(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 104636466 A
(43)申请公布日 2015.05.20

(21)申请号 201510071993.6

(22)申请日 2015.02.11

(71)申请人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号

(72)发明人 程学旗 贾岩涛 赵泽亚 王元卓 熊锦华 李曼玲 林海伦 许洪波

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇 李科

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

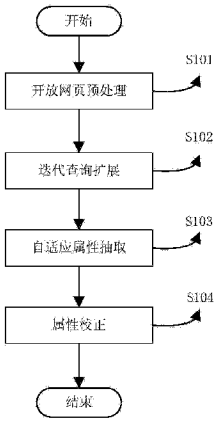
权利要求书2页 说明书16页 附图7页

(54)发明名称

一种面向开放网页的实体属性抽取方法和系统

(57)摘要

本发明提供一种面向开放网页的实体属性抽取方法和系统。其中，所述方法包括：提取开放网页的文本，从中获得目标实体的候选文本集合；以及，根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。本发明能够提高开放网页实体属性抽取的准确率和召回率，并且不依赖于网页结构，能够适应开放网页类型的变化。



CN 104636466 A

1. 一种面向开放网页的实体属性抽取方法，包括：

步骤 1)、提取开放网页的文本，从中获得目标实体的候选文本集合；

步骤 2)、根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。

2. 根据权利要求 1 所述的方法，其中，步骤 1) 包括：

步骤 11)、从开放网页中提取出非结构化文本，对该非结构化文本进行分词，得到词与所述非结构化文本之间的相关度；

步骤 12)、获得目标实体的上下文中距离该目标实体最近的一个或多个初始查询扩展词，将与目标实体及所述一个或多个初始查询扩展词相关度最高的一个或多个非结构化文本作为第一文本集合；

步骤 13)、从所述第一文本集合中选择词频最高的一个或多个二次查询扩展词，将与目标实体及所述一个或多个二次查询扩展词相关度最高的一个或多个非结构化文本作为第二文本集合；

步骤 14)、将所述第一文本集合和所述第二文本集合的并集作为目标实体的候选文本集合。

3. 根据权利要求 2 所述的方法，其中，多个词与非结构化文本的相关度为所述多个词中的每个词与该非结构化文本的相关度之和。

4. 根据权利要求 1-3 中任何一个所述的方法，其中，步骤 2) 包括：

计算目标实体属性在训练文本集合中出现的频率，如果该频率超过预定的阈值，则根据构造的统计模型来抽取目标实体属性的值，否则根据构造的层叠有穷状态自动机来抽取目标实体属性的值；其中，所述训练文本集合用于训练所述统计模型。

5. 根据权利要求 4 所述的方法，其中，根据以下步骤构造层叠有穷状态自动机：

步骤 a)、在所述候选文本集合中进行实体识别并生成概念文件；其中，所述概念文件包括指示实体类型和属于该类型的、从所述候选文本集合中识别出的实体的基本概念；指示待抽取变量的正则表达式；以及，指示实体与属性之间的关系的标志词；

步骤 b)、生成包括所述概念文件和关联规则的规则文件；其中，关联规则包括单个规则或者嵌套了多个子规则的规则，用于指示所述概念文件中基本概念、正则表达式以及标志词之间的关系；

步骤 c)、根据所述规则文件中的关联规则，构造层叠有穷状态自动机；其中，所述层叠有穷状态自动机的初始状态为基本概念、正则表达式或者标志词；其他状态包括关联规则以及关联规则中的子规则。

6. 根据权利要求 5 所述的方法，其中，根据构造的层叠有穷状态自动机来抽取目标实体属性的值包括：

将所述候选文本集合与所述层叠有穷状态自动机从初始状态开始匹配，对每个状态在所述候选文本集合中匹配到的内容建立倒排索引；

匹配完成后，从建立的倒排索引中得到目标实体属性的值。

7. 根据权利要求 4 所述的方法，其中，根据以下步骤构造统计模型：

步骤 A)、从在线百科中获得训练实体和相应的训练属性；

步骤 B)、从训练开放网页中获得所述训练实体的训练文本集合；

步骤 C)、在所述训练文本集合中提取特征，将所述训练属性的特征进行回标得到各属性的训练数据；

步骤 D)、根据所述训练数据，生成与每个属性相对应的统计模型。

8. 根据权利要求 7 所述的方法，其中，步骤 B) 包括：

步骤 B1)、从训练开放网页中提取出非结构化文本，对该非结构化文本进行分词，得到词与非结构化文本之间的相关度；

步骤 B2)、根据训练实体在训练开放网页中的上下文信息获得距离该训练实体最近的 n 个初始查询扩展词，将与训练实体及初始查询扩展词相关度最高的 K 个非结构化文本作为第三文本集合；其中，n 和 K 为正整数；

步骤 B3)、从所述第三文本集合中选择词频最高的 m 个二次查询扩展词，将与训练实体及二次查询扩展词相关度最高的 L 个非结构化文本作为第四文本集合；其中，m 和 L 为正整数；

步骤 B4)、取所述第三文本集合和所述第四文本集合的并集，作为训练文本集合。

9. 根据权利要求 7 所述的方法，其中，步骤 C) 还包括：

去除所述训练数据中的杂质，以及控制所述训练数据中的正例与反例的比例。

10. 根据权利要求 7 所述的方法，其中，所述特征包括词语、词语之间的依存关系、词语的词频及词性。

11. 根据权利要求 7-10 中任何一个所述的方法，其中，根据构造的统计模型来抽取目标实体属性的值包括：

按照在构造所述统计模型时提取特征的方式来提取所述候选文本集合的特征；

将提取出的特征输入目标实体属性对应的统计模型，得到目标实体属性的值。

12. 根据权利要求 1-3 中任何一个所述的方法，还包括：

步骤 3)、根据目标实体属性的类型、词性或者取值范围，校正抽取出的目标实体属性的值。

13. 一种面向开放网页的实体属性抽取系统，包括：

网页预处理模块，用于提取开放网页的文本；

查询扩展模块，用于从提取出的文本中获得目标实体的候选文本集合；

属性抽取模块，用于根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。

一种面向开放网页的实体属性抽取方法和系统

技术领域

[0001] 本发明涉及数据挖掘技术领域，特别地，涉及一种面向开放网页的实体属性抽取方法和系统。

背景技术

[0002] 开放网页是指数据源不固定、包含多种网络数据的非结构化互联网网页，如博客、论坛、新闻、聊天记录、电子邮件等，其信息的性质和量值出现的位置不固定，所有内容都是不可预知的。随着网络技术的发展，特别是 Internet 和 Intranet 技术的飞快发展，开放网页以其结构灵活的自身特点，在数量快速增大的同时，也为其文本理解带来困难：

[0003] 1、文本结构不固定，没有特定的上下文语法；

[0004] 2、关键词范围不固定，涉及的学科领域多样；

[0005] 3、文本长度不固定，上下文信息量差距较大；

[0006] 4、数据源不固定，语言现象复杂。

[0007] 实体是指客观存在并可相互区别的事物，可以是具体的客观对象，也可以是抽象的事件。实体属性是指实体本身的性质，实体属性抽取通过将不同信息源对于某一实体的属性集中起来，从不同的角度反映这个实体的相关情况，完善对该实体的认识，在信息抽取、事件跟踪、人名消歧等研究中有着重要作用，并且已成为文本理解的关键技术。

[0008] 针对开放网页的特点，传统的实体属性抽取方法在以下方面存在限制：

[0009] 第一、开放网页的文本结构不固定，实体及其描述没有固定规律可循，且多数在自由文本中，不易抽取分析；

[0010] 第二、传统的面向规则的属性抽取方法，规则定义死板，过于依赖上下文语法，且匹配效率低下；

[0011] 第三、开放网页的数据源不固定，语言现象复杂，普通规则难以涵盖，传统的基于规则的属性抽取不支持规则的嵌套匹配；

[0012] 第四、传统的基于统计的实体属性抽取方法，训练数据的准备过于依赖人工，效率不高，且准确率和召回率较低；

[0013] 第五、传统的属性抽取多局限在某个领域或学科里面进行，不能将系统直接移植到其他领域或学科进行使用，缺乏具有通用性的关联特征，不易移植和扩展。

发明内容

[0014] 为解决上述问题，根据本发明的一个实施例，提供一种面向开放网页的实体属性抽取方法，包括：

[0015] 步骤 1)、提取开放网页的文本，从中获得目标实体的候选文本集合；

[0016] 步骤 2)、根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。

[0017] 上述方法中，步骤 1) 包括：

[0018] 步骤 11)、从开放网页中提取出非结构化文本，对该非结构化文本进行分词，得到词与所述非结构化文本之间的相关度；

[0019] 步骤 12)、获得目标实体的上下文中距离该目标实体最近的一个或多个初始查询扩展词，将与目标实体及所述一个或多个初始查询扩展词相关度最高的一个或多个非结构化文本作为第一文本集合；

[0020] 步骤 13)、从所述第一文本集合中选择词频最高的一个或多个二次查询扩展词，将与目标实体及所述一个或多个二次查询扩展词相关度最高的一个或多个非结构化文本作为第二文本集合；

