(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 102254014 B
(45)授权公告日 2013.06.05

(21)申请号 201110205137.7
(22)申请日 2011.07.21
(73)专利权人 华中科技大学
地址 430074 湖北省武汉市洪山区珞喻路1037号
(72)发明人 金海 李毅 赵峰 严奉伟 陈恒
(74)专利代理机构 华中科技大学专利中心 42201
代理人 曹葆青
(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 101727498 A,2010.06.09,
JP 2004-46312 A,2004.02.12,
US 2010083095 A1,2010.04.01,
CN 101620608 A,2010.01.06,

审查员 徐琳

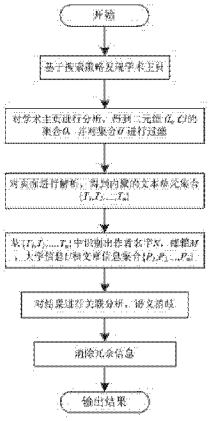
权利要求书3页 说明书7页 附图5页

(54)发明名称

一种网页特征自适应的信息抽取方法

(57)摘要

本发明公开了一种从学术主页中抽取信息的方法，其步骤为：(1)在互联网中发现学术主页；(2)对学术主页进行爬取和解析，使用启发式策略减少无关页面的爬取，加快解析速度；(3)将页面解析成DOM树的形式，并按照元素的属性和内容进行划分，得到内聚的文本单元列表；(4)使用信息识别器对文本单元进行识别，每种信息识别器只识别一种信息类型，对于文章信息还需要进行子字段提取。(5)对抽取结果进行关联分析，利用信息的关联性消除歧义，对缺失字段进行补全；(6)将抽取结果与数据库进行匹配，消除冗余数据，抽取结果以语义数据的形式保存在语义数据库中。本发明通过结合使用启发式规则，机器学习方法和条件概率模型能够高效准确的从学术主页中抽取学术信息。



CN 102254014 B

1. 一种网页特征自适应的信息抽取方法，其特征在于，该方法包括下述步骤：

第 1 步从互联网中搜寻类型为学术主页的站点；

第 2 步对搜寻的学术主页进行分析，将学术主页的页面看成是二元组 (L,C) 的集合，其中 L 是学术主页中链接的 URL，C 是学术主页中链接的上下文，再检查 L 和 C 中是否包含关键字，如果包含，则进入第 3 步，否则过滤掉该链接；

第 3 步对所述链接的页面进行分析，得到页面的文档树结构，根据树节点的属性和内容对页面进行划分，分成文本单元 T，构成文本单元集合 $\{T_1, T_2, \cdots, T_n\}$，步骤如下：

(a) 首先使用 HTML 解析器对页面进行解析，得到页面的文档树；文档树的节点即对应于页面里的 HTML 标签，文档树以树形结构展现出页面里各个 HTML 标签之间的关系；

(b) 然后对页面进行划分；HTML 标签分为块级元素和内联元素，HTML 页面被看做是块级元素的集合，块级元素之间拥有两种关系：父子关系和兄弟关系，块级元素和内联元素之间能够相互嵌套，文档树就是以树节点的形式呈现出这些关系，文档树中含有块级元素的节点称为块级节点，其他节点称为非块级节点，对文档树的节点进行遍历，通过判断节点的类别来对页面进行划分，划分步骤如下：

(b1) 初始，文本单元集合为空；

(b2) 对文档树进行深度优先遍历，找出所有的块级节点，对每一个块级节点 Ni，生成一个文本单元 Ti，并将 Ni 在页面中相应的内容划分至 Ti；

(b3) 对每一个块级节点 Ni，判断其在文档树中是否有非块级子节点，如果有则将其所有非块级子节点在页面中相应的内容划分至 Ti；

(b4) 将 Ti 加入文本单元集合中；

(b5) 结束；

(c) 遍历结束后，完成页面的划分，得到文本单元集合 $\{T_1, T_2, \cdots, T_n\}$；

第 4 步从文本单元集合 $\{T_1, T_2, \cdots, T_n\}$ 中抽取出作者名字 N，邮箱 M，机构信息 U 和文章信息集合 {P1,P2,…,Pm} 这四个目标字段，作为初步抽取结果；针对不同类型的目标字段，其抽取方法如下：

作者名字 N 的抽取过程如下：

(a1) 使用支持向量机分类算法对文本单元集合 $\{T_1, T_2, \cdots, T_n\}$ 里的文本单元进行分类，保留类别为作者名字的文本单元集合 T_{name}；

(a2) 使用作者名字数据库从 T_{name} 中匹配出作者名字部分，使用该数据库从 T_{name} 中匹配出候选的作者名字集合；

(a3) 提取出作者学术主页标题中的文字，提取作者学术主页标题中的作者名字 XXX；

(a4) 用 (a3) 得到的作者名字 XXX 对 (a2) 得到的候选作者名字进行匹配，选择与 XXX 匹配程度最高的名字作为作者名字 N 输出；

邮箱 M 的抽取过程如下：

(b1) 首先使用支持向量机分类器从文本单元集合 $\{T_1, T_2, \cdots, T_n\}$ 中找出可能的邮箱候选文本单元集合 T_{Email}；支持向量机算法根据特征向量对 T_{Email} 中邮箱候选文本单元进行判定，如果分类结果为肯定，则进行 (b2) 处理，否则直接过滤掉；

