(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 103823824 A
(43)申请公布日 2014.05.28

(21)申请号 201310314269.2

(22)申请日 2013.11.12

(71)申请人 哈尔滨工业大学深圳研究生院
地址 518000 广东省深圳市南山区西丽镇深圳大学城哈工大校区

(72)发明人 陈清财 张亮 王丹丹 王晓龙

(74)专利代理机构 深圳市科吉华烽知识产权事务所（普通合伙）44248
代理人 罗志强 黄震

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

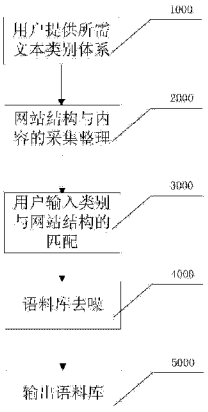
权利要求书2页 说明书8页 附图6页

(54)发明名称

一种借助互联网自动构建文本分类语料库的方法及系统

(57)摘要

本发明公开了一种借助互联网自动构建文本分类语料库的方法及系统，该方法包括如下步骤：用户提供所需的文本类别体系，网站结构与内容的采集整理，用户输入类别与网站结构的匹配，语料库去噪，输出语料库。本发明的技术效果是：利用互联网上各类网站上存在的类别标注信息，无需专业的人员手工标注，自适应不同用户的分类体系需求。它改变了传统的语料库构建系统需要大量的具有一定专业知识的人员加入标注的方式，采用对互联网上的丰富信息进行自动的抽取和挖掘的方式，可快速的构建出大容量的精准的文本分类语料库。



CN 103823824 A

1. 一种借助互联网自动构建文本分类语料库的方法，其特征在于，包括如下步骤：

用户提供所需的文本类别体系：即包含一个或多个层次的文本类别树，用户可以指定或不指定所涉及的领域；

网站结构与内容的采集整理：从互联网中采集大量网站，提取并分析网站的内容层次结构和每个主题词对应的网页内容信息；

用户输入类别与网站结构的匹配：将用户输入的文本类别与网站内容结构进行自动匹配，一个网站中与某个文本类别匹配上的网站内容单元所包含的网页作为该文本类别的候选语料；

语料库去噪：将从多个网站中匹配上的同一文本类别的候选语料合并成每个类别的候选语料库，并对候选语料库中每个类别下的文本进行去噪处理，提高语料库的质量；

输出语料库：输出去噪后的类别标准文本语料。

2. 根据权利要求 1 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在网站结构与内容的采集整理步骤中，对网站结构与内容的采集整理包括如下步骤：

初始种子链接获取：使用门户网站和领域内的网站作为初始种子链接，或者由用户指定；

网页数据爬取和存储：设置参数，参数包括并行通信数，每个站点的访问时间间隔，递归爬取深度；对上述初始种子链接进行递归抓取，同时记录页面间的跳转信息，将爬取得到的网页的源码文件存储在本地，用于后续对网页的离线分析和处理；

导航栏提取：识别提取体现网站内部组织结构的导航栏；

有效链接提取：统计各链接被赋予类别的次数，设置阈值，被赋予类别次数多于阈值的作为无效链接过滤掉，其余作为和页面主题相关的链接即有效链接；

网页内容提取：根据非标签字符和标签数的比值来提取网页的正文。

3. 根据权利要求 2 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在导航栏提取步骤中，包括如下步骤：

网页的页面分割：根据网页的 DOM 树将其分割成若干个块；

基于规则过滤的导航栏抽取：通过制定规则对网页的各个分块进行过滤和排序来提取导航栏；

基于图结构的导航栏抽取：将网页之间的链接关系组织成图结构，寻找极大连通子图，根据子图信息对页面的块结构进行过滤来提取导航栏。

4. 根据权利要求 3 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在网页的页面分割步骤中，包括如下步骤：

网页 DOM 树构造：利用 DOM 解析器将网页解析成 DOM 树；

网页 DOM 树化简：利用三个规则对网页 DOM 树进行化简。

5. 根据权利要求 3 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在基于规则过滤的导航栏抽取步骤中，包括如下步骤：

网页块过滤：根据链接类型、链接唯一性、样式表、锚文本在源码中的距离、每个锚文本包含的最大词数特征对网页块进行过滤；

网页剩余块排序：通过制定公式对网页中过滤后剩余的块进行打分并排序；

输出候选导航栏：根据网页剩余各块的排名和各个块的得分置信度来输出候选导航

栏。

6. 根据权利要求 3 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在基于图结构的导航栏抽取步骤中，包括如下步骤：

构造页面的链接关系图：将网页之间的链接指向关系表示成图；

获取极大完全子图：从页面的链接关系图中找出所有的极大完全子图，即：子图中任意两个节点直接相邻，并且此子图不被其他具有上述属性的子图包含；

输出候选导航栏：根据极大完全子图对页面的块结构进行过滤来得到候选导航栏。

7. 根据权利要求 2 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在网页内容提取步骤中，包括如下步骤：

网页源码行特征抽取：对网页源码中的每一行抽取二维特征，即：标签比和标签比导数；

网页源码行聚类获取正文：利用抽取出的二维特征和 k-Means 方法对网页源码中的所有行进行聚类，保证对非正文的去除效果。

8. 根据权利要求 7 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在网页源码行特征抽取步骤中，包括如下步骤：

计算初始行标签比，即：行包含的非 HTML 标签字符数和同一行标签数目的比值；

行标签比平滑：采用高斯过滤器对标签比的数据进行过滤归一，用这个过滤器与标签比进行卷积运算来平滑行标签比；

行标签比求导：计算行标签比的近似导数。

9. 根据权利要求 1 所述的借助互联网自动构建文本分类语料库的方法，其特征在于，在用户输入类别与网站结构的匹配步骤中，包括如下步骤：

相似度计算：向量化每个导航项和每个类别，计算它们之间的余弦相似度；

获取导航项所属类别：根据导航项和类别之间的相似度，结合导航项的链接 URL 决定导航项所属的类别。

10. 一种借助互联网自动构建文本分类语料库的系统，其特征在于，包括：

用户文本分类体系获取单元，用于提供所需的文本类别体系，包含一个或多个层次的文本类别树，用户可以指定或不指定所涉及的领域；

网站结构与内容的采集整理单元，用于从互联网中采集大量网站，提取并分析网站的内容层次结构和每个主题词对应的网页内容信息；

用户输入类别与网站结构的匹配单元，用于将用户输入的文本类别与网站内容结构进行自动匹配，一个网站中与某个文本类别匹配上的网站内容单元所包含的网页作为该文本类别的候选语料；

