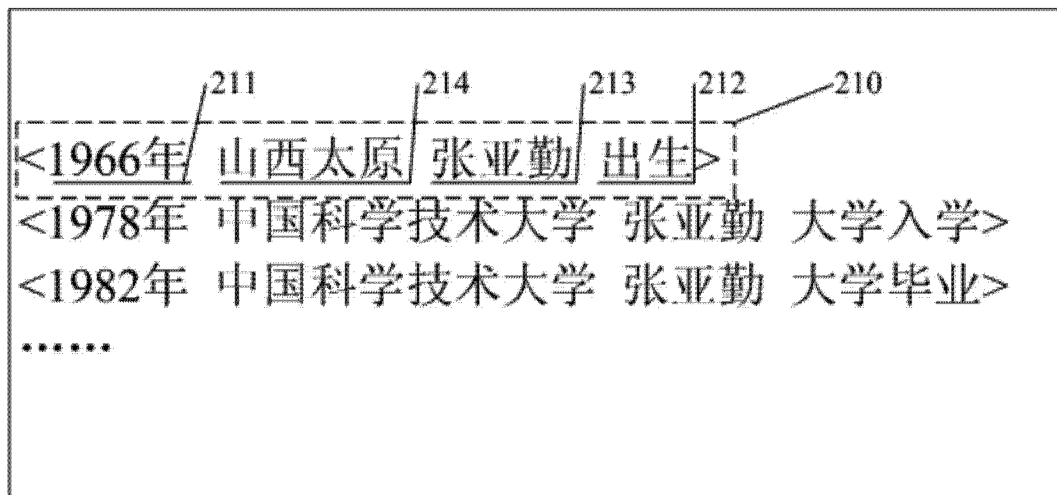


图 2

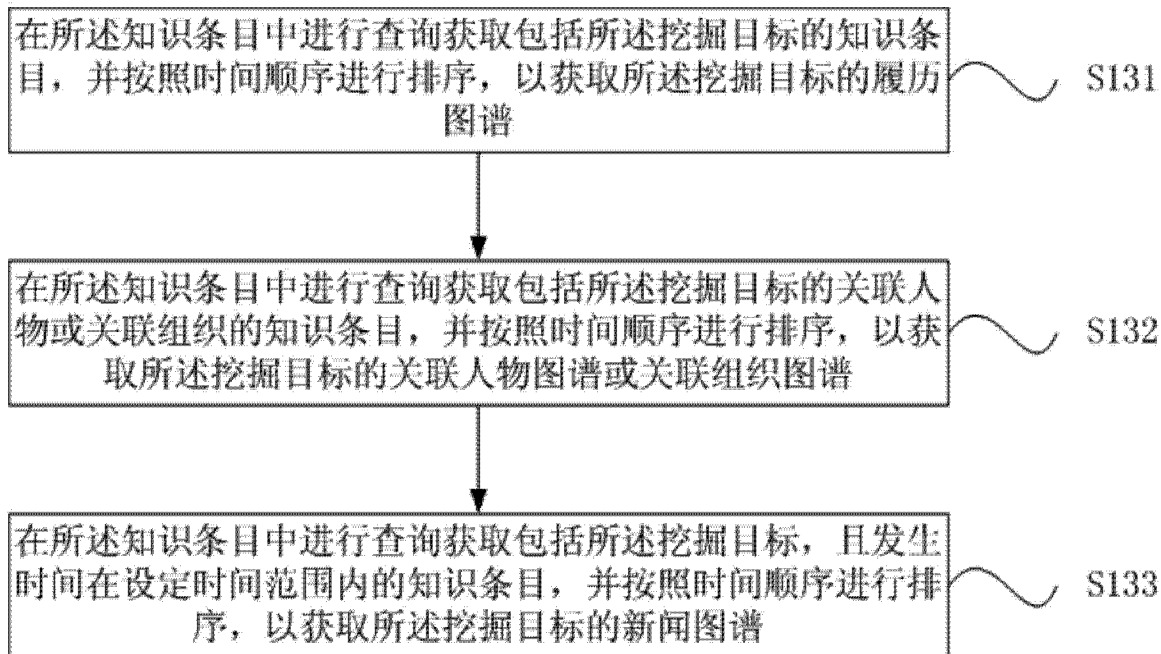