(b2) 去掉邮箱候选文本单元中多余的部分；

(b3) 采用模糊匹配状态机算法对邮箱候选文本单元与标准的邮箱进行匹配，一个标准

的邮箱有如下字段：用户名 @（提供商域名 .）+. 顶级域名，生成不同的匹配结果；

(b4) 将邮箱候选文本单元的各个字段和匹配结果进行比对，选取匹配程度最大的结果作为最终结果，并按照标准的邮箱字段将其转换为规范的合法邮箱格式输出；

机构信息 U 的抽取过程如下：

(c1) 首先从互联网上收集全球大学和研究所的数据，包括机构的名字和其对应的主页链接，建立一个机构主页数据库；为数据库建立倒排索引；

(c2) 使用支持向量机分类器从文本单元集合 $\{T_1, T_2, \cdots, T_n\}$ 中找出可能的机构信息文本单元集合 T_U，将 T_U 中的机构信息文本单元转换为文本形式，将其作为关键字在索引中查找，取得排名前三的检索结果；将前三个检索结果和相应的机构信息文本单元进行模糊匹配，如果能够匹配上则确定该文本是对应该机构的，将匹配程度最高的匹配结果输出，否则如果均无法匹配上，则转 (c3) 处理；

(c3) 利用学术主页的 URL 进行寻找，学术站点通常是机构站点的子站点，因此将学术主页的域名与机构主页数据库进行匹配，如果存在匹配的记录，则认为作者属于该机构，将匹配的记录作为结果输出；

文章信息 {P1,P2,…,Pm} 的抽取的过程如下：

(a) 首先使用支持向量机分类算法对文本单元进行分类，筛选出可能包含文章信息的文本单元；

(b) 然后对候选文章信息文本单元进行序列标注，提取候选文本中各个子字段：从候选文章信息文本单元中提取出文本类特征、模式特征、词典特征和术语特征，用于条件随机场模型，条件随机场模型中的特征函数使用真值形式，即函数输出是或者否；经过条件随机场模型的计算，给出候选文章信息文本单元的最可能的标注形式；具有相同标签的符号会被合并成相应的子字段，然后分别对这些字段进行相应的后续处理；

(c) 作者名字段包含了整个作者列表，需要分割成单个作者的形式；分割算法基于启发式规则，依据名字的长度，缩写形式以及标点符号；分割后的结果被保存在数组中；

标题字段经过规范化裁剪作为最终的结果；

会议期刊名和文献期刊数据库中的进行匹配；首先提取出待识别字段 中大写字母缩写部分，在数据库中进行查找，如果匹配则将匹配的全称与待识别字段进行模糊匹配，防止缩写形式冲突的情况导致的错误；若匹配则直接输出结果；否则为会议期刊名建立索引，将待识别字段在索引中进行检索，将检索结果与待识别字段做模糊匹配；若找到匹配则输出结果；

年份字段使用规则化方法，使用正则表达式在输入文本中寻找合法的年份模式；

第 5 步对第 4 步得到的初步抽取结果进行关联分析，利用信息的关联性消除歧义，对缺失字段进行补全，得到抽取结果，存至结果数据库中；

第 6 步将文章信息集合 {P1,P2,…,Pm} 中的元素与结果数据库中的记录进行匹配，消除冗余数据；

第 7 步输出抽取结果。

2. 根据权利要求 1 所述的信息抽取方法，其特征在于，第 1 步分为两个阶段：寻找阶段和判定阶段；

在寻找阶段，首先从已有的文献数据中导出作者名字的数据集作为种子数据，然后以

数据集中的每一个作者名作为关键字在搜索引擎中进行检索，搜索引擎以列表形式返回检索结果，每一条检索结果由标题，链接特征和摘要文本组成，并将返回结果中的第一页的检索结果的链接特征和摘要文本存放在候选结果列表中；

在判定阶段，首先根据检索结果的链接特征和摘要文本对候选结果列表按下述方式进行过滤，首先检查链接是否存在于屏蔽链接数据库中，将位于该数据库中的结果直接排除，然后，对剩余的检索结果，检查其链接特征是否呈现为“～”+ 作者名字的模式，如果是则保留，否则则直接排除，经过这两步过滤之再依次对剩余的每一条检索结果进行如下操作：根据其链接特征发出页面请求，使用支持向量机分类算法判定返回的页面是否是作者学术主页，如果是，则直接将其保存为作者学术主页，判定结束，否则继续对下一条检索结果进行相同的操作。

一种网页特征自适应的信息抽取方法

技术领域

[0001] 本发明属于信息抽取系统领域，具体涉及一种网页特征自适应的信息抽取方法，该方法尤其适用于从学术主页中抽取作者名字，邮箱，机构信息和发表文章等信息。

背景技术

[0002] 信息时代的来临使得网络逐渐成为人们分享和获取信息的主要途径，各种信息以网页的形式发布在互联网上供人们阅读。然而随着互联网信息的爆炸性增长，人们发现在互联网中找到所需的信息变得越来越困难，一方面信息量巨大，另一方面信息呈现的方式非常灵活和自由，这增加了人们辨别目标信息的成本。因此，网页信息抽取技术成为信息时代值得研究的领域。

[0003] 网页信息抽取技术是从传统的文本信息抽取上发展起来的。跟文本信息不同，网页内容是用超文本标记语言 (HTML) 表述的，包含文本，图片和其他多媒体信息，且标记之间允许相互嵌套形成树状的结构。网页信息抽取任务的主要目的是从半结构化的网页文本中抽取出目标信息。网页信息通常具有如下特征 :(1) 离散化，信息并不集中在某一站点，而是由不同的人发布到不同的站点上。(2) 异构性，即使是同类的信息在不同的网站上也会使用不同的方式呈现。(3) 冗余性，相同的信息可能会在多个站点上重复出现。针对网页信息的这些特征，网页信息抽取系统需要能够具有较强的适应能力和辨别能力。