语料库去噪单元，用于将从多个网站中匹配上的同一文本类别的候选语料合并成每个类别的候选语料库，并对候选语料库中每个类别下的文本进行去噪处理，提高语料库的质量；

输出语料库单元，用于输出去噪后的类别标准文本语料。

一种借助互联网自动构建文本分类语料库的方法及系统

技术领域

[0001] 本发明涉及一种自动构建文本分类语料库的方法及系统。

背景技术

[0002] 随着互联网信息的高速增长，搜索引擎已成为人们浏览网络信息必不可少的工具。2012 年 7 月发布的《中国互联网络发展状况统计报告》显示：在网民日常使用中，搜索引擎虽然排名有所下滑，但依然超越了网络音乐和新闻，成为规模第二大的应用。

[0003] 目前，基于人工编撰目录并对其进行索引和维护的第一代搜索引擎技术基本退出历史舞台，取代它的是基于向量空间模型、概率语言模型等模型的第二代信息检索技术，在其中由于引入了 PageRank 和 LinkAnalysis 等技术，利用机群对大量互联网网页进行索引和检索，满足了用户对于检索系统的基本要求。不过，现有系统最大的问题是用户的需求是通过关键词来进行描述的，很多情况下很难找到准确描述检索目标的关键词，因此严重影响了返回结果的准确率。同时，由于需要索引的网页数目过于庞大，为了兼顾准确率与召回率，传统的通用搜索引擎往往返回属于不同主题的搜索结果，这种策略很难满足单个用户的检索需求。为了缓解这个问题，出现了专注于某一领域的垂直搜索引擎，如学术搜索、金融搜索、音乐搜索等。这类搜索引擎通过限定爬取和索引的网页范围来达到较高的检索精度，通过用户指定的分类来进行搜索，可以更好的满足用户的不同需求。

[0004] 然而，对于某个概念进行分类的方法往往是多样的，比如计算机学科，可以分为软件和硬件，也可以按照涉及的不同子学科分为体系结构，操作系统，计算机网络等等。现有的垂直搜索引擎，一般是根据领域专家事先定义好的分类方式，通过人工的方法标注出训练语料来训练分类器。这个过程费时费力，结果容易受到标注人个人倾向的影响，而且一旦分类方式发生变化，这一切又得从头再来。所以，这样的分类方式难以满足人们对不同领域的分类需求，更无法随用户需求的改变进行灵活调整。同时，网络上有一些网页是具有某些标注信息的或者是已经经过初步分类的，比如门户网站的导航栏一般都分为新闻、军事、博客等等很多子版块，关注于某个特定领域的网站一般都会按照相关领域的某种分类方式来构造。如何构造一种方法，使其能够自动利用这些已有的网页分类信息来自动构建分类语料库，是本发明要重点研究和探讨的问题。

发明内容

[0005] 为了解决现有技术中的问题，本发明提供了一种借助互联网自动构建文本分类语料库的方法。

[0006] 本发明提供了一种借助互联网自动构建文本分类语料库的方法，包括如下步骤：

[0007] 用户提供所需的文本类别体系：即包含一个或多个层次的文本类别树，用户可以指定或不指定所涉及的领域；

[0008] 网站结构与内容的采集整理：从互联网中采集大量网站，提取并分析网站的内容层次结构和每个主题词对应的网页内容信息；

[0009] 用户输入类别与网站结构的匹配：将用户输入的文本类别与网站内容结构进行自动匹配，一个网站中与某个文本类别匹配上的网站内容单元所包含的网页作为该文本类别的候选语料；

[0010] 语料库去噪：将从多个网站中匹配上的同一文本类别的候选语料合并成每个类别的候选语料库，并对候选语料库中每个类别下的文本进行去噪处理，提高语料库的质量；

[0011] 输出语料库：输出去噪后的类别标准文本语料。

[0012] 本发明的进一步技术方案是：在网站结构与内容的采集整理中，包括如下步骤：

[0013] 初始种子链接获取：使用门户网站和领域内的网站作为初始种子链接，或者由用户指定；

[0014] 网页数据爬取和存储：设置参数，参数包括并行通信数，每个站点的访问时间间隔，递归爬取深度，对上述初始种子链接进行递归抓取，同时记录页面间的跳转信息，将爬取得到的网页的源码文件存储在本地，用于后续对网页的离线分析和处理；