[0021] 步骤 14)、将所述第一文本集合和所述第二文本集合的并集作为目标实体的候选文本集合。

[0022] 上述方法中，多个词与非结构化文本的相关度为所述多个词中的每个词与该非结构化文本的相关度之和。

[0023] 上述方法中，步骤 2) 包括：计算目标实体属性在训练文本集合中出现的频率，如果该频率超过预定的阈值，则根据构造的统计模型来抽取目标实体属性的值，否则根据构造的层叠有穷状态自动机来抽取目标实体属性的值；其中，所述训练文本集合用于训练所述统计模型。

[0024] 上述方法中，根据以下步骤构造层叠有穷状态自动机：

[0025] 步骤 a)、在所述候选文本集合中进行实体识别并生成概念文件；其中，所述概念文件包括指示实体类型和属于该类型的、从所述候选文本集合中识别出的实体的基本概念；指示待抽取变量的正则表达式；以及，指示实体与属性之间的关系的标志词；

[0026] 步骤 b)、生成包括所述概念文件和关联规则的规则文件；其中，关联规则包括单个规则或者嵌套了多个子规则的规则，用于指示所述概念文件中基本概念、正则表达式以及标志词之间的关系；

[0027] 步骤 c)、根据所述规则文件中的关联规则，构造层叠有穷状态自动机；其中，所述层叠有穷状态自动机的初始状态为基本概念、正则表达式或者标志词；其他状态包括关联规则以及关联规则中的子规则。

[0028] 上述方法中，根据构造的层叠有穷状态自动机来抽取目标实体属性的值包括：

[0029] 将所述候选文本集合与所述层叠有穷状态自动机从初始状态开始匹配，对每个状态在所述候选文本集合中匹配到的内容建立倒排索引；

[0030] 匹配完成后，从建立的倒排索引中得到目标实体属性的值。

[0031] 上述方法中，根据以下步骤构造统计模型：

[0032] 步骤 A)、从在线百科中获得训练实体和相应的训练属性；

[0033] 步骤 B)、从训练开放网页中获得所述训练实体的训练文本集合；

[0034] 步骤 C)、在所述训练文本集合中提取特征，将所述训练属性的特征进行回标得到各属性的训练数据；

[0035] 步骤 D)、根据所述训练数据，生成与每个属性相对应的统计模型。

[0036] 上述方法中，步骤 B) 包括：

[0037] 步骤 B1)、从训练开放网页中提取出非结构化文本，对该非结构化文本进行分词，得到词与非结构化文本之间的相关度；

[0038] 步骤 B2)、根据训练实体在训练开放网页中的上下文信息获得距离该训练实体最近的 n 个初始查询扩展词，将与训练实体及初始查询扩展词相关度最高的 K 个非结构化文本作为第三文本集合；其中，n 和 K 为正整数；

[0039] 步骤 B3)、从所述第三文本集合中选择词频最高的 m 个二次查询扩展词，将与训练实体及二次查询扩展词相关度最高的 L 个非结构化文本作为第四文本集合；其中，m 和 L 为正整数；

[0040] 步骤 B4)、取所述第三文本集合和所述第四文本集合的并集，作为训练文本集合。

[0041] 上述方法中，步骤 C) 还包括：去除所述训练数据中的杂质，以及控制所述训练数据中的正例与反例的比例。

[0042] 上述方法中，所述特征包括词语、词语之间的依存关系、词语的词频及词性。

[0043] 上述方法中，根据构造的统计模型来抽取目标实体属性的值包括：

[0044] 按照在构造所述统计模型时提取特征的方式来提取所述候选文本集合的特征；

[0045] 将提取出的特征输入目标实体属性对应的统计模型，得到目标实体属性的值。

[0046] 上述方法中，还包括：

[0047] 步骤 3)、根据目标实体属性的类型、词性或者取值范围，校正抽取出的目标实体属性的值。

[0048] 根据本发明的一个实施例，还提供一种面向开放网页的实体属性抽取系统，包括：

[0049] 网页预处理模块，用于提取开放网页的文本；

[0050] 查询扩展模块，用于从提取出的文本中获得目标实体的候选文本集合；

[0051] 属性抽取模块，用于根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。

[0052] 本发明具有如下的有益效果：

[0053] 1、提出一种基于层叠有穷状态自动机的实体属性抽取方法，实现了复杂嵌套规则的抽取；

[0054] 2、在基于层叠有穷状态自动机的抽取过程中，对该自动机每个状态抽取的内容建立倒排索引，大大提升了规则匹配效率；

[0055] 3、制定了一套无关文法的概念定义和规则定义语言，使得实体属性抽取脱离上下文语言环境，实现声明式信息抽取，提升了系统的兼容性；

[0056] 4、对 CRF 模型训练提出了一套句级的文本特征，能够提升属性抽取中机器学习的效果；

[0057] 5、提出了根据在线百科属性框 (Infobox) 的已有属性信息回标，自动生成 CRF 训练数据的方法，并针对回标效果提出需人工校验的部分，提升了训练数据的效率和准确性；

[0058] 6、提供一种迭代查询扩展的方法，经验证能够提高开放网页的实体属性抽取的准确率和召回率；

[0059] 7、根据属性的出现频率自适应地采用基于规则或基于统计的抽取方法，实现对开放网页的实体属性抽取。

附图说明

[0060] 以下参照附图对本发明实施例作进一步说明，其中：

[0061] 图 1 是根据本发明一个实施例的面向开放网页的实体属性抽取方法的流程图；

[0062] 图 2 是根据本发明一个实施例的迭代查询扩展方法的流程图；

[0063] 图 3 是根据本发明一个实施例的自适应实体属性抽取方法的流程图；

[0064] 图 4 是根据本发明一个实施例的构建层叠有穷状态自动机以及基于层叠有穷状态自动机的关联规则进行属性抽取的方法的流程图；

[0065] 图 5 是根据本发明一个实施例的层叠有穷状态自动机的示意图；

[0066] 图 6 是根据本发明一个实施例的初始倒排索引的示意图；

[0067] 图 7 是根据本发明一个实施例的层叠有穷状态自动机与候选文本集合匹配的示意图；

[0068] 图 8 是根据本发明一个实施例的匹配完成时的倒排索引的示意图；

[0069] 图 9 是根据本发明一个实施例的构建条件随机场模型以及基于条件随机场模型的属性抽取方法的流程图。

具体实施方式

[0070] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图通过具体实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0071] 根据本发明的一个实施例，提供一种面向开放网页的实体属性抽取方法。

[0072] 概括而言，该方法包括：提取开放网页的文本，从中获得目标实体的候选文本集合；根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从所述候选文本集合中抽取目标实体属性的值。

[0073] 在描述该面向开放网页的实体属性抽取方法之前，首先对实体属性、规则和统计模型进行说明。其中，实体属性包括目标实体、属性名和属性值三个部分；规则包括规则类型、目标名称及参数、以及规则体，统计模型使用的特征的文本来源包括属性名前的文本、属性名和属性值之间的文本以及属性值后的文本。

[0074] 现结合图 1 对该面向开放网页的实体属性抽取方法的各步骤进行详细描述。需要说明的是，说明书中描述的方法的各个步骤并非一定是必须的，而是可以根据实际情形来省略或替换其中的一个或多个步骤，另外，各个步骤之间的顺序也是可以调整的。

[0075] 步骤 S101：开放网页预处理

[0076] 根据本发明的一个实施例，开放网页的预处理过程包括：

[0077] 1、获得待抽取的开放网页集合，抽取网页内容，得到待抽取的非结构化文本。

[0078] 2、将待抽取的非结构化文本进行分词，计算词与每个非结构化文本的相关度，得到与每个词对应的最高相关度（或称匹配度）非结构化文本集合，并根据以上信息建立倒排索引。

[0079] 在一个实施例中，根据词频等特征来计算词与非结构化文本之间的相关度。例如，可利用 TF-IDF 方法得到一个词与所有非结构化文本的相关度，接着将相关度最高的 k(k 为正整数）个非结构化文本作为该词的最高相关度非结构化文本集合。

[0080] 步骤 S102 :通过迭代查询扩展得到候选文本集合

[0081] 根据步骤 S101 中建立的倒排索引,通过两次查询扩展并融合目标实体的上下文信息和词频信息,生成候选文本集合。图 2 描述了迭代查询扩展方法的步骤的一个实施例,包括 :

[0082] 步骤 S201、根据目标实体 E 的上下文信息,获取上下文中距离 E 最近的 n(n 为正整数) 个实体 (词),称作查询扩展词。

[0083] 在一个实施例中,选取 n = 1,即将目标实体 E 的前后词语 E1 和 E2 作为查询扩展词。

[0084] 步骤 S202、初始查询扩展。

[0085] 在步骤 S101 中建立的倒排索引中查询目标实体和查询扩展词,得到与目标实体和查询扩展词相关度最高的非结构化文本的文本集合 U_1。

[0086] 在一个实施例中,将目标实体、查询扩展词与某个非结构化文本的相关度之和作为该目标实体和查询扩展词与该非结构化文本的相关度并排序,从而得到文本集合 U_1(如包括 50 篇文本)。在另一个实施例中,分别找到目标实体和查询扩展词的最高相关度非结构化文本集合,取交集得到文本集合 U_1。通过实验发现,初始查询扩展过程可提升实体属性抽取的准确率。

[0087] 步骤 S203、从 U_1 中选出词频最高的 m(m 为正整数) 个词。

[0088] 在一个实施例中,选取 m = 2,即选取 U_1 中词频最高的两个实体 E3 和 E4 进行第二次查询扩展。

[0089] 步骤 S204、二次查询扩展。

[0090] 将词频最高的 m 个词与目标实体 E 一起再次在倒排索引中查询,得到与它们相关度最高的文本集合 U_2。例如,采用步骤 S202 中的方法来得到 U_2。通过实验发现,该步骤能有效提升实体属性抽取的召回率和准确率。