[0004] 早期的网页信息抽取研究集中探索了规则化方法，从基于正则表达式的脚本化抽取方法，到之后发展起来的专有的抽取语言，其核心思想是提取出包含目标信息的特定模式。模式的提取的方法是这类系统的主要不同，一些系统使用手工方式来提取模式，这样的好处是提取的模式更加准确，不过在处理复杂抽取任务时需要提取模式将非常之多，因此人工成本较高。为了降低模式提取的成本，人们提出了基于自动训练的模式学习系统，系统需要接受一组训练样例，样例由人工标识出其中的目标信息块，学习系统自动的根据从样例中总结出可能的匹配模式，模式经过验证和筛选后被用于实际的抽取任务。该方法具有了一定的自动提取能力，但是由于底层仍然依赖于规则化方法，因此对复杂的抽取任务无法达到较高的准确率。最近几年来，抽取方法逐渐转向于机器学习模型，一些原本在处理自然语言理解过程中的方法被应用来处理信息抽取问题，取得了很好的效果。

[0005] 学术主页是学术领域内的研究人员用来展示自己个人基本信息和研究成果的站点。不同的作者根据自己的喜好制作不同的页面模板呈现个人信息。尽管页面风格各不相同，但是学术主页上通常包含了类似的信息，如作者名字，机构信息，联系方式，项目，文章信息等。使用信息抽取系统将这些信息收集起来是十分有价值的。

发明内容

[0006] 本发明的目的是提供一种网页特征自适应的信息抽取方法，该方法能够从不同风格的学术主页中提取所需的信息，并且具有适应能力强，准确率高，以及扩展性强特点。

[0007] 本发明提供的一种网页特征自适应的信息抽取方法，其特征在于，该方法包括下

述步骤：

[0008] 第 1 步从互联网中搜寻类型为学术主页的站点；

[0009] 第 2 步对搜寻的学术主页进行分析，将学术主页的页面看成是二元组 (L，C) 的集合，其中 L 是链接的 URL，C 是链接的上下文，再检查 L 和 C 中是否包含关键字，如果包含，则进入第 3 步，否则过滤掉该链接；

[0010] 第 3 步对所述链接进行分析，得到页面的文档树结构，根据树节点的属性和内容对页面进行划分，分成文本单元 T，构成文本单元集合 $\{T_1, T_2, \ldots, T_n\}$

[0011] 第 4 步从文本单元集合 $\{T_1, T_2, \ldots, T_n\}$ 中抽取出作者名字 N，邮箱 M，机构信息 U 和文章信息集合 $\{P_1, P_2, \ldots, P_n\}$ 这四个目标字段，作为初步抽取结果；

[0012] 第 5 步对第 4 步得到的初步抽取结果进行关联分析，利用信息的关联性消除歧义，对缺失字段进行补全，得到抽取结果，存至结果数据库中；

[0013] 第 6 步将文章信息集合 $\{P_1, P_2, \ldots, P_n\}$ 中的元素与结果数据库中的记录进行匹配，消除冗余数据；

[0014] 第 7 步输出抽取结果。

[0015] 本发明提供的一种网页特征自适应的信息抽取方法，该方法结合使用了机器学习算法，概率模型和规则化方法，能够从不同风格的学术主页中提取出作者的名字，邮箱，机构信息和发表文章等信息。具体而言，本发明有以下效果和优点：

[0016] (1) 适应性强

[0017] 学术主页的编写者是许多不同的研究者，内容和排版各式各样。本发明能够很好的解决页面格式不统一的的问题，自动的适应各种变化情况；

[0018] (2) 准确度高

[0019] 本发明的核心算法基于机器学习算法和概率模型，并结合使用了启发式规则，对各个目标字段的抽取都能够达到很高的准确率；

[0020] (3) 可扩展性强

[0021] 本发明能够被扩展来提取出页面中的其他字段，其识别过程也能够被应用来解决其他类似问题，扩展过程简单，通用性强。

附图说明

[0022] 图 1 为本发明的抽取过程的整体流程图；

[0023] 图 2 为本发明对作者名进行抽取的流程图；

[0024] 图 3 为本发明对邮箱进行抽取的流程图；

[0025] 图 4 为本发明对机构信息进行抽取的流程图；

[0026] 图 5 为本发明对文章信息进行抽取的流程图。

具体实施方式

[0027] 下面结合附图和实例对本发明进行详细说明。

[0028] 本发明提供的一种网页特征自适应的信息抽取方法，其步骤包括：

[0029] (1) 从互联网中搜寻类型为学术主页的站点，该过程可以分为两个阶段：寻找阶段和判定阶段。

[0030] 在寻找阶段，首先从已有的文献数据中导出作者名字的数据集作为种子数据，然后以数据集中的每一个作者名作为关键字在搜索引擎中进行检索，搜索引擎以列表形式返回检索结果，每一条检索结果通常由标题，链接特征和一小段摘要文本组成，搜索引擎通常会返回多页结果，将第一页的检索结果的链接特征和摘要文本存放在候选结果列表中。

[0031] 在判定阶段，首先根据链接特征和摘要文本对候选结果列表中的检索结果进行过滤。过滤过程中用到了一个数据库，该数据库包含了检索结果中经常出现的混淆站点，称之为屏蔽链接数据库。过滤策略包含两个步骤，首先检查检索结果是否存在于屏蔽链接数据库中，将位于该数据库中的检索结果直接排除。然后，对剩余的检索结果，检查其链接特征是否呈现为“～”+ 作者名字的模式，如果是则保留，否则则直接排除，经过这两步过滤之再依次对剩余的每一条检索结果进行如下操作：根据其链接特征发出页面请求，使用支持向量机分类算法判定返回的页面是否是作者学术主页，如果是，则直接将其保存为作者学术主页，判定结束，否则继续对下一条检索结果进行相同的操作。