[0015] 导航栏提取：识别提取体现网站内部组织结构的导航栏；

[0016] 有效链接提取：统计各链接被赋予类别的次数，设置阈值，被赋予类别次数多于阈值的作为无效链接过滤掉，其余作为和页面主题相关的链接即有效链接；

[0017] 网页内容提取：根据非标签字符和标签数的比值来提取网页的正文。

[0018] 本发明的进一步技术方案是：在导航栏提取步骤中，包括如下步骤：

[0019] 网页的页面分割：根据网页的 DOM 树将其分割成若干个块；

[0020] 基于规则过滤的导航栏抽取：通过制定规则对网页的各个分块进行过滤和排序来提取导航栏；

[0021] 基于图结构的导航栏抽取：将网页之间的链接关系组织成图结构，寻找极大连通子图，根据子图信息对页面的块结构进行过滤来提取导航栏。

[0022] 本发明的进一步技术方案是：在网页的页面分割步骤中，包括如下步骤：

[0023] 网页 DOM 树构造：利用 DOM 解析器将网页解析成 DOM 树；

[0024] 网页 DOM 树化简：利用三个规则对网页 DOM 树进行化简。

[0025] 本发明的进一步技术方案是：在基于规则过滤的导航栏抽取步骤中，包括如下步骤：

[0026] 网页块过滤：根据链接类型、链接唯一性、样式表、锚文本在源码中的距离、每个锚文本包含的最大词数特征对网页块进行过滤；

[0027] 网页剩余块排序：通过制定公式对网页中过滤后剩余的块进行打分并排序；

[0028] 输出候选导航栏：根据网页剩余各块的排名和各个块的得分置信度来输出候选导航栏。

[0029] 本发明的进一步技术方案是：在基于图结构的导航栏抽取步骤中，包括如下步骤：

[0030] 构造页面的链接关系图：将网页之间的链接指向关系表示成图；

[0031] 获取极大完全子图：从页面的链接关系图中找出所有的极大完全子图，即：子图中任意两个节点直接相邻，并且此子图不被其他具有上述属性的子图包含；

[0032] 识别候选导航栏：根据极大完全子图对页面的块结构进行过滤来得到候选导航栏。

[0033] 本发明的进一步技术方案是：在网页内容提取步骤中，包括如下步骤：

[0034] 网页源码行特征抽取：对网页源码中的每一行抽取二维特征，即：标签比和标签比导数；

[0035] 网页源码行聚类获取正文：利用抽取出的二维特征和 k-Means 方法对网页源码中的所有行进行聚类，保证对非正文的去除效果。

[0036] 本发明的进一步技术方案是：在网页源码行特征抽取步骤中，包括如下步骤：

[0037] 计算行标签比，即：行包含的非 HTML 标签字符数和同一行标签数目的比值；

[0038] 行标签比平滑：采用高斯过滤器对标签比的数据进行过滤归一，用这个过滤器与标签比进行卷积运算来平滑行标签比；

[0039] 行标签比求导：计算行标签比的近似导数。

[0040] 本发明的进一步技术方案是：在用户输入类别与网站结构的匹配步骤中，包括如下步骤：

[0041] 相似度计算：向量化每个导航项和每个类别，计算它们之间的余弦相似度；

[0042] 获取导航项所属类别：根据导航项和类别之间的相似度，结合导航项的链接 URL 决定导航项所属的类别。

[0043] 本发明还提供了一种借助互联网自动构建文本分类语料库的系统，包括：

[0044] 用户文本分类体系获取单元，用于提供所需的文本类别体系，包含一个或多个层次的文本类别树，用户可以指定或不指定所涉及的领域；

[0045] 网站结构与内容的采集整理单元，用于从互联网中采集大量网站，提取并分析网站的内容层次结构和每个主题词对应的网页内容信息；

[0046] 用户输入类别与网站结构的匹配单元，用于将用户输入的文本类别与网站内容结构进行自动匹配，一个网站中与某个文本类别匹配上的网站内容单元所包含的网页作为该文本类别的候选语料；

[0047] 语料库去噪单元，用于将从多个网站中匹配上的同一文本类别的候选语料合并成每个类别的候选语料库，并对候选语料库中每个类别下的文本进行去噪处理，提高语料库的质量；

[0048] 输出语料库单元，用于输出去噪后的类别标准文本语料。

[0049] 本发明的技术效果是：本发明提出一种借助互联网自动构建文本分类语料库的方法及系统，利用互联网上各类网站上存在的类别标注信息，无需专业的人员手工标注，自适应不同用户的分类体系需求。它改变了传统的语料库构建系统需要大量的具有一定专业知识的人员加入标注的方式，采用对互联网上的丰富信息进行自动的抽取和挖掘的方式，可快速的构建出大容量的精准的文本分类语料库。

附图说明

[0050] 图 1 为本发明流程图。

[0051] 图 2 为本发明网站结构与内容的采集整理的流程图。

[0052] 图 3 为本发明导航栏提取的流程图。

[0053] 图 4 为本发明网页页面分割的流程图。

[0054] 图 5 为本发明经过简化后的百度首页源码实例图。

[0055] 图 6 为本发明使用 DOM 解析器解析后的 DOM 树的实例图。

[0056] 图 7 为本发明网页 DOM 树化简规则的示意图。

[0057] 图 8 为本发明基于规则过滤的导航栏提取的流程图。

[0058] 图 9 为本发明基于图结构的导航栏提取的流程图。

[0059] 图 10 为本发明极大完全子图的实例图。

[0060] 图 11 为本发明网页内容提取的流程图。

[0061] 图 12 为本发明网页源码行特征抽取的流程图。

[0062] 图 13 为本发明用户输入类别与网站结构的匹配的流程图。

[0063] 图 14 为本发明的系统原理图。

具体实施方式

[0064] 下面结合具体实施例，对本发明技术方案进一步说明。

[0065] 如图 1 所示，本发明的具体实施方式是：提供一种自动构建文本分类语料库的方法，包括如下步骤：

[0066] 步骤 1000：用户提供所需的文本类别体系，即：包含一个或多个层次的文本类别树，用户可以指定或不指定所涉及的领域。

[0067] 步骤 2000：网站结构与内容的采集整理，即：从互联网中采集大量网站，提取并分析网站的内容层次结构和每个主题词对应的网页内容信息。

[0068] 如图 2 所示，在网站结构与内容的采集整理步骤中，包括如下步骤：

[0069] 步骤 2100：初始种子链接获取：使用门户网站和领域内的网站作为初始种子链接，或者由用户指定；

[0070] 步骤 2200：网页数据爬取和存储：设置参数，参数包括并行通信数，每个站点的访问时间间隔，递归爬取深度，对初始种子链接进行递归抓取，同时记录页面间的跳转信息，将爬取得到的网页的源码文件存储在本地，用于后续对网页的离线分析和处理。

[0071] 步骤 2300：导航栏提取，即：识别提取体现网站内部组织结构的导航栏。如图 3 所示，具体在导航栏提取步骤中，包括如下步骤：

[0072] 步骤 2310：网页的页面分割，即：根据网页的 DOM 树将其分割成若干个块。如图 4 所示，具体在网页的页面分割步骤中，包括如下步骤：

[0073] 步骤 2311：网页 DOM 树构造，即：利用 DOM 解析器将网页解析成 DOM 树。DOM 是一种独立于使用平台和语言的接口标准，它由 W3C 组织提出，目的是为程序提供一种在运行过程中动态访问并改变其中的内容、结构或样式的方法。一个网页文件经过 DOM 解析器后形成的 DOM 呈树形结构，因此也有文献将其称为 DOM 树（DOM tree）。图 5 采用缩进形式显示了经过简化后的百度首页源码。

[0074] 使用 DOM 解析器解析后的 DOM 树将会具有如图 6 所示的树形结构。树中的每个内部节点都有具有指定名称（在网页源码的标签中定义）和属性。这些节点也可以通过一条从根节点到该节点的路径进行访问。

[0075] 将网页解析成 DOM 树以后，可以很方便的利用 XPath 来查找具有指定属性的节点，比如链接节点和文本节点。也可以直接通过遍历树中节点，对其进行操作（增、删、改等）。