[0091] 步骤 S205、将两次查询扩展的结果取并集,作为目标实体 E 的实体属性抽取的候选文本集合 U(简称候选文本集合 U)。

[0092] 步骤 S103 :自适应实体属性抽取

[0093] 概括而言,自适应实体属性抽取过程包括 :根据目标实体属性 (或称目标属性) 在训练文本集合中出现的频率,自适应地选择不同的实体属性抽取方法。其中,训练文本集合是用于训练统计模型 (该模型用于基于统计的实体属性抽取方法,并将在下文中描述) 的文本集合,可根据下文的标志词在训练文本集合中出现的频率来得到目标属性在训练文本集合中出现的频率。这里,如果出现频率高于预定阈值,则采用基于统计的实体属性抽取方法,否则采用基于规则的实体属性抽取方法。这样做的原因在于 :对于出现频率较低的实体属性,采用基于规则的方法的精度和执行效率更好 ;而对于出现频率较高的属性,选择基于统计的方法更全面。

[0094] 其中,基于规则的实体属性抽取方法可通过构造层叠有穷状态自动机,实现规则嵌套,并对层叠有穷状态自动机每个状态 (或称节点) 匹配到的文本内容建立倒排索引,快速实现复杂文本模式的匹配,得到实体属性值 ;基于统计的实体属性抽取方法可根据条件随机场原理进行有监督的机器学习,选取文本特征,训练统计模型 (如条件随机场模型) 来抽取实体属性。如图 3 所示,自适应实体属性抽取过程可包括以下子步骤 :

[0095] 步骤 S301、构建层叠有穷状态自动机。

[0096] 对候选文本集合 U 进行实体识别，制定一套无关文法的声明式语言规范，其中定义概念集合和关联规则集合，根据规则的嵌套依赖关系，构造出层叠有穷状态自动机。

[0097] 步骤 S302、训练统计模型。

[0098] 选择文本特征，生成训练数据，通过训练得到统计模型，如 CRF 模型 M_{CRF}。

[0099] 步骤 S303、计算目标实体属性在训练文本集合中出现的频率，判断是否超过预定阈值。

[0100] 步骤 S304、若步骤 S303 的判断结果为否，则采用基于层叠有穷状态自动机的关联规则进行属性抽取（即基于规则的实体属性抽取方法）。

[0101] 步骤 S305、若步骤 S303 的判断结果为是，则采用基于条件随机场的机器学习进行属性抽取（即基于统计的实体属性抽取方法）。对候选文本集 U 进行句子级的特征抽取，生成特征向量，输入步骤 S302 中生成的统计模型，抽取出目标属性值。

[0102] 应理解，上述子步骤的顺序是可以调换的，例如，训练统计模型的子步骤可以在构建层叠有穷状态自动机之前的任何时候进行或者与构建层叠有穷状态自动机同时进行。

[0103] 上文对自适应实体属性抽取过程进行了概括描述，下面将分别对构建层叠有穷状态自动机，基于层叠有穷状态自动机进行属性抽取；训练统计模型，基于统计模型（尤其是条件随机场模型 M_{CRF}）进行属性抽取的过程进行详细描述。

[0104] 图 4 描述了构建层叠有穷状态自动机以及基于层叠有穷状态自动机的关联规则进行属性抽取的方法，以下是该方法的各个子步骤：

[0105] 步骤 S401、实体识别。

[0106] 在候选文本集合 U 中进行命名实体识别，得到实体集合，并确定实体的类型，如人、地点、机构等。

[0107] 步骤 S402、生成概念文件。

[0108] 概念文件是所有概念的集合，包括 CONCEPT 基本概念、REGEX 正则表达式和 CONCEPT 标志词。使用与上下文文法无关的语言，进行声明式定义：

[0109] (1)CONCEPT 基本概念即待抽取的类型变量，如人物、地点、组织等（一种类型定义为一个概念对象）。定义格式为 CONCEPT:[概念名称]:[实例取值]，其中，概念名称即类型变量，实例取值即该类型的实体集合，可认为是<类型变量－类型变量取值范围>的概念对。其中，基本概念由步骤 S401 生成的实体集合得到。例如：将文中所有“组织”类型的实体归为一个概念，概念名为 ORG，其取值范围是文中类型为组织的实例，例如含中国科学技术大学研究生院、北京大学、中科院计算所等，该 ORG 概念定义如表 1 所示。

[0110] 表 1

[0111]

CONCEPT:ORG:中国科学技术大学研究生院
CONCEPT:ORG:中科院计算所
CONCEPT:ORG:北京大学

[0112] (2)REGEX 正则表达式即为待抽取的变量的正则表达式，格式为 REGEX:[概念名称]:[正则表达式内容]，表 2 给出了 REGEX 正则表达式的一个例子。

[0113] 表 2

[0114]

REGEX:DATE:([\d]{4} 年){0,1}([\d]{1,2} 月){0,1}([\d]{1,2} 日){0,1}

[0115] 其中，DATE 为该正则表达式的名称，([\d]{4} 年){0,1}([\d]{1,2} 月){0,1}([\d]{1,2} 日){0,1} 为 DATE 所指代的正则表达式。

[0116] (3)CONCEPT 标志词是与待抽取属性相关的标识词，即实体与属性的关系标志词，用于关联规则的制定，格式为 CONCEPT:[概念名称]:[标志词取值]。例如，要抽取人的属性“出生日期”时，所需用到的标志词可如表 3 所示。

[0117] 表 3

[0118]

CONCEPT:BIRTH_OR:生于
CONCEPT:BIRTH_OR:出生
CONCEPT:BIRTH_OR:诞生

[0119] 将以上 CONCEPT 和 REGEX 的集合合并，成为总的概念集合，即生成概念文件，如表 4 所示。

[0120] 表 4

[0121]

…
CONCEPT:NAME:李国杰
CONCEPT:NAME:程学旗
CONCEPT:ORG:中国科学技术大学研究生院

[0122]

CONCEPT:ORG:中科院计算所
CONCEPT:ORG:北京大学
CONCEPT:ORG:清华大学
REGEX:DATE:([\d]{4}年){0,1}([\d]{1,2}月){0,1}([\d]{1,2}日){0,1}
REGEX:DATE:[\d]{4}[\.-][\d]{1,2}([\.-][\d]{1,2}){0,1}
CONCEPT:BIRTH_OR:生于
CONCEPT:BIRTH_OR:出生
CONCEPT:BIRTH_OR:诞生
…

[0123] 步骤 S403、生成规则文件。

[0124] 规则文件由关联规则 MCONCEPT_RULE 集合和步骤 S402 生成的概念文件合并生成。关联规则 MCONCEPT_RULE 表征概念之间的关系，是通过对概念进行布尔逻辑约束和上下文约束得到的，格式为 MCONCEPT_RULE:[规则名称]([待输出变量]) :([约束],["_ 待输出变量 { 概念 }"],[标志词])。待输出变量即匹配到的概念的实例（包括实体实例，以及属性实例，即属性值），其中约束包括但不限于：

[0125] (1)AND :所有子句都出现的字符串才会被匹配；

[0126] (2)OR :只要有一个子句出现，该字符串就会被匹配；

[0127] (3)SENT :所有子句都出现在同一个句子，该字符串才会被匹配；

[0128] (4)ORD :所有子句按规则定义的顺序同时出现，该字符串才会被匹配；

[0129] (5)DIST_n :所有子句同时出现在字符串，且相邻子句实例距离（间隔距离）不超过 n 个词时，该字符串会被匹配。

[0130] MCONCEPT_RULE 通过定义概念之间的关系，将满足该关系的概念的实例抽取出来。步骤 S402 中生成的 CONCEPT 和 REGEX 都是待抽取的概念，其匹配到的文本即该概念的实例，如表 5 示出了 NAME_BIRTHDAY 规则，其抽取人的“出生日期。”

[0131] 表 5

[0132]

```
MCONCEPT_RULE:NAME_BIRTHDAY(person,birthday):(DIST_20,
```

[0133]

```
"_person{NAME}","BIRTH_OR","_birthday{DATE}")
```

[0134] 其中，NAME_BIRTHDAY 为规则名称，(person,birthday) 表示输出 person 和 birthday 两个变量，其中 person 是 NAME 概念的实例，birthday 是 DATE 变量的实例。NAME_BIRTHDAY 的含义是：若 NAME、BIRTH_OR、DATE 概念的实例同时出现且距离不超过 20 个词，则将 NAME 匹配的子句（即 NAME 的实例）作为 person 输出，DATE 匹配的子句作为 birthday 输出。

[0135] 特别地，本发明支持规则的嵌套。根据规则之间的依赖关系，形成复杂嵌套规则。例如，表 6 示出了 NAME_COLLEGE 规则，抽取人的“毕业院校及毕业时间”。

[0136] 表 6

[0137]

```
MCONCEPT_RULE:NAME_COLLEGE(person, graduatetime,
college):(SENT, "_person{NAME}", "_graduatetime{DATE}",
"_college{ORG}", (ORD, "DEGREE_GET_OR","DEGREE"))
```