图 3

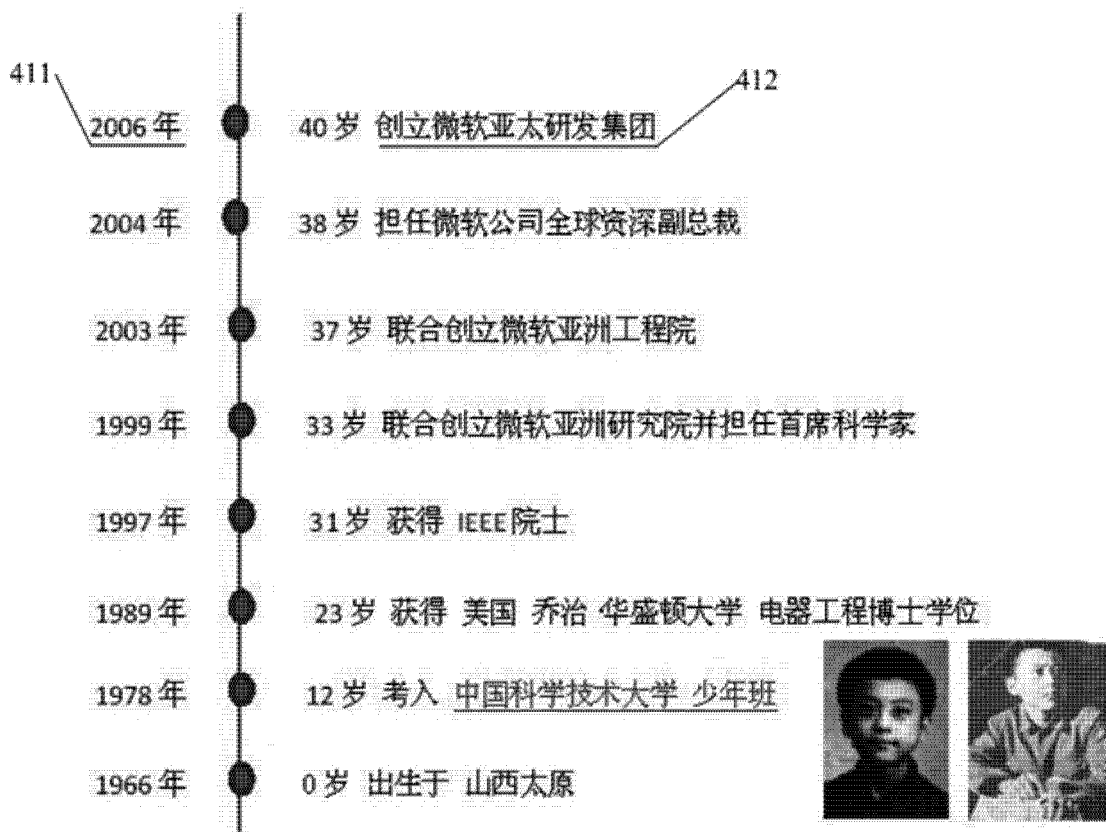


图 4a