[0076] 步骤 2312：网页 DOM 树化简，即：利用三个规则对网页 DOM 树进行化简。涉及如下

三个规则：

[0077] （1）删掉叶子节点中非链接节点的部分。

[0078] （2）如果某节点是其父节点唯一的孩子节点，则将其父节点删除，直接将该节点与其祖先节点连接起来。

[0079] （3）如果某节点有两个孩子节点，并且第一个孩子节点是链接节点，而另一个不是，则将该节点删除，并将两个孩子节点与其祖先节点连接起来。

[0080] 图 7 形象化的阐释了上面三条规则，其中左上角表示规则（1），右上角表示规则（2），下部表示规则（3）。

[0081] 网页的 DOM 树经过上述化简以后，将具有相同父节点的叶子节点合并为一个块，就完成了将整个页面进行分割的任务。

[0082] 步骤 2320：基于规则过滤的导航栏抽取，即：通过制定规则对网页的各个分块进行过滤和排序来提取导航栏。如图 8 所示，具体在基于规则过滤的导航栏提取步骤中，包括以下步骤：

[0083] 步骤 2321：网页块过滤：根据链接类型、链接唯一性、样式表、锚文本在源码中的距离、每个锚文本包含的最大词数、块内最少项目数等特征对网页块进行过滤；

[0084] 步骤 2322：网页剩余块排序：通过制定公式对网页中过滤后剩余的块进行打分并排序。

[0085] 二、块内锚文本所含词数的一致性：导航栏中的每个项目所含的单词数一般是一致的，外观越整齐的项目，越有可能属于同一导航栏 .

[0086] 三、块内剩余的锚文本所占比例：从前面的过滤过程可以发现，如果属于一个页面块的某些项被去掉了，那么这个块是导航栏的可能性应该会降低，而且该块中被过滤掉的项目越多，该块是导航栏的可能性就应该越低。

[0087] 步骤 2323：输出候选导航栏：根据网页剩余各块的排名和各个块的得分置信度来输出候选导航栏；

[0088] 步骤 2330：基于图结构的导航栏抽取，即：将网页之间的链接关系组织成图结构，寻找极大连通子图，根据子图信息对页面的块结构进行过滤来提取导航栏。

[0089] 步骤 2331：构造页面的链接关系图，即：将网页之间的链接指向关系表示成图。每个页面用一个节点表示，如果页面 A 中存在一条指向页面 B 的链接，则用一条由 A 指向 B 的有向边表示。将所有相关的页面都处理完以后可以生成整个网站的页面链接关系图。具有公共导航栏的页面将会是双向链接的方式呈现在图中，那么删掉图中的单向边，保留双向边，为了简化计算可以把具有双向边的有向图简化成无向图的方式进行处理。

[0090] 步骤 2332：获取极大完全子图，即：从页面的链接关系图中找出所有的极大完全子图，即：子图中任意两个节点直接相邻，并且此子图不被其他具有上述属性的子图包含。如图 10 所示，极大完全子图有 {1,2,3}，{2,3,4} 和 {4,5}。

[0091] 步骤 2333：识别候选导航栏，即：根据极大完全子图对页面的块结构进行过滤来得到候选导航栏。由于寻找最大完全子图的算法本质上是 NP 问题，当页面的链接关系图中双向链接的顶点超过一定数量，算法运行时间将变得不可接受，所以从复杂度的角度考虑，链接关系图中双向连接的顶点数少于 100 个时，该发明具体实施例中此部分采用但不限于如下的导航栏识别提取方法：

[0092] 输入：子图队列 MCQueuePage，首页的块集合 PageSec；

[0093] 输出：候选导航栏集合 CandNav；

[0094] Step1：将 MCQueuePage 中的所有元素标识为未处理。

[0095] Step2：从 MCQueuePage 中选取一个未处理的子图 SubGraph，如果全部处理完，则转步骤 4。

[0096] Step3：对 PageSec 中的所有元素进行过滤，将不在 SubGraph 的元素去掉，结果存入 CandNav，转步骤 2。

[0097] Step4：将 CandNav 中的各个块，按照包含的元素从多到少排序。

[0098] Step5：从头到尾查看 CandNav 中的各个块，如果当前块包含当前位置以后的某个块 Sec 的所有元素，则将 Sec 删除。

[0099] Step6：结束。

[0100] 双向连接的顶点数多于 100 个的时候，导航栏识别提取的方法为先投票再聚类的近似方法，该方法具体实施例中此部分过程如下但并不限于如下方法：

[0101] 输入：首页指向的所有页面集合 SetPages，每个页面有 URL 和入度两个属性。

[0102] 输出：形成导航栏的顶点集合 SetNav。

[0103] Step1：将 SetPages 中所有页面的入度置为 1，并将集合 SetPages 中的所有元素读入队列 QueuePages 中。

[0104] Step2：若队列为空，转到 Step4；否则转到 Step3。

[0105] Step3：从队列中取出一个页面，分析该页面包含的所有链接，如果当前页面包含的某个链接 URL_i 在 SetPages 中，则 SetPages 中 URL_i 对应的入度加 1。转到 Step2。

[0106] Step4：将 SetPages 中每个页面的入度进行聚类，方法为 k-Means，聚类中心数为 3。将所有属于中间簇的页面读入到 SetNav 中。

[0107] Step5：结束。

[0108] 步骤 2400：有效链接提取，即：提取和页面主题相关的链接。一般情况下导航栏中的每个锚文本都是指向另一个链接型页面(页面本身没有明显正文，而是某一主题下的链接的集合)。这样一个页面的类别实际上代表了内部包含的有效链接的类别。所谓有效链接是指跟页面本身主题相关的链接，像 Login、About Us、Sitemap 等链接就不属于有效链接，下文称之为无效链接。一般地，指向站外的链接不会是有效链接，可以把它们过滤掉。将剩余的链接与链入本页的导航栏锚文本赋予相同的类别。在网站内部，指向 Login、Sitemap 等无效链接的锚文本非常多，因此，无效链接被赋予的类别的次数也将明显多于有效链接。基于此，我们可以统计各链接被赋予类别的次数，设置一定的阈值，将被赋予类别次数多于阈值的作为无效链接过滤掉，其余的作为有效链接。