[0138] 表 6 表示如下：首先，必须匹配成功 (ORD,"DEGREE_GET_OR","DE GREE") 子规则，即存在 "DEGREE_GET_OR" 后出现 "DEGREE" 的子句；其次，子规则匹配成功后，若该子句所在的句子中，同时出现了 NAME、DATE、ORG 概念的实例，则将 NAME 匹配的子句（即 NAME 的实例）作为 person 输出，DATE 匹配的子句作为 graduatetime 输出，ORGANIZATI ON 匹配的子句作为 college 输出。

[0139] 规则文件由所有 MCONCEPT_RULE 和所有概念集合生成，以“#”开头代表该行被注释，表 7 示出了抽取人的属性“出生日期”、“毕业院校”、“联系方式”的规则文件。

[0140] 表 7

[0141]

```
#birthday
REGEX:DATE:([\d]{4}年){0,1}([\d]{1,2}月){0,1}([\d]{1,2}日){0,1}
REGEX:DATE:[\d]{4}[\.-][\d]{1,2}([\.-][\d]{1,2}){0,1}
CONCEPT:BIRTH_OR:生于
CONCEPT:BIRTH_OR:出生
CONCEPT:BIRTH_OR:诞生
```

[0142]

```
MCONCEPT_RULE:NAME_BIRTHDAY(person, birthday):(DIST_20,
"_person{NAME}", "BIRTH_OR", "_birthday{DATE}")

#教育经历
CONCEPT:GRADUATE:毕业
CONCEPT:DEGREE_GET_OR:获
CONCEPT:DEGREE_GET_OR:取得
CONCEPT:DEGREE:学士
CONCEPT:DEGREE:硕士
CONCEPT:DEGREE:博士
CONCEPT:DEGREE:博士后
MCONCEPT_RULE:NAME_COLLEGE(person,college):(SENT,
"_person{NAME}", (DIST_3, "GRADUATE", "_college{ORG}"))
MCONCEPT_RULE:NAME_COLLEGE(person,graduatetime,
college):(SENT,"_person{NAME}",(DIST_10, "_graduatetime{DATE}",
"GRADUATE", "_college{ORG}"))
MCONCEPT_RULE:NAME_COLLEGE(person,graduatetime,
college):(SENT,"_person{NAME}","_graduatetime{DATE}",
"_college{ORG}", (ORD, "DEGREE_GET_OR","DEGREE"))
MCONCEPT_RULE:NAME_COLLEGE(person,college):(SENT,
"_person{NAME}", "_college{ORG}", (ORD, "DEGREE_GET_OR",
"DEGREE"))

#phone number
CONCEPT:FIXEDPHONENAME:固定电话
CONCEPT:FIXEDPHONENAME:座机
CONCEPT:FIXEDPHONENAME:phone
CONCEPT:FIXEDPHONENAME:电话
CONCEPT:MOBILEPHONENAME:移动电话
CONCEPT:MOBILEPHONENAME:mobile
CONCEPT:MOBILEPHONENAME:手机
REGEX:FIXEDPHONE:([\d]{3,4}-){0,1}[\d]{7,8}
```

[0143]

```
REGEX:MOBILEPHONE:[\d]{11}
REGEX:MOBILEPHONE:\+86[\d]{11}
MCONCEPT_RULE:NAME_FIXEDPHONE(person,
fixedphone):(ORD,(DIST_50,"_person{NAME}",(DIST_5,
"FIXEDPHONENAME", "_fixedphone{FIXEDPHONE}")))
MCONCEPT_RULE:NAME_MOBILEPHONE(person,
mobilephone):(ORD,(DIST_50,"_person{NAME}",(DIST_5,
"MOBILEPHONENAME", "_mobilephone{MOBILEPHONE}")))

#email
CONCEPT:EMAILNAME:电子邮箱
CONCEPT:EMAILNAME:邮箱
CONCEPT:EMAILNAME:email
CONCEPT:EMAILNAME:Email
REGEX:EMAIL:\w+@\w+(\.\w+)+
MCONCEPT_RULE:NAME_EMAIL(person, email):(ORD, (DIST_50,
"_person{NAME}", (DIST_2, "EMAILNAME", "_email{EMAIL}")))
```

[0144] 步骤 S404、构建层叠有穷状态自动机。

[0145] 根据规则之间的嵌套依赖关系，将规则转移为一组彼此有依赖关系的有穷状态自动机，每个概念是初始状态，通过概念之间的约束关系和规则嵌套依赖关系，逐步生成层叠的有穷状态自动机。该层叠有穷状态自动机呈树状，底层为初始状态，其可转移成的状态可以看作其父状态；初始状态为概念，其父状态为规则或子规则，转移函数即为规则或子规则的约束条件。通过规则的约束条件和嵌套关系，从初始状态开始，逐步向上转移为规则或子规则状态，构成层叠有穷状态自动机，如图 5 所示。

[0146] 步骤 S405、将候选文本集合 U 与层叠有穷状态自动机的初始状态匹配，建立初始倒排索引。

[0147] 本步骤将候选文本集合 U 与初始状态匹配，即匹配各个概念，并对初始状态匹配到的文本建立倒排索引。其中，每个状态作为词项，该状态匹配到的文本作为该词项的倒排记录表，每个词项有一个指针指向其倒排记录表，如图 6 所示。

[0148] 步骤 S406、根据层叠有穷状态机进行状态转移，得到实体属性。

[0149] 以初始状态为起点，自底向上判断层叠有穷状态自动机中的每一个状态可否进行状态转移。状态转移函数为父状态代表的规则或子规则的约束条件，通过判断规则或子规则所需的其他状态是否在倒排索引中存在，可以得出是否能够进行状态转移。若可以转移

到该状态，则对该状态匹配的文本建立倒排索引，并追加在步骤 S405 中生成的倒排索引后，继续向上判断是否可以状态转移；若不能转移，则该状态是匹配成功的最复杂规则，终止向上匹配，并根据倒排索引将该规则包含的概念的实例输出，获取该规则所代表的属性值。其中，在匹配的过程中，动态维护一个各状态所匹配的文本内容的倒排索引。

[0150] 例如，输入步骤 S403 得到的规则文件得到层叠有穷状态自动机，与候选文本集合 U 进行匹配，得到属性值，如表 8 所示。

[0151] 表 8

[0152]

候选文本集合	规则文件
… … 李国杰，男，1943 年 5 月生于湖南邵阳，1968 年本科毕业于北京大学,1981 年硕士毕业于中国科学技术大学研究生院，1985 年获美国普渡大学博士学位。 … … … …	… MCONCEPT_RULE:NAME_COLLEGE1(person, graduatetime, college):(SENT, "_person{NAME}", (DIST_10, "_graduatetime{DATE}", "GRADUATE", "_college{ORG}")) MCONCEPT_RULE:NAME_COLLEGE2(person, graduatetime, college):(SENT, "_person{NAME}", "_graduatetime{DATE}", "_college{ORG}", (ORD, "DEGREE_GET_OR","DEGREE")) MCONCEPT_RULE:NAME_BIRTHDAY(person, birthday):(DIST_20, "_person{NAME}", "BIRTH_OR", "_birthday{DATE}") … CONCEPT:NAME:李国杰 CONCEPT:NAME:程学旗

[0153]

	CONCEPT:ORG:中国科学技术大学研究生院 CONCEPT:ORG:中科院计算所 CONCEPT:ORG:北京大学 CONCEPT:ORG:普渡大学
匹配结果: person:李国杰 birthday:1943 年 5 月 ----- person:李国杰 graduatetime:1968 年 college: 北京大学 ----- person:李国杰 graduatetime:1981 年 college: 中国科学技术大学研究生院 ----- person:李国杰 graduatetime:1985 年 college: 普渡大学	

[0154] 在该示例中，输入上述 3 条规则 NAME_COLLEGE1、NAME_COLLE GE2 和 NAME_BIRTHDAY，对该段文本匹配时，首先根据规则的依赖关系构建层叠有穷状态自动机，如图 5 所示。

[0155] 根据构建的层叠有穷状态自动机，将初始状态与候选文本进行匹配，对 NAME、DATE、BIRTH_OR、GRADUATE、ORG、DEGREE_GET_OR、DEGREE 匹配到的内容建立倒排索引，如图 6 所示。

[0156] 根据层叠有穷状态自动机，进行状态转移（即 S406）。从第一个匹配到的初始状态 NAME 开始，能够转移的状态有 NAME_BIRTHDAY、NAME_COLLEGE1 规则、NAME_COLLEGE2 规则，依次查看转移所需的其余状态是否满足，例如 :NAME_BIRTHDAY 还需在间距 20 以内有 DATE、BIRTH_OR，因其满足，已为终止状态，停止匹配，将 NAME 变量的实例“李国杰”作为 person 输出，将 DATE 变量的实例“1943 年 5 月”作为 birthday 输出，同时对 NAME_BIRTHDAY 建立倒排索引，方便后续作为嵌套规则的一部分进行查找 ;NAME_COLLEGE1 规则还需 NAME_COLLEGE1 子规则，在倒排索引中没有，则终止匹配，NAME_COLLEGE2 规则同理终止匹配。NAME 状态终止后，从下一个匹配到的初始状态 DATE 状态开始继续向上匹配，可转移 NAME_COLLEGE1 子规则、NAME_COLLEGE2 规则，依次查看是否满足转移条件，NAME_COLLEGE1 子规则还需