[0032] (2) 对作者学术主页进行分析，作者学术主页通常是一个完整的站点，包含了许多子页面，其中有些包含了系统需要的目标信息，有些则是完全无关的。为了提高爬取效率，避免过多的无用页面被后续模块进行深入解析，消耗计算资源，本发明使用了一种基于启发式策略的过滤算法。该算法将页面看成是二元组 (L，C) 的集合，其中 L 是链接的 URL，C 是链接的上下文，该算法检查 L 和 C 中是否包含 publication，paper，research 等关键字，如果包含则进一步解析该链接 (进入步骤 (3))，否则过滤掉该链接。

[0033] (3) 对待解析页面进行分析，得到网页的文档树结构，根据文档树节点的属性和内容对页面进行划分，分成若干个小单元，称之为文本单元 T，划分结果为文本单元集合 $\{T_1, T_2, \ldots, T_n\}$，步骤如下。

[0034] (a) 首先使用 HTML 解析器对页面进行解析，得到页面的文档树。文档树的节点即对应于页面里的 HTML 标签，文档树以树形结构展现出页面里各个 HTML 标签之间的关系。

[0035] (b) 然后对页面进行划分。HTML 标签可以分为块级元素和内联元素，常见的块级元素如 BR，DIV，H1，H2，LI，UL，TH，TD，TR，TABLE 等，常见的内联元素如 SPAN，BOLD，A，FONT，IMG 等。HTML 页面可以被看做是块级元素的集合，块级元素之间拥有两种关系：父子关系和兄弟关系。块级元素和内联元素之间可以相互嵌套。文档树就是以树节点的形式呈现出这些关系，文档树中含有块级元素的节点称为块级节点，其他节点称为非块级节点，对文档树的节点进行遍历，通过判断节点的类别来对页面进行划分，划分步骤如下：

[0036] (b1) 初始，文本单元集合为空；

[0037] (b2) 对文档树进行深度优先遍历，找出所有的块级节点，对每一个块级节点 Ni，生成一个文本单元 Ti，并将 Ni 在页面中相应的内容划分至 Ti；

[0038] (b3) 对每一个块级子节点 Ni，判断其在文档树中是否有非块级子节点，如果有则将其所有非块级子节点在页面中相应的内容划分至 Ti；

[0039] (b4) 将 Ti 加入文本单元集合中；

[0040] (b5) 结束。

[0041] (c) 遍历结束后，完成页面的划分，得到文本单元集合 $\{T_1, T_2, \ldots, T_n\}$。

[0042] (4) 从文本单元集合 $\{T_1, T_2, \ldots, T_n\}$ 中抽取出作者名字 N，邮箱 M，机构信息 U 和文章信息集合 $\{P_1, P_2, \ldots, P_n\}$ 这四个目标字段，作为初步抽取结果；

[0043] 针对不同类型的目标字段，下面分别介绍不同字段的抽取方法：

[0044] 作者名字 N 的抽取过程如图 2 所示，其基本步骤如下：

[0045] (a1) 使用支持向量机分类算法对文本单元集合 $\{T_1, T_2, \ldots, T_n\}$ 里的文本单元进行分类，保留类别为作者名字的文本单元集合 T_{name}；

[0046] (a2) 使用作者名字数据库从 T_{name} 中匹配出作者名字部分，作者名字数据库是一个事先准备好的数据库，该数据库收集和整理了常见的英文男女人名和一些中文拼音，使用该数据库从 T_{name} 中匹配出候选的作者名字集合；

[0047] (a3) 提取出作者学术主页标题中的文字，大多数时候作者学术主页的标题会以“XXX’s Hompage”的形式包含作者的名字 XXX，提取作者学术主页标题中的作者名字 XXX；

[0048] (a4) 用 (a3) 得到的作者名字 XXX 对 (a2) 得到的候选作者名字进行匹配，选择与 XXX 匹配程度最高的名字作为作者名字 N 输出。

[0049] 邮箱 M 的抽取过程如图 3 所示，其基本步骤如下：

[0050] (b1) 首先使用支持向量机分类器从文本单元集合 $\{T_1, T_2, \ldots, T_n\}$ 中找出可能的邮箱候选文本单元集合 T_{Email}。支持向量机的输入特征包括邮箱信息中的常见符号，如“Email”,“@”,“.”等。在 T_{Email} 中寻找这些特征符号，生成特征向量。支持向量机算法根据特征向量对 T_{Email} 中邮箱候选文本单元进行判定，如果分类结果为肯定，则进行 (b2) 处理，否则直接过滤掉。

[0051] (b2) 去掉邮箱候选文本单元中多余的部分，如提示性前缀“Email：”，去除这些信息有利于后续步骤获得合法的邮箱信息。

[0052] (b3) 接下来采用模糊匹配状态机算法对邮箱候选文本单元进行匹配，一个标准的邮箱有如下字段：用户名 @(提供商域名 .)+. 顶级域名。该算法为每一个字段建立一个匹配节点，使用状态机枚举可能的匹配形式，生成许多不同的匹配结果，通常有几十个。

[0053] (b4) 将邮箱候选文本单元的各个字段和匹配结果进行比对，选取匹配程度最大的结果作为最终结果，并按照标准的邮箱字段将其转换为规范的合法邮箱格式输出。