[0109] 步骤 2500：网页内容提取：根据非标签字符和标签数的比值来提取网页的正文。如图 11 所示，具体在网页内容提取步骤中，包括如下步骤：

[0110] 步骤 2510：网页源码行特征抽取：对网页源码中的每一行抽取二维特征，即：标签比和标签比导数。如图 12 所示，具体在网页源码行特征抽取步骤中，包括如下步骤：

[0111] 步骤 2511：计算行标签比，即：行包含的非 HTML 标签字符数和同一行标签数目的比值。

[0112] 步骤 2512：行标签比平滑，即：采用高斯过滤器对标签比的数据进行过滤归一。

[0113] 步骤 2513 :行标签比求导 :计算行标签比的近似导数。

[0114] 步骤 2520 :网页源码行聚类获取正文 :利用抽取出的二维特征和 k-Means 方法对网页源码中的所有行进行聚类,保证对非正文的去除效果。将一个聚类中心始终固定在原点(坐标 0,0)处,迭代的终止条件是达到最大迭代次数或者两次迭代间的簇中心变化幅度小于阈值。最后将属于中心的原点的簇的行去掉,将其他行中的正文信息保存下来,作为正文提取的结果。

[0115] 步骤 3000 :用户输入类别与网站结构的匹配,即 :计算每个导航项和各个类别之间的相似度,根据相似度以及导航项链接 URL 决定导航项所属的类别。如图 13 所示,具体在用户输入类别与网站结构的匹配步骤中,包括如下步骤 :

[0116] 步骤 3100 :相似度计算 :向量化每个导航项和每个类别,计算它们之间的相似度。

[0117] 步骤 3200 :获取导航项所属类别 :根据导航项和类别之间的相似度,结合导航项的链接 URL 决定导航项所属的类别。首先将导航项与各个类别的相似度进行计算,并对相似度进行排序。如果最大相似度唯一,则将对应类别作为锚文本类别,并将锚文本和分类信息存储起来 ;否则,需要将该锚文本对应的链接输入到一个 URL 类别判定模块。如果该模块输出非空,则将该锚文本及其对应的分类信息存储起来。

[0118] 步骤 4000 :语料库去噪 :将候选语料库中的各个类别下的文本进行聚类,去除类别内部的噪音,提高语料库的质量。理想情况下,候选语料库中的所有网页文本应该都是属于相关类别的。而由于网站本身权威程度和网站管理人员员水平的差异,一些原本不属于某个主题的页面也可能被划分到该类别下。另外,某些与指定主题不相关的信息也可能被引入,有些登录或者网站协议等界面由于锚文本的表述与过滤条件不同,从而可能被保留下来。这些噪声网页的存在令候选语料库的质量大大下降,因此需要对候选语料库进行去噪。

[0119] 由于噪声文本与所在主题并无太大关联,因此它们在文本的特征向量空间中会以离群点的形式出现。而聚类能够去除集合中的离群点,所以本文采用文本聚类算法对每个类别下剩余的网页文本进行聚类,保留聚类结果中的较大的簇,将较小的簇作为噪声去掉。在本发明的实施例中使用了 k-means 算法聚类,但具体实践中并不限于 k-means 聚类方法。

[0120] 步骤 5000 :输出语料库 :输出最终的各个类别下的标准文本语料。

[0121] 如图 14 所示,本发明还公开了一种借助互联网自动构建文本分类语料库的系统,包括 :

[0122] 用户文本分类体系获取单元 11,用于提供所需的文本类别体系,包含一个或多个层次的文本类别树,用户可以指定或不指定所涉及的领域 ;

[0123] 网站结构与内容的采集整理单元 21,用于从互联网中采集大量网站,提取并分析网站的内容层次结构和每个主题词对应的网页内容信息 ;

[0124] 用户输入类别与网站结构的匹配单元 22,用于将用户输入的文本类别与网站内容结构进行自动匹配,一个网站中与某个文本类别匹配上的网站内容单元所包含的网页作为该文本类别的候选语料 ;

[0125] 语料库去噪单元 23,用于将从多个网站中匹配上的同一文本类别的候选语料合并成每个类别的候选语料库,并对候选语料库中每个类别下的文本进行去噪处理,提高语料库的质量 ;

[0126] 输出语料库单元 12，用于输出去噪后的类别标准文本语料。

[0127] 本发明提出一种借助互联网自动构建文本分类语料库的方法及系统，利用互联网上各类网站上存在的类别标注信息，无需专业的人员手工标注，自适应不同用户的分类体系需求。它改变了传统的语料库构建系统需要大量的具有一定专业知识的人员加入标注的方式，采用对互联网上的丰富信息进行自动的抽取和挖掘的方式，可快速的构建出大容量的精准的文本分类语料库。

[0128] 本发明解决的技术问题是：传统的通过人工标注来对文档进行分类的方式需要不同程度的专业知识，耗费大量的人力物力，并且缺乏灵活性，不能很好的适应用户需求的变化。本发明提供了一种借助互联网中包含的形式各样的网站结构和内容信息、丰富的链接关系，自动寻找到高质量的网站信息源作为候选，从中获得相关网页，并利用导航栏识别、内容抽取、文本聚类等过程构建满足用户分类需求的语料库。用户只需提供确定的分类体系，自动构建文本分类语料库系统将会帮助用户搜集到大量精准的各个类别的语料。

[0129] 以上内容是结合具体的优选实施方式对本发明所作的进一步详细说明，不能认定本发明的具体实施只局限于这些说明。对于本发明所属技术领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干简单推演或替换，都应当视为属于本发明的保护范围。