GRADUTE 和 ORG 在间距 10 以内，在倒排索引中存在且满足，NAME_COLLEGE1 子规则匹配成功，建立倒排索引，并向上继续匹配 NAME_COLLEGE1 规则，NAME_COLLEGE1 规则还需 NAME 同时出现在该子句中，满足，NAME_COLLEGE1 规则匹配成功，建立倒排索引，已为终止状态，停止匹配，输出 NAME 实例“person：李国杰”、DATE 实例“graduatetime:1968 年”、ORG 实例“college :北京大学”。继续查看可以转移的状态，自底向上实现复杂嵌套规则的匹配，如图 7 所示。

[0157] 其中，在匹配过程中动态维护各状态匹配文本的倒排索引，最终生成的倒排索引如图 8 所示。

[0158] 以上子步骤通过对所有规则生成统一的层叠有穷状态自动机，自底向上遍历，避免了相同子规则的重复匹配，并在匹配过程中，自底向上建立倒排索引，加快了匹配的速度。

[0159] 参考图 9，描述了构建统计模型（具体地，条件随机场模型）以及基于该统计模型的属性抽取方法。

[0160] 概括而言，该方法测试不同文本特征对条件随机场模型的效果，选取最佳文本特征（即下文的词语、依存关系、词性、词频），设置模板文件参数。提取在线百科页面属性框中已有的实体属性关系，回标以自动生成训练数据，对每个属性分别训练条件随机场模型 M_{CRF}，并对得到的文本候选集 U 进行属性抽取。

[0161] 详细来说，该方法包括以下子步骤：

[0162] 步骤 S501、获取训练实体及训练属性。

[0163] 获取在线百科的属性框（Infobox）的内容，生成已知的＜实体－属性＞集合，从而得到训练实体和训练属性。

[0164] 其中，Infobox 是在线百科词条页面中结构化描述词条属性的一个表格性区域。

[0165] 步骤 S502、将训练实体在训练用开放网页集合中分别按照步骤 S102 中描述的方法进行迭代查询扩展，从而得到用于训练的文本集合（或称训练文本集合）。

[0166] 步骤 S503-S507 是提取训练数据特征的步骤。其中：

[0167] 步骤 S503、对训练文本集合中的文本进行句子切分，以句子为单位进行模型训练；

[0168] 步骤 S504、对句子进行分词，得到句子中含有的词语；

[0169] 步骤 S505、对每个词标注词性，如名词、代词、动词、形容词等；

[0170] 步骤 S506、对每个词进行依存关系分析，处理词语之间的支配关系，例如可使用依存关系树完成该过程；

[0171] 步骤 S507、计算每个词的词频，即每个词在文本中出现的次数。

[0172] 将句子的词语、依存关系、词性、词频作为特征提取出来，作为机器学习的特征。

[0173] 步骤 S508、生成训练数据。

[0174] 将步骤 S501 生成的每个已知＜实体－属性＞对，对其特征数据进行回标（其中，“回标”是指标出各句子特征是该属性的正例还是反例，将回标完成的特征数据作为训练数据），从而生成各属性的训练数据。

[0175] 其中，生成训练数据包括正例的生成和反例的生成，具体的实现过程如下：已知“李国杰”的工作单位是“中科院计算所”，则将实体“李国杰”的候选文本集合中，所有含有

“中科院计算所”的句子的特征数据回标为正例；而对剩余句子进行命名实体识别，将含有组织的句子的特征数据回标为反例。

[0176] 步骤 S509、人工校正。

[0177] 人工校正子步骤 S508 中生成的训练数据，包括但不限于：

[0178] (1) 去除杂质。例如，已知某实体的工作单位是中科院计算所，在自动生成的训练数据中，会把“中科院计算所”全部标注为正例，如“在中科院计算所工作”，标注为正例，而“中科院计算所位于中科院路”，会被错误地标注成了正例，需人工校正，去除错误的标注；

[0179] (2) 控制正例、反例比例。若正例过多，抽取结果会引入较多杂质；若反例过多，抽取结果召回率会过低，因此需控制正例、反例的比例（如 1 :3）。

[0180] 步骤 S510、制定模板文件，据上下文信息，确定行、列、条件概率的窗口大小。

[0181] 步骤 S511、输入步骤 S509 生成的训练数据，通过有监督的机器学习，为每个属性生成 CRF 模型 M_{CRF}。

[0182] 步骤 S512、根据生成的 CRF 模型抽取属性值。

[0183] 其中，抽取步骤 S102 中得到的候选文本集合 U 的特征，该特征抽取过程同上述子步骤 S503、步骤 S504、步骤 S505、步骤 S506 和步骤 S507。

[0184] 按目标属性选择相应的 M_{CRF}，根据抽取的特征及相应的 M_{CRF} 对目标实体的目标属性进行抽取，得到属性值。

[0185] 步骤 S104 :属性校正

[0186] 根据属性本身的词性、类型、范围等限制，对抽取出的属性值进行校正，剔除不符合要求的属性。例如，人的儿女、配偶、父母等属性，类型也应为人；再例如，描述年龄的属性，通常为 1-120 之间的数字。

[0187] 在一个实施例中，属性的校正规则包括但不限于：

[0188] 1)、属性类型校验：通过某属性对类型的限制，判断抽取的属性值的正误，剔除与类型不匹配的属性值；

[0189] 2)、取值范围校验：根据某属性的词性（如名词、数量词、日期等）及数据范围，剔除超出词性范围的属性值。

[0190] 根据本发明的一个实施例，还提供一种面向开放网页的实体属性抽取系统，包括网页预处理模块、查询扩展模块、属性抽取模块和属性校正模块。

[0191] 其中，网页预处理模块用于提取开放网页的文本，建立倒排索引；查询扩展模块用于从提取出的文本中获得目标实体的候选文本集合；属性抽取模块用于根据目标实体属性在训练文本集合中出现的频率，选择基于规则的方式或者基于统计的方式从候选文本集合中抽取目标实体属性的值；以及，属性校正模块用于对抽取出的属性值进行校正。

[0192] 为验证本发明提供的面向开放网页的实体属性抽取方法和系统的有效性，发明人使用 2014 年 TAC KBP 的 Slot Filling 评测的数据进行了实验，实验参数如下：

[0193] 实验数据集包括 100 个实体（其中“人”类型 50 个，“组织”类型 50 个），要抽取的属性共 41 个（其中“人”类型目标属性 25 个，“组织”类型目标属性 16 个）。其中，有 30 种属性出现频率高，具有充足的训练数据，选用 CRF 的方法训练模型进行属性抽取，剩下的 11 种属性出现频率较低，采用基于 CFT 制定规则的方法对属性进行抽取。该实验数据集中，包含的属性值共计 1001 个。

[0194] 在实验过程中，发现了最佳参数配置。其中，查询扩展时选取扩展窗口为 1，每次扩展选取前 50 篇文本；CRF 训练选取 4 个文本特征：句子包含的词、各词的词性、各词的词频、各词之间的依存关系。在生成 CRF 模型的训练数据时，采用的正例、反例的比例为 1 :3，通过验证发现该配置能够达到最高的召回率和准确率。

[0195] 经过实验，得出如下结果：

[0196] 共抽取结果为 412 个，命中 243 个，准确率为 58.98％。而现有抽取技术中，Stanford 大学的自然语言处理组的通过对实体对之间关系词的机器学习方法表现最好，准确率达到 58.54％；Rensselaer Polytechnic Institute 的 RPI BLENDER 团队多策略综合搜索的方法表现较好，准确率达到 47.80％。本发明的准确率均高于现有抽取技术，因此，本发明比现有属性抽取技术更为精确。

[0197] 应当理解，虽然本说明书是按照各个实施例描述的，但并非每个实施例仅包含一个独立的技术方案，说明书的这种叙述方式仅仅是为清楚起见，本领域技术人员应当将说明书作为一个整体，各实施例中的技术方案也可以经适当组合，形成本领域技术人员可以理解的其他实施方式。

[0198] 以上所述仅为本发明示意性的具体实施方式，并非用以限定本发明的范围。任何本领域的技术人员，在不脱离本发明的构思和原则的前提下所作的等同变化、修改与结合，均应属于本发明保护的范围。