[0054] 机构信息 U 的抽取过程如图 4 所示，其基本步骤如下：

[0055] (c1) 首先从互联网上收集全球大学和研究所的数据，包括机构的名字和其对应的主页链接，建立一个机构主页数据库。为数据库建立倒排索引。倒排索引支持快速的关键字查找，能够快速确定包含一组关键字的条目。

[0056] (c2) 使用支持向量机分类器从文本单元集合 $\{T_1, T_2, \ldots, T_n\}$ 中找出可能的机构信息文本单元集合 T_U，将 T_U 中的机构信息文本单元转换为文本形式，将其作为关键字在索引中查找，取得排名前三的检索结果。将前三个检索结果和相应的机构信息文本单元进行模糊匹配，如果能够匹配上则确定该文本是对应该机构的，将匹配程度最高的匹配结果输出，否则如果均无法匹配上，则转 (c3) 处理。

[0057] (c3) 利用主页的 URL 进行寻找，学术站点通常是机构站点的子站点，因此将主页的域名与机构主页数据库进行匹配，如果存在匹配的记录，则认为作者属于该所机构，将匹配的记录作为结果输出。

[0058] 文章信息 $\{P_1, P_2, \ldots, P_n\}$ 的抽取的过程如图 5 所示，其基本步骤如下：

[0059] (a) 首先使用支持向量机分类算法对文本单元进行分类，筛选出可能包含文章信息的文本单元。分类算法的准确率与文章信息的最终识别准确率关系密切，分类算法需要

过滤掉课程信息，专利，项目等容易发生混淆的相似信息。分类算法的准确率主要依赖于两个方面：训练样例和特征的选取。训练样例的构建按照迭代法，通过不断的将错误样例添加到训练集中来更正原有模型。特征向量由一组具有区分能力的词汇向量构成。经过分类算法的筛选，无关的文本单元被排除掉，得到候选文章信息文本单元。

[0060] (b) 然后对候选文章信息文本单元进行序列标注，提取候选文本中各个子字段，包括：作者名字，标题，会议期刊名，年份。序列标注的算法基于条件随机场模型，模型中用了下列特征：

[0061] ①文本类特征

[0062] a) 词条本身，包括原始形式和词根形式

[0063] b) 大小写特征，包括首字母大写，全大写，单个大写字母

[0064] c) 数字特征，全数字，数字和字母的混合，罗马字母

[0065] d) 标点特征，逗号，引号，句号等

[0066] e)HTML 标签特征，标签起始，中间部分和结束部分

[0067] ②模式特征

[0068] a) 年份特征，19XX 或者 20XX

[0069] b) 页模式，XXX-XXX

[0070] ③词典特征

[0071] 作者名字，地理位置，出版社，时间，会议期刊名，机构名

[0072] ④术语特征

[0073] 文献数据中常用的词汇，如 pp/editor/volume 等

[0074] 从候选文章信息文本单元中提取出上述特征，条件随机场模型中的特征函数使用真值形式，即函数输出是或者否。经过模型的计算，给出候选文章信息文本单元的最可能的标注形式。具有相同标签的符号会被合并成相应的子字段，如作者名字字段，标题字段，会议期刊字段，年份字段等，然后分别对这些字段进行相应的后续处理。

[0075] (c) 作者名字段包含了整个作者列表，需要分割成单个作者的形式。分割算法基于启发式规则，主要依据与名字的长度，缩写形式以及标点符号。分割后的结果被保存在数组中。

[0076] 标题字段需要经过规范化裁剪才能作为最终的结果。裁剪的主要目的是为了去除掉前缀和后缀的非法字符，比如标点符号，边界错误等。

[0077] 会议期刊名在实际中存在多种表达方式，如大写字母的缩写和常见的习惯称呼等。直接提取的会议期刊字段不能作为最终的结果，需要和数据库中的进行匹配。文献期刊数据库收集了常见的会议和期刊名以及相应的缩写形式。首先提取出待识别字段中大写字母缩写部分，在数据库中进行查找，如果匹配则将匹配的全称与输入字段进行模糊匹配，防止缩写形式冲突的情况导致的错误。若匹配则直接输出结果。否则为会议期刊名建立索引，将待匹配字段在索引中进行检索，将检索结果与待匹配字段做模糊匹配。若找到匹配则输出结果。

[0078] 年份字段使用规则化方法，使用正则表达式在输入文本中寻找合法的年份模式。合法年份模式有两种形式：第一种以 19 或者 20 开始，并且为四位数字；第二种以会议期刊名字的大写字母缩写形式开始，接着引号和年份。使用这两种模式能够处理实际中的绝大

部分情况，识别准确率超过百分之九十九。

[0079] (5) 对步骤 (4) 得到的初步抽取结果（包括作者名字 N，邮箱 M，机构信息 U 和文章信息集合 $\{P_1, P_2, \ldots, P_n\}$）进行缺失字段补全和歧义消除，得到最终的抽取结果，存至结果数据库中。

[0080] 实际页面中包含的信息可能存在一定程度的缺失和不规范的情况，对相同信息项可能识别出多个结果需要进一步判定。该过程利用信息之间的关联关系，对抽取结果进行补全，对存在歧义的结果进行进一步判定。信息关联包含如下情况：

[0081] (a) 作者名和邮箱用户名之间的关联；

[0082] (b) 机构信息与主页域名之间的关联；

[0083] (c) 作者名和文章信息中作者列表的关联；

[0084] 根据上述关联，可以对抽取结果进行补全，如当机构信息存在缺失时，可以将主页链接在数据库中进行查询，获得对应的机构信息。在信息的歧义消除方面，当存在多个邮箱时，可以利用作者名和用户名之间的对应关系，排除掉错误的结果。