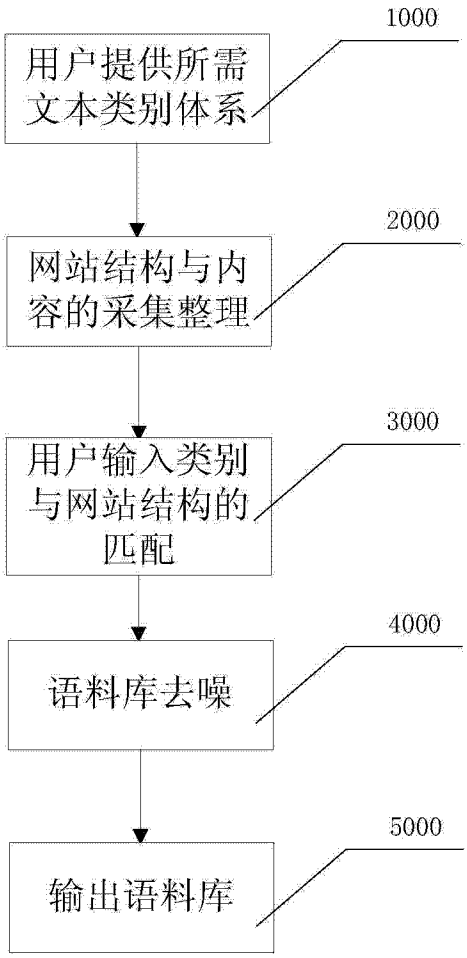

用户提供所需文本类别体系
1000
网站结构与内容的采集整理
2000
用户输入类别与网站结构的匹配
3000
语料库去噪
4000
输出语料库
5000

图 1

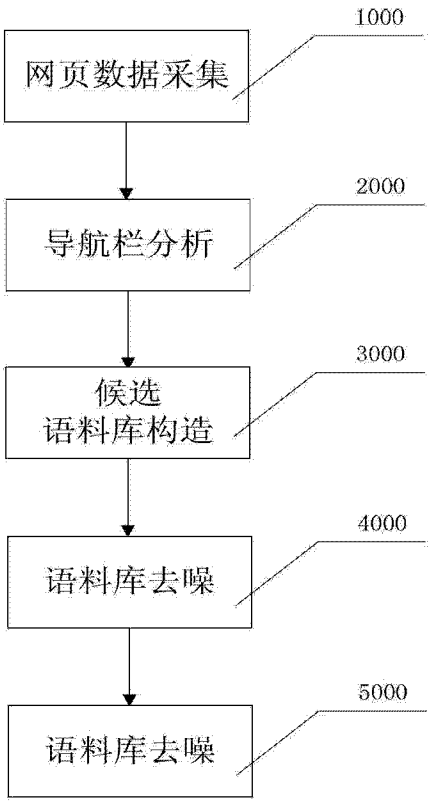

网页数据采集
1000
导航栏分析
2000
候选语料库构造
3000
语料库去噪
4000
语料库去噪
5000

图 2

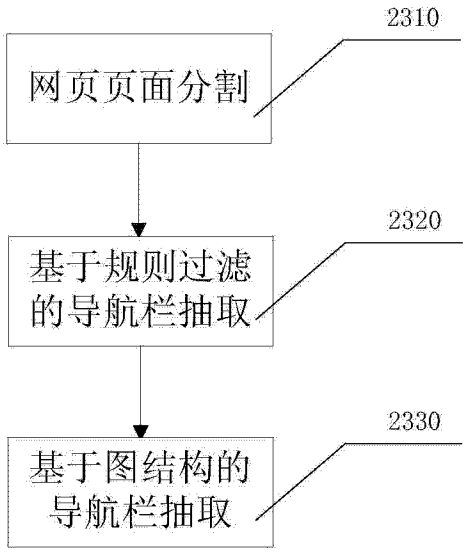

网页页面分割
2310
基于规则过滤的导航栏抽取
2320
基于图结构的导航栏抽取
2330

图 3

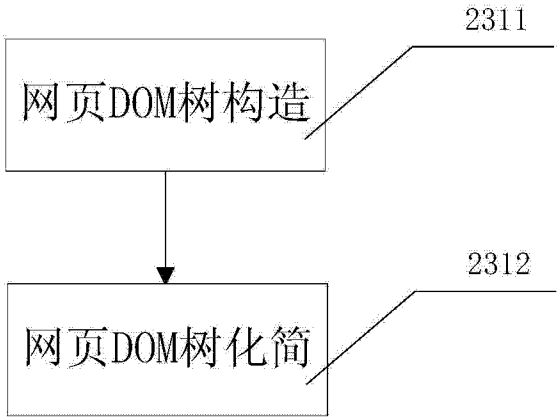

网页DOM树构造
2311
网页DOM树化简
2312

图 4

```
<html>
    <head>
        <meta http-equiv="Content-Type" content="text/html;charset=gb2312">
        <title>百度一下，你就知道</title>
    </head>
    <body>
        <p >
            <a id="seth">把百度设为主页</a>
            <a>把百度添加到桌面</a>
        </p>
        <p id="lh">
            <a href="http://top.baidu.com">搜索风云榜</a>
            |
            <a href="http://home.baidu.com">关于百度</a>
            |
            <a href="http://ir.baidu.com">About Baidu</a>
        </p>
    </body>
</html>
```