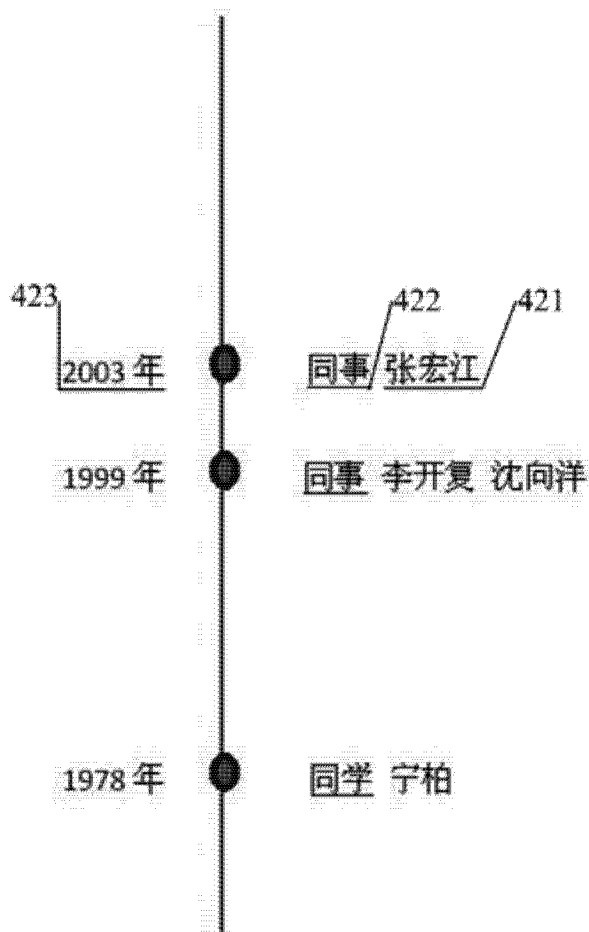


图 4b

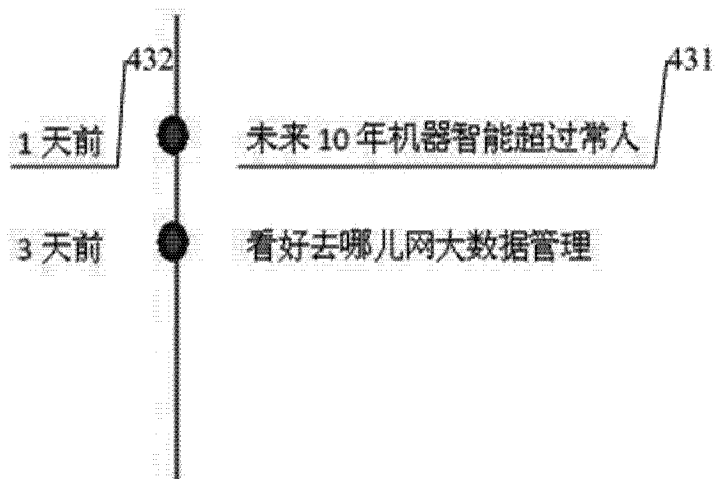