[0085] (6) 将文章信息集合 $\{P_1, P_2, \ldots, P_n\}$ 中的元素与结果数据库中的记录进行匹配，消除冗余数据。

[0086] 虽然经过关联分析之后，抽取过程就已经完成，但是结果中可能存在重复的冗余信息。本步骤将抽取结果与结果数据库中的记录进行匹配。当找到匹配结果时，将两者进行模糊比对，如果结果数据库中的记录存在相关字段的缺失，则对该字段进行补全。如果在结果数据库中没有找到匹配结果，则将抽取结果添加到结果数据库中。

[0087] (7) 输出抽取结果。

[0088] 实例：

[0089] 以从学术主页 http://www.cs.uiuc.edu/～hanj/ 中抽取信息的过程为例，首先使用 Jiawei Han 作为搜索关键字在搜索引擎中进行检索，首先根据屏蔽数据库的，排除掉 Wikipedia 和 DBLP 的结果，然后选取排名前三的结果发出页面请求，经过分类器判定，选择第一个搜索结果即为该作者的学术主页。

[0090] 使用 HTML 解析器对页面进行解析，获取其中的子链接，根据链接关键字和上下文选定如下子页面进一步分析：

[0091] http://www.cs.uiuc.edu/homes/hanj/pubs/index.htm

[0092] https://agora.cs.illinois.edu/display/cs591han/Research+Publications+-+Data+Mining+Researc h+Group+at+CS%2C+UIUC

[0093] 对每一个待分析的页面进行文本单元的划分，以首页的页面为例，得到如下结果：

[0094]

“Jiawei Han”

“Professor, Department of Computer Science”

“Univ. of Illinois at Urbana-Champaign”

“E-mail: hanj[at]cs.uiuc.edu”

“Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han, Graph Cube: On Warehousing and OLAP Multidimensional Networks, Proc. of 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11), Athens, Greece, June 2011”

[0095] .....

[0096] 使用支持向量机对上述文本单元进行分类，分别判定为作者名字，无关数据，大学信息，邮箱，文章信息。根据判定的类别按照不同的提取流程进行进一步的提取，无关数据则直接放弃。

[0097] 作者名字的提取过程分别找到主页标题部分 (Jiawei Han)，正文中的作者名字 (Jiawei Han)，以及文章信息中包含的作者名字 (Jiawei Han，Xiaofei He，Deng Cai)，经过交叉匹配，确定 Jiawei Han 为最终的结果。

[0098] 邮箱信息的提取首先去掉前缀部分 (E-mail)：之后使用模糊匹配自动机枚举所有可能的邮箱匹配结果，如：

[0099] Hanj(用户名)at(@分隔符)cs(域名).(点)uiuc(域名).(点)edu(域名)

[0100] 按照匹配的符合程度对结果进行评分，选取最优结果作为邮箱的合法形式，之后转换为合法形式输出。

[0101] 机构信息的提取过程将被分类为机构信息的文本单元在机构索引中进行检索，在本例中以“Univ.of Illinois at Urbana-Champaign”为关键字进行检索，得到的检索结果中第一条记录即为“University of Illinois at Urbana-Champaign”，经过模糊匹配判定两者相符，因此可以直接输出结果。

[0102] 文章信息需要使用序列标注算法对文章信息进行标注，识别出其中的作者名，比如对于前面找到的文章信息，将其标注为如下形式：

[0103] <作者>Peixiang Zhao，Xiaolei Li，Dong Xin，and Jiawei Han，</作者><标题>Gfaph Cube：On Warehousing and OLAP Multidimensional Networks，</标题><会议>Proc.of 2011 ACM SIGMOD Int.Conf.on Management of Data(SIGMOD' 11)，</会议><地点>Athens，Greece，</地点><时间>June 2011</时间>

[0104] 将各个子字段分别识别出来即完成了文章信息的识别过程。之后根据信息之间的相关关联对存在缺失和歧义的结果进行补全和判定，将结果与结果数据库进行合并。

[0105] 本发明不仅局限于上述具体实施方式，本领域一般技术人员根据本发明公开的内容，可以采用其它多种具体实施方式实施本发明，因此，凡是采用本发明的设计结构和思路，做一些简单的变化或更改的设计，都落入本发明保护的范围。

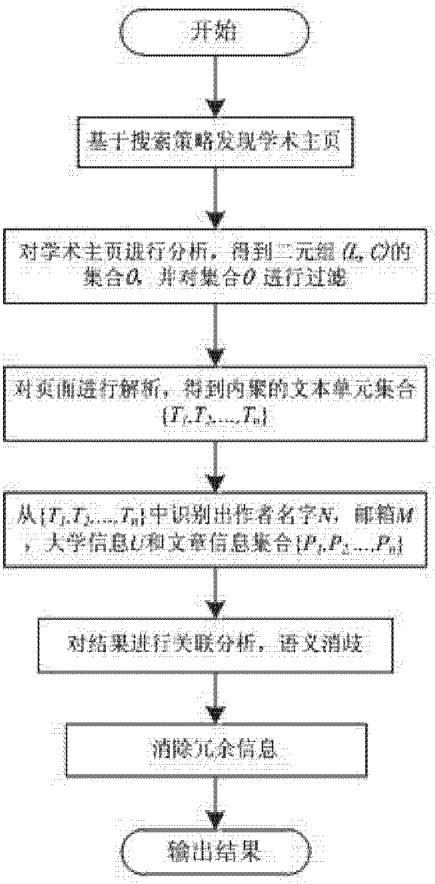

开始
基于搜索策略发现学术主页
对学术主页进行分析，得到二元组 (L, C)的集合O，并对集合O 进行过滤
对页面进行解析，得到内聚的文本单元集合 {T1,T2,...,Tn}
从{T1,T2,...,Tn}中识别出作者名字N，邮箱M，大学信息U和文章信息集合{P1,P2,...,Pn}
对结果进行关联分析，语义消歧
消除冗余信息
输出结果