图 5

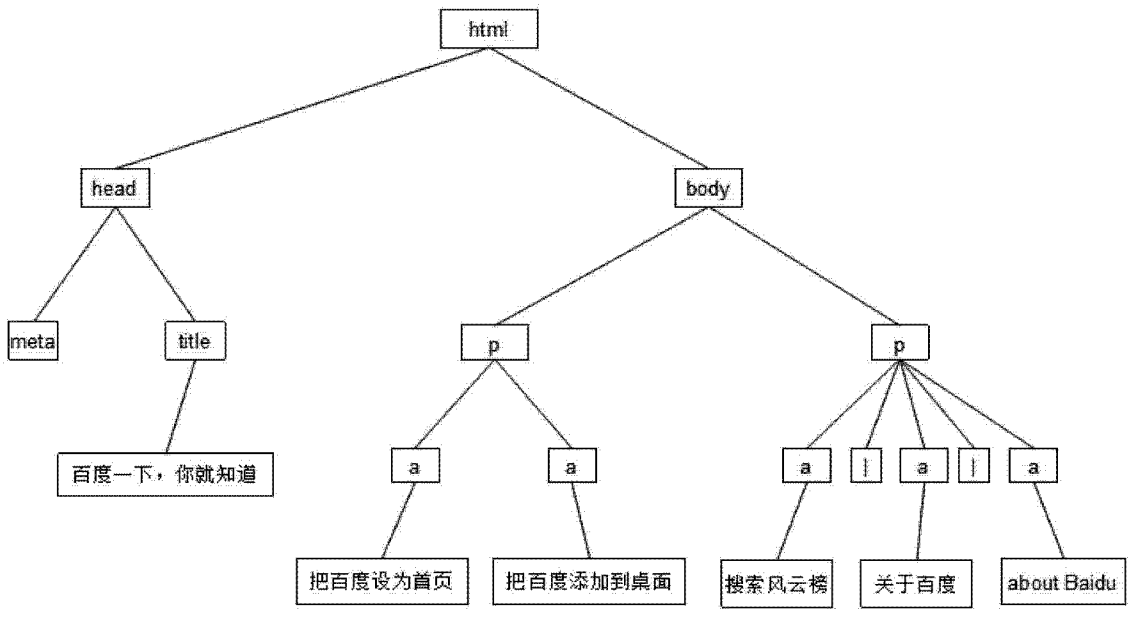


图 6

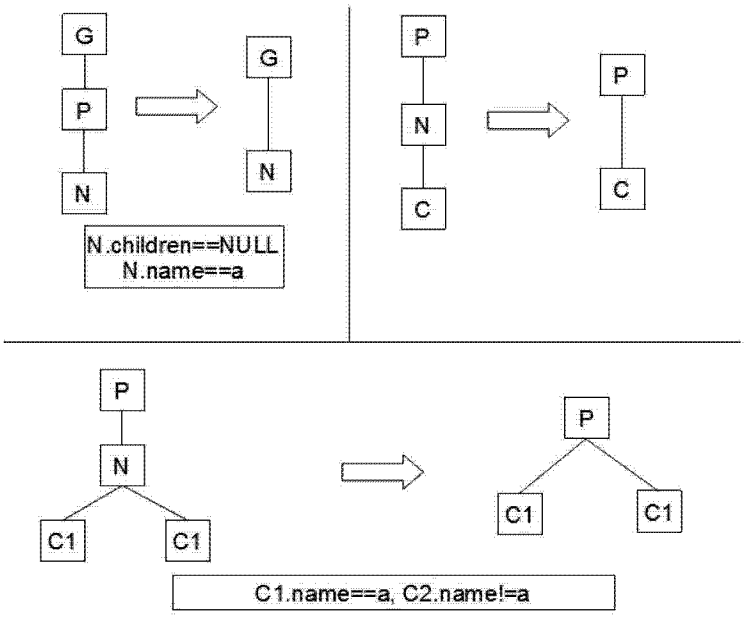

G
P
N
G
N
N.children==NULL
N.name==a
P
N
C
P
C
P
N
C1
C1
C1.name==a, C2.name!=a
P
C1
C1