图 4c

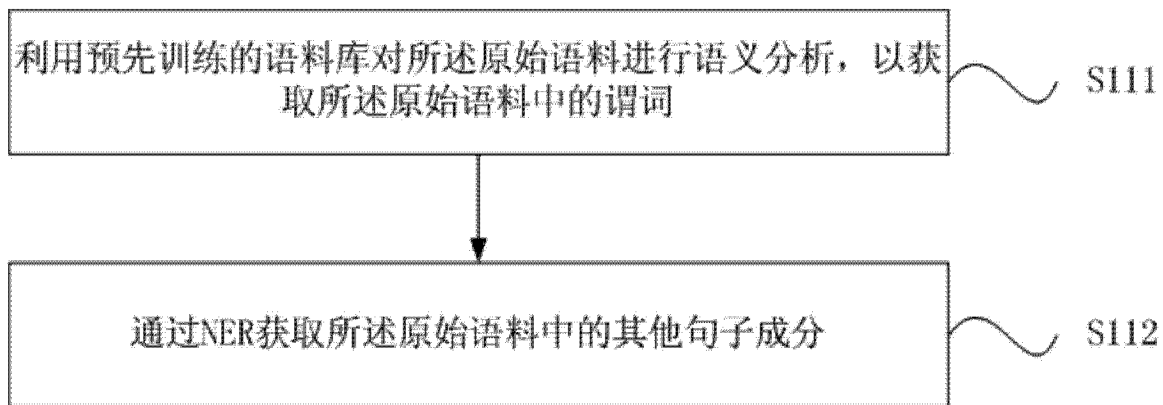


图 5

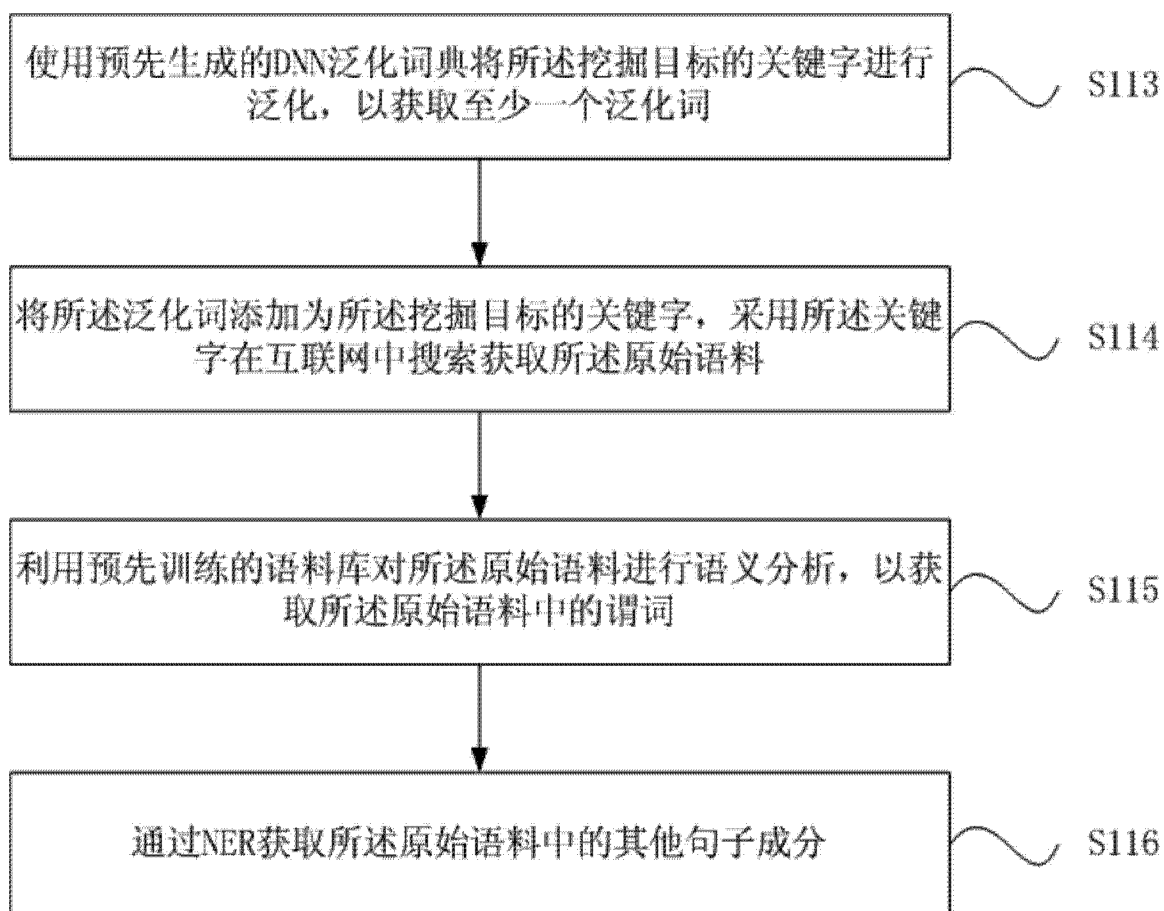