图 1

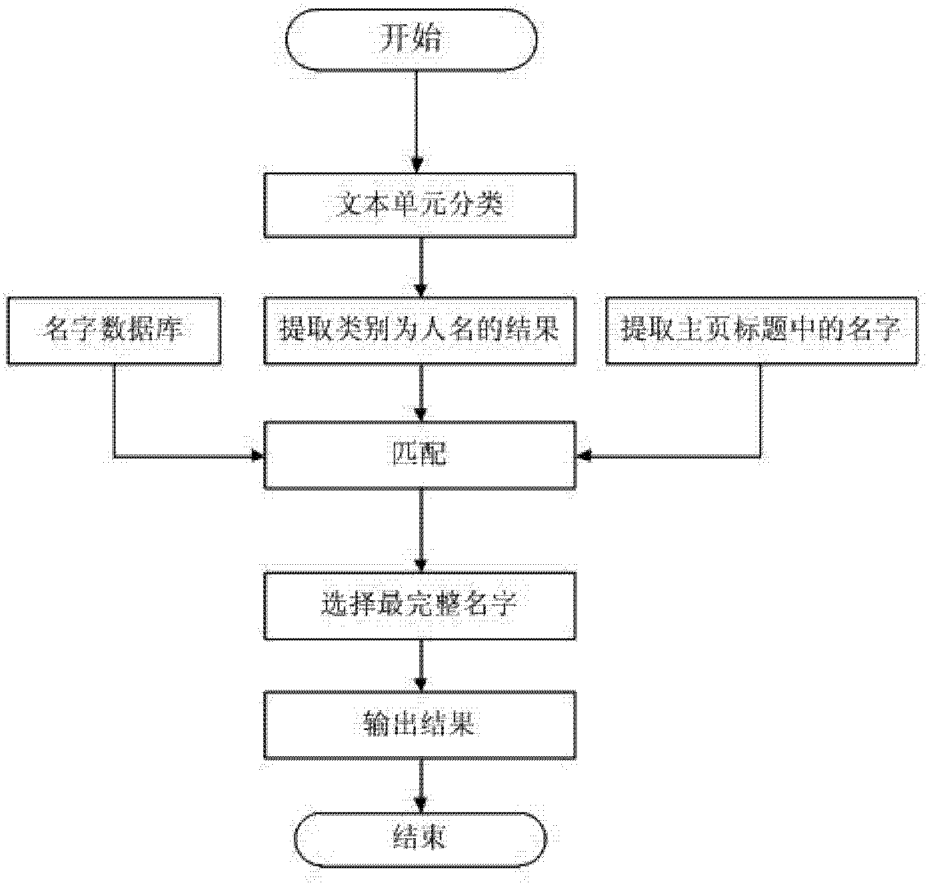


图 2

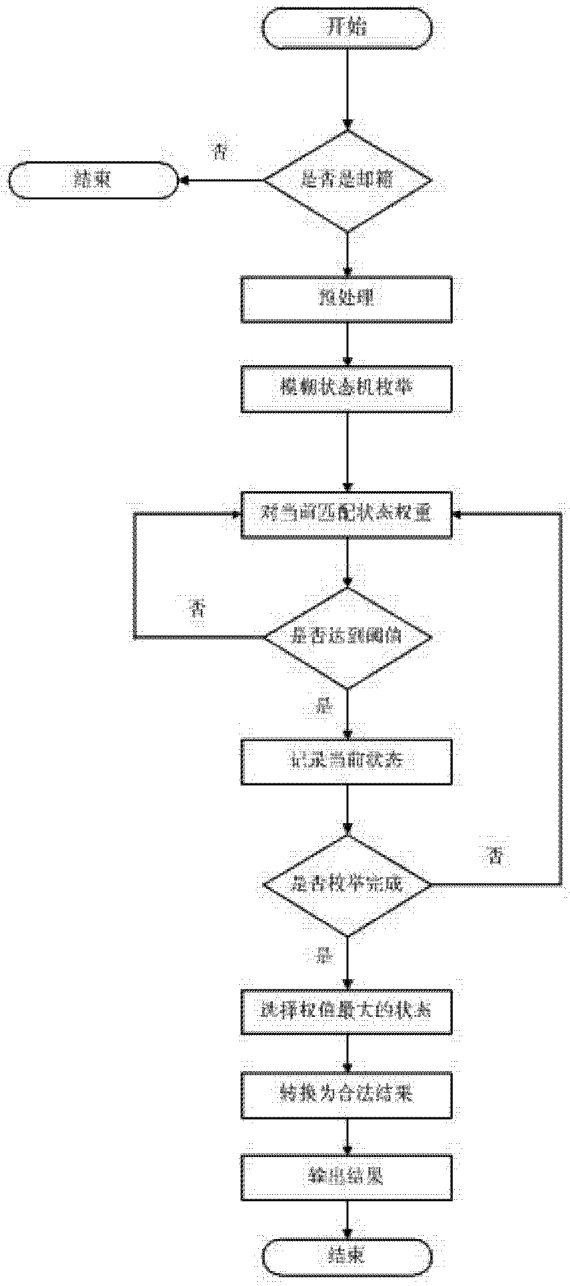

开始
是否是邮箱
否
结束
预处理
模糊状态机枚举
对当前匹配状态权重
是否达到阈值
否
是
记录当前状态
是否枚举完成
否
是
选择权值最大的状态
转换为合法结果
输出结果
结束