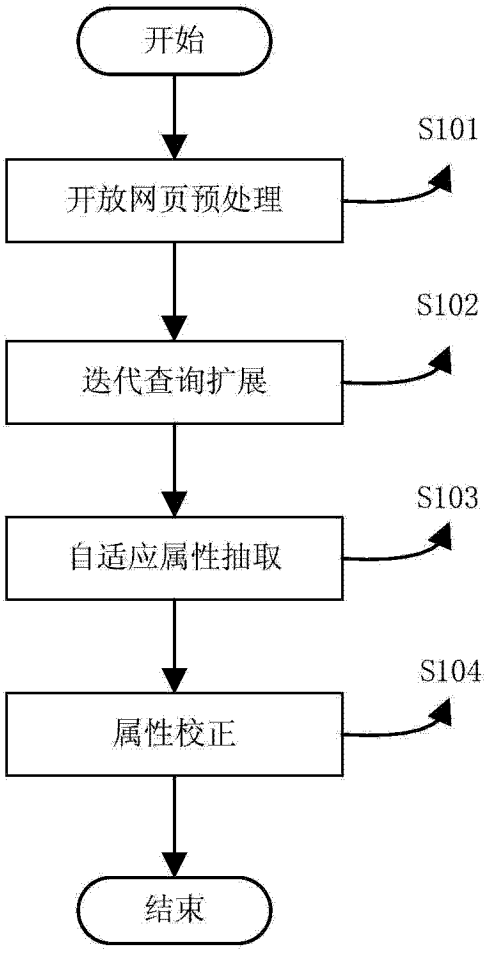


图 1

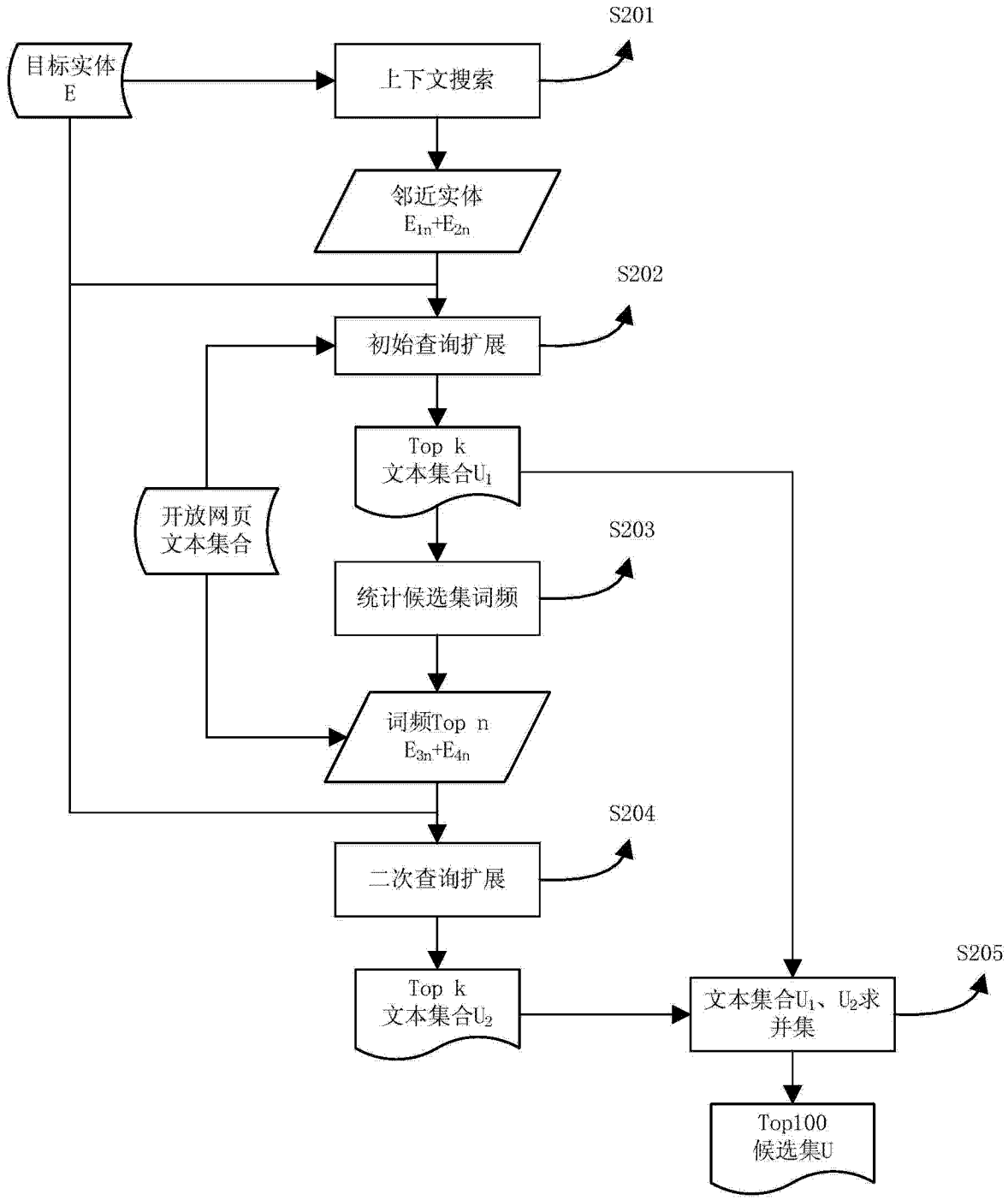

目标实体
E
上下文搜索
S201
邻近实体
E_{1n}+E_{2n}
S202
初始查询扩展
Top k
文本集合U_1
开放网页
文本集合
S203
统计候选集词频
词频Top n
E_{3n}+E_{4n}
S204
二次查询扩展
Top k
文本集合U_2
文本集合U_1、U_2求
并集
S205
Top100
候选集U

图 2

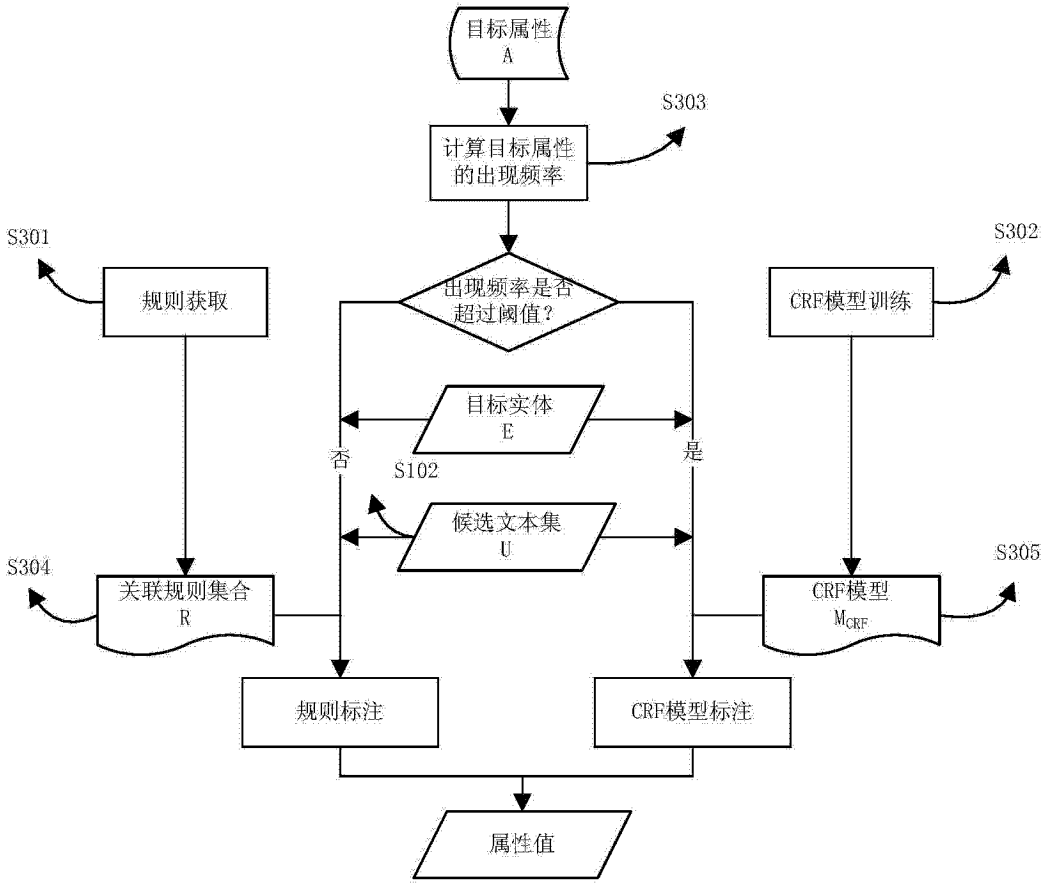

目标属性
A
S303
计算目标属性
的出现频率
S301
规则获取
出现频率是否
超过阈值?
S302
CRF模型训练
目标实体
E
否
S102
是
候选文本集
U
S304
关联规则集合
R
CRF模型
M_{CRF}
S305
规则标注
CRF模型标注
属性值