图 6

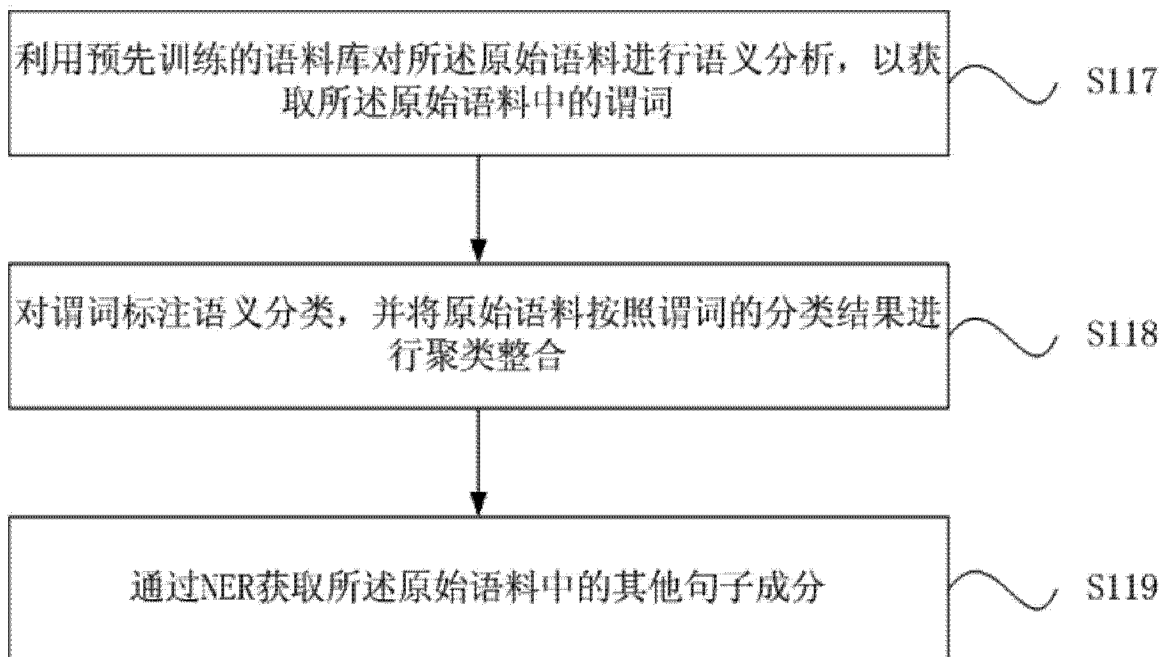


图 7