图 3

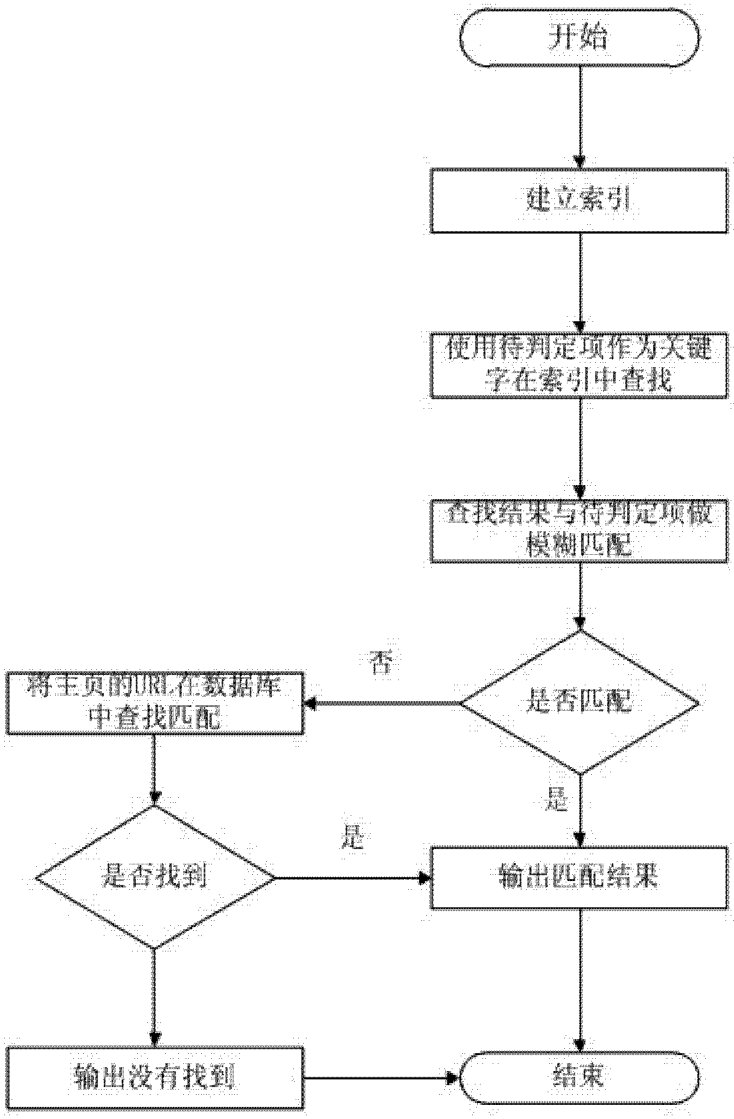

开始
建立索引
使用待判定项作为关键字在索引中查找
查找结果与待判定项做模糊匹配
是否匹配
否
将主页的URL在数据库中查找匹配
是
是否找到
是
输出匹配结果
输出没有找到
结束

图 4

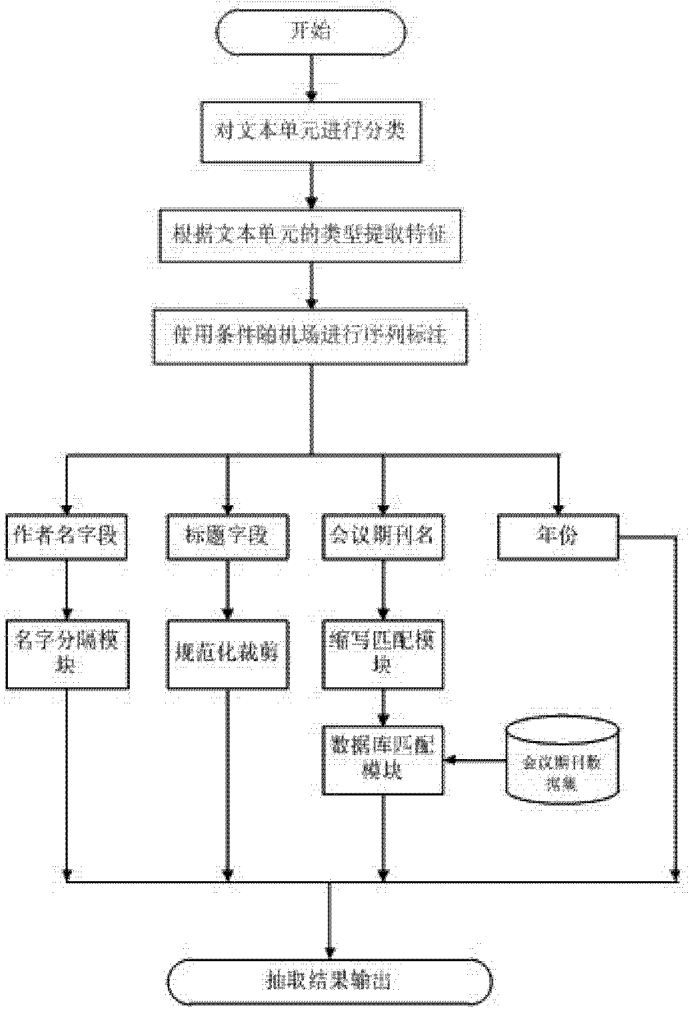

开始
对文本单元进行分类
根据文本单元的类型提取特征
使用条件随机场进行序列标注
作者名字段
标题字段
会议期刊名
年份
名字分隔模块
规范化裁剪
缩写匹配模块
数据库匹配模块
会议期刊数据集
抽取结果输出

图 5