图 3

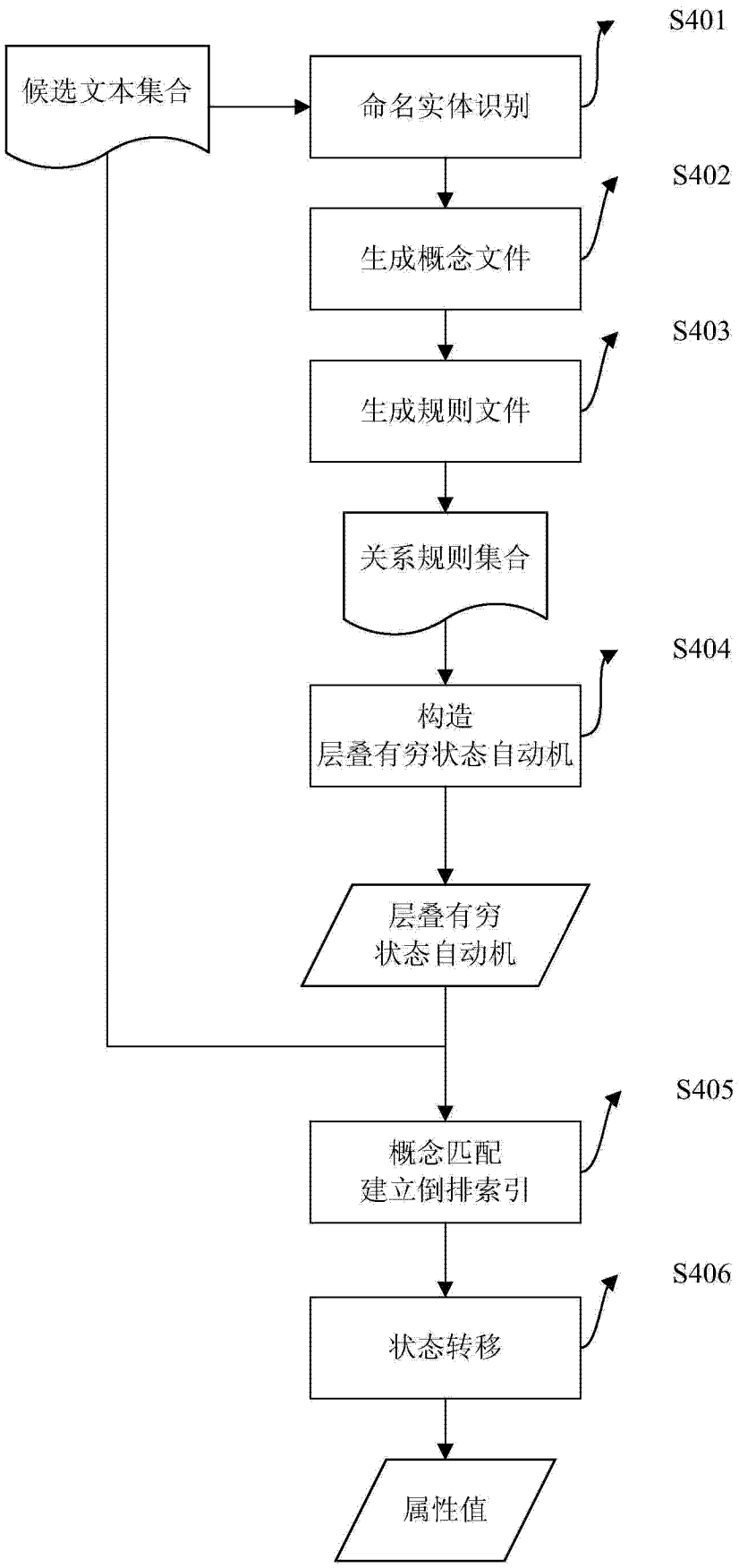

候选文本集合
命名实体识别
S401
生成概念文件
S402
生成规则文件
S403
关系规则集合
构造
层叠有穷状态自动机
S404
层叠有穷
状态自动机
概念匹配
建立倒排索引
S405
状态转移
S406
属性值

图 4

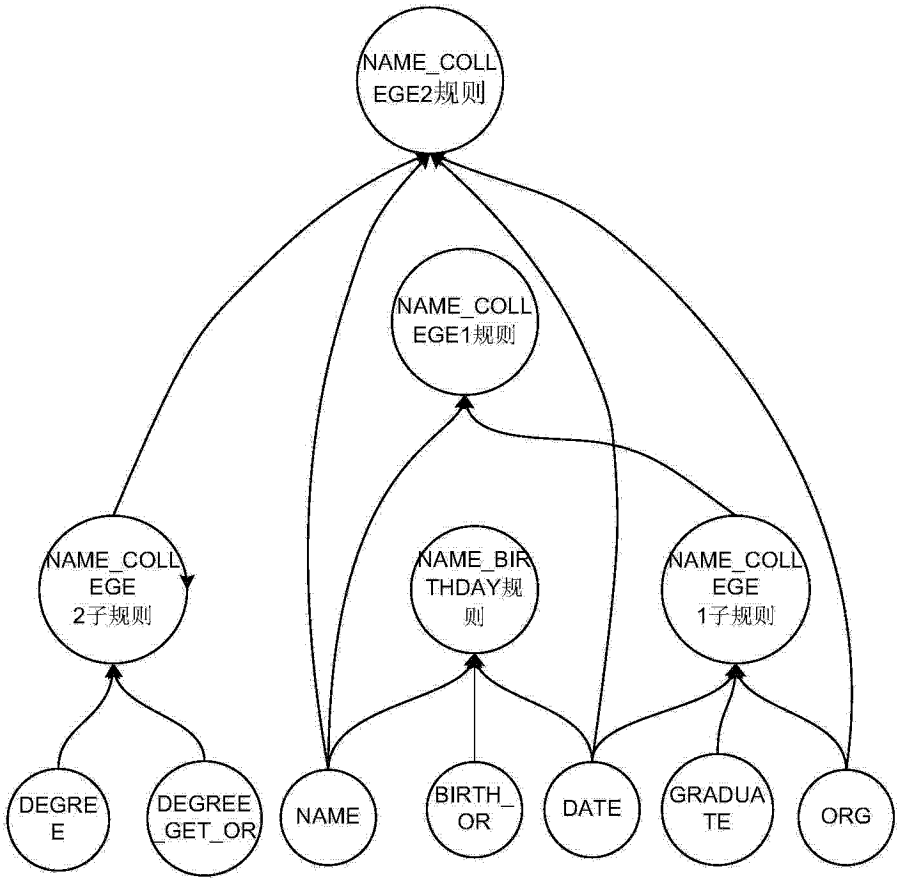

NAME_COLLEGE2规则
NAME_COLLEGE1规则
NAME_COLLEGE 2子规则
NAME_BIRTHDAY规则
NAME_COLLEGE 1子规则
DEGREE
DEGREE_GET_OR
NAME
BIRTH_OR
DATE
GRADUATE
ORG

图 5

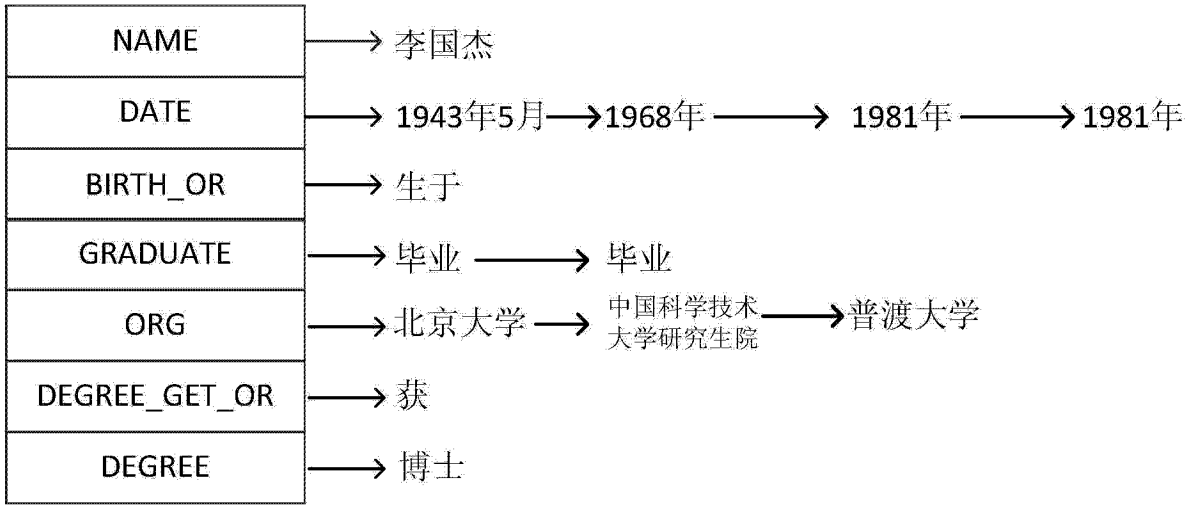

NAME
李国杰
DATE
1943年5月
1968年
1981年
1981年
BIRTH_OR
生于
GRADUATE
毕业
毕业
ORG
北京大学
中国科学技术大学研究生院
普渡大学
DEGREE_GET_OR
获
DEGREE
博士

图 6

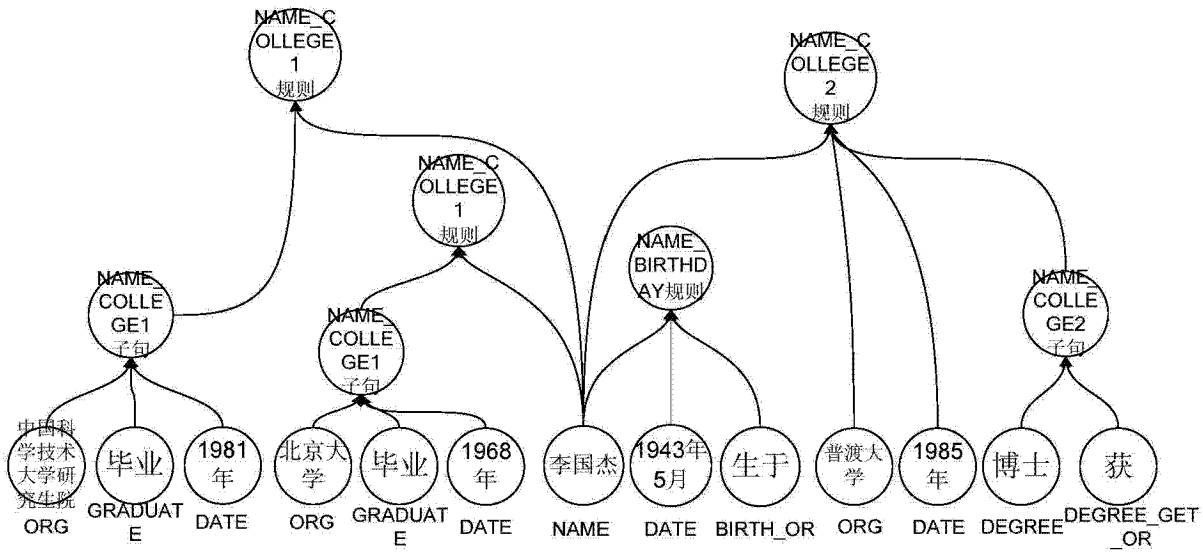

NAME_COLLEGE1规则
NAME_COLLEGE1规则
NAME_COLLEGE2规则
NAME_BIRTHDAY规则
NAME_COLLEGE1子句
NAME_COLLEGE1子句
NAME_COLLEGE2子句
中国科学技术大学研究生院
毕业
1981年
北京大学
毕业
1968年
李国杰
1943年5月
生于
普渡大学
1985年
博士
获
ORG
GRADUATE
DATE
ORG
GRADUATE
DATE
NAME
DATE
BIRTH_OR
ORG
DATE
DEGREE
DEGREE_GET_OR