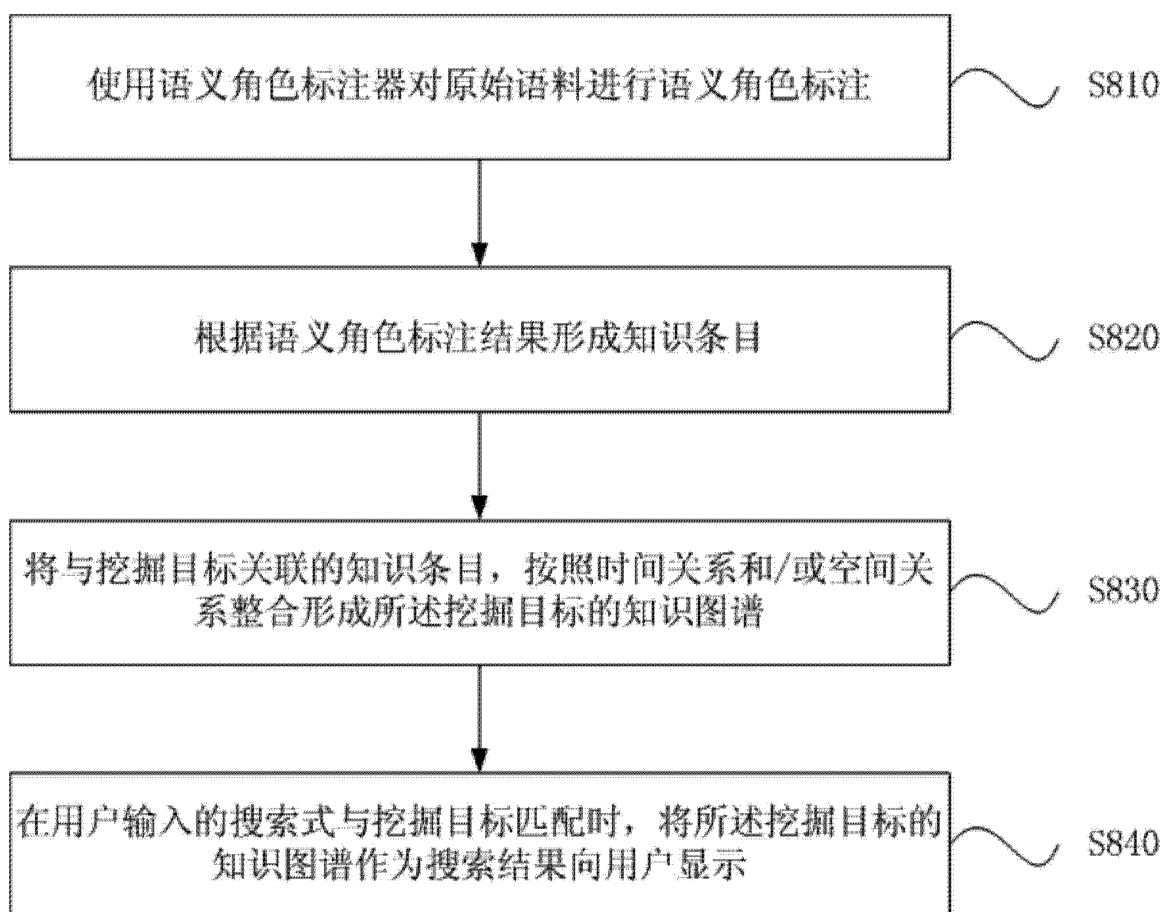


图 8

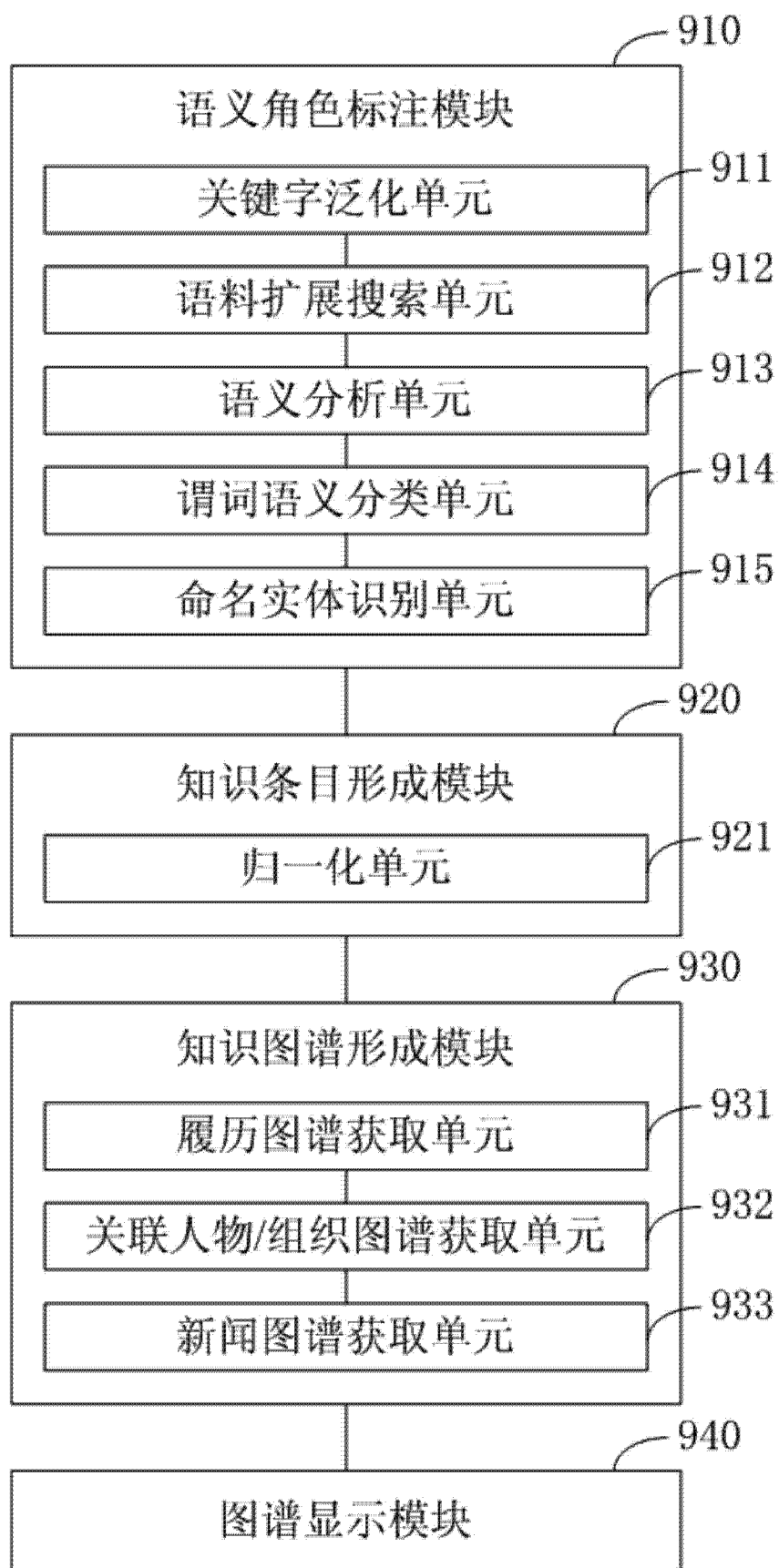


图9