图 7

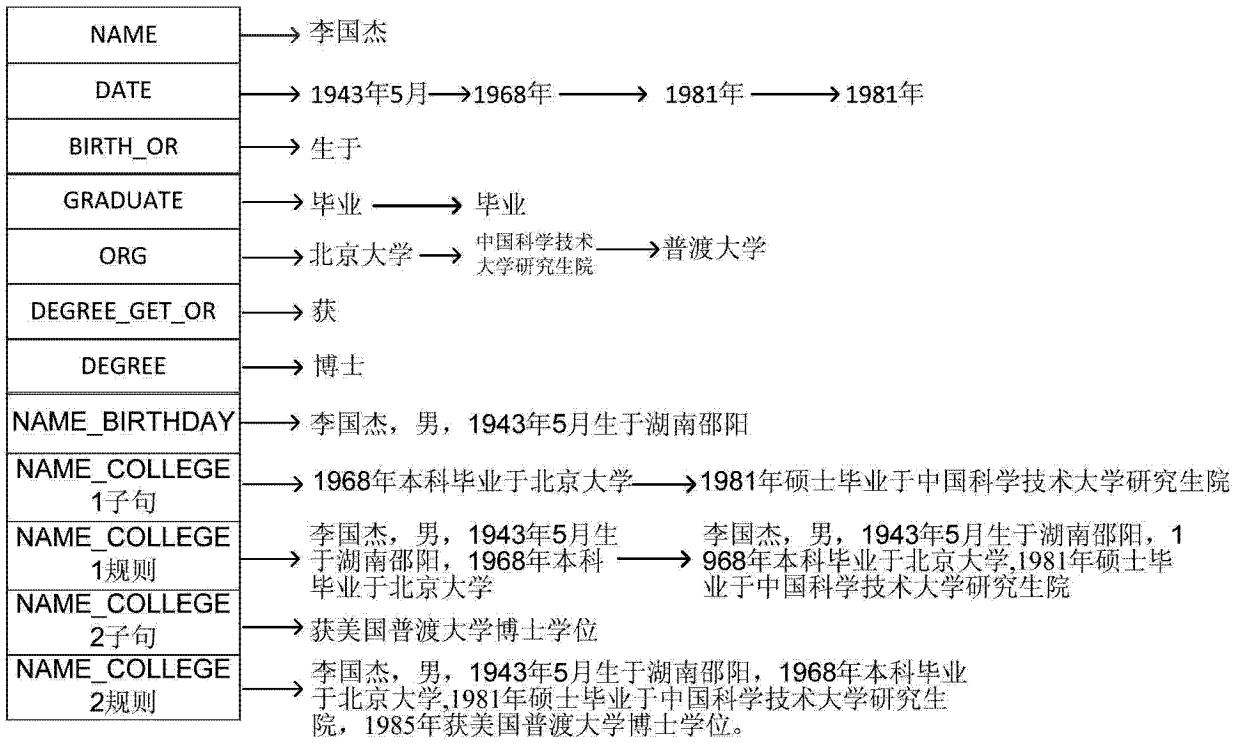

NAME
李国杰
DATE
1943年5月→1968年→1981年→1981年
BIRTH_OR
生于
GRADUATE
毕业→毕业
ORG
北京大学→中国科学技术大学研究生院→普渡大学
DEGREE_GET_OR
获
DEGREE
博士
NAME_BIRTHDAY
李国杰，男，1943年5月生于湖南邵阳
NAME_COLLEGE 1子句
1968年本科毕业于北京大学→1981年硕士毕业于中国科学技术大学研究生院
NAME_COLLEGE 1规则
李国杰，男，1943年5月生于湖南邵阳，1968年本科毕业于北京大学
李国杰，男，1943年5月生于湖南邵阳，1968年本科毕业于北京大学,1981年硕士毕业于中国科学技术大学研究生院
NAME_COLLEGE 2子句
获美国普渡大学博士学位
NAME_COLLEGE 2规则
李国杰，男，1943年5月生于湖南邵阳，1968年本科毕业于北京大学,1981年硕士毕业于中国科学技术大学研究生院，1985年获美国普渡大学博士学位。

图 8

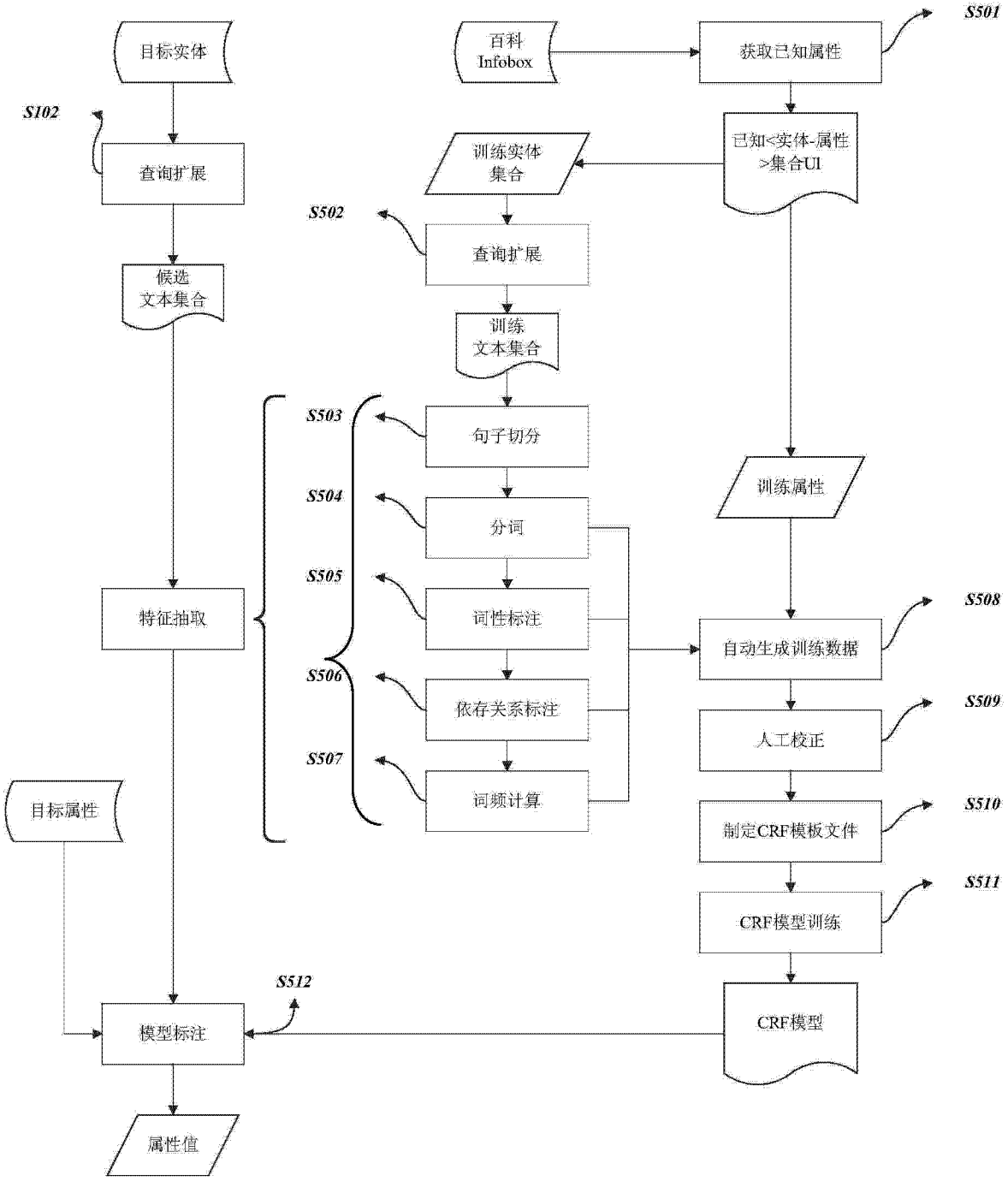

目标实体
S102
查询扩展
候选
文本集合
特征抽取
目标属性
模型标注
属性值
S512
百科
Infobox
获取已知属性
S501
已知<实体-属性
>集合UI
训练实体
集合
S502
查询扩展
训练
文本集合
S503
句子切分
S504
分词
S505
词性标注
S506
依存关系标注
S507
词频计算
训练属性
S508
自动生成训练数据
S509
人工校正
S510
制定CRF模板文件
S511
CRF模型训练
CRF模型

图 9