图 7

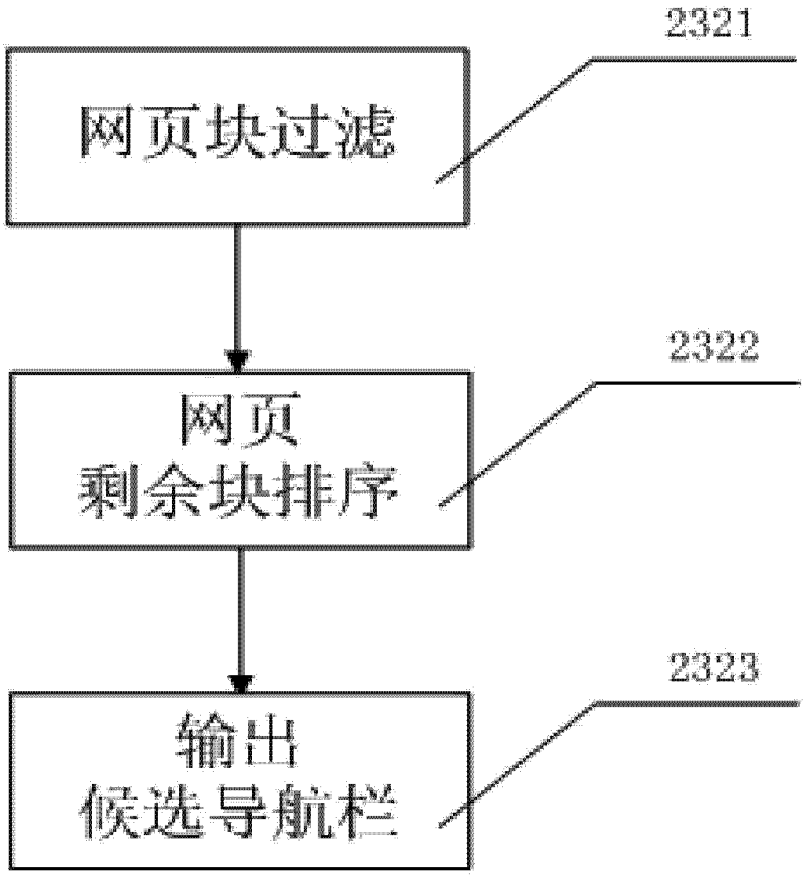

2321
网页块过滤
2322
网页
剩余块排序
2323
输出
候选导航栏

图 8

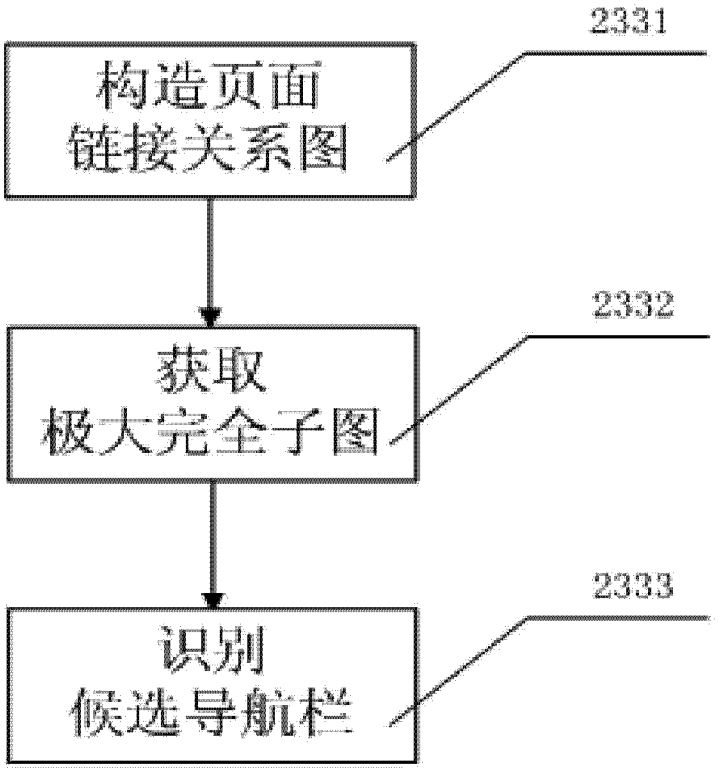

构造页面
链接关系图
2331
获取
极大完全子图
2332
识别
候选导航栏
2333

图 9

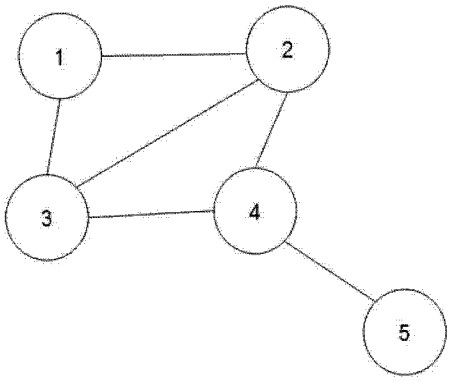

1
2
3
4
5

图 10

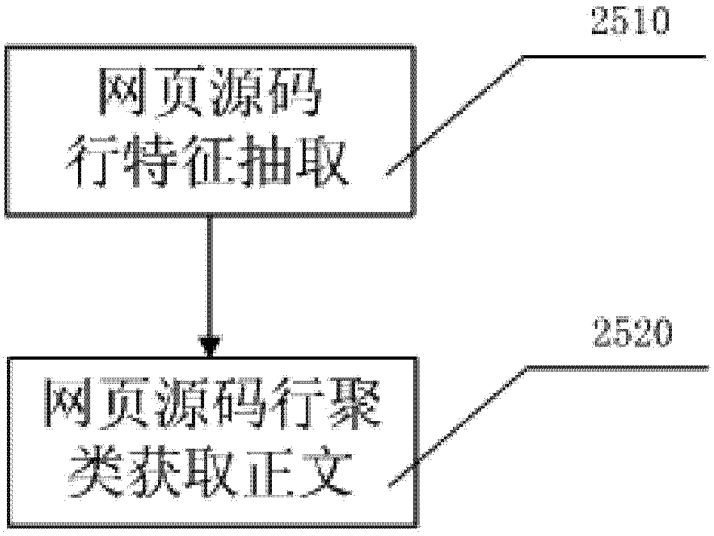

网页源码
行特征抽取
2510
网页源码行聚
类获取正文
2520

图 11

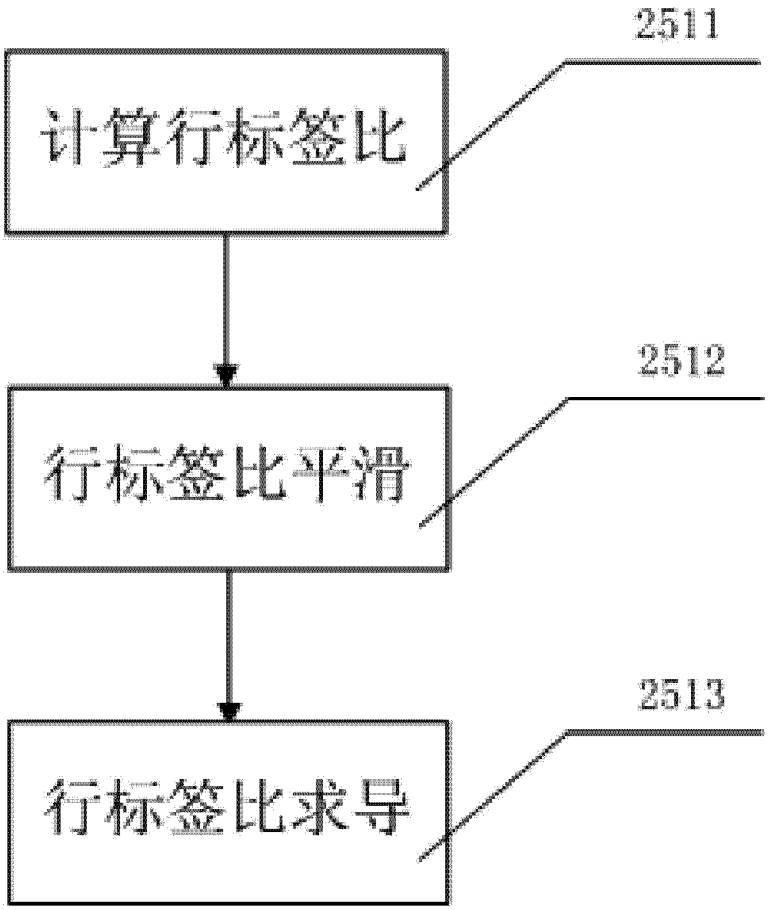

2511
计算行标签比
2512
行标签比平滑
2513
行标签比求导

图 12

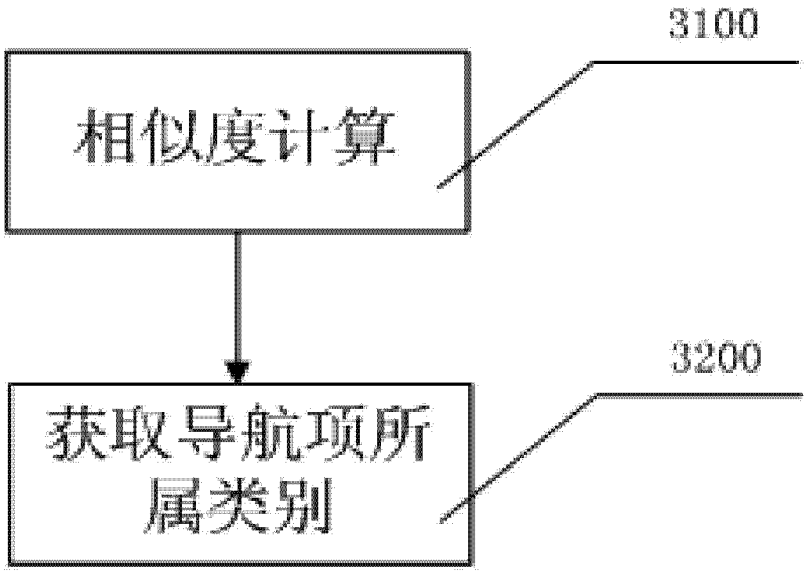

3100
相似度计算
3200
获取导航项所属类别

图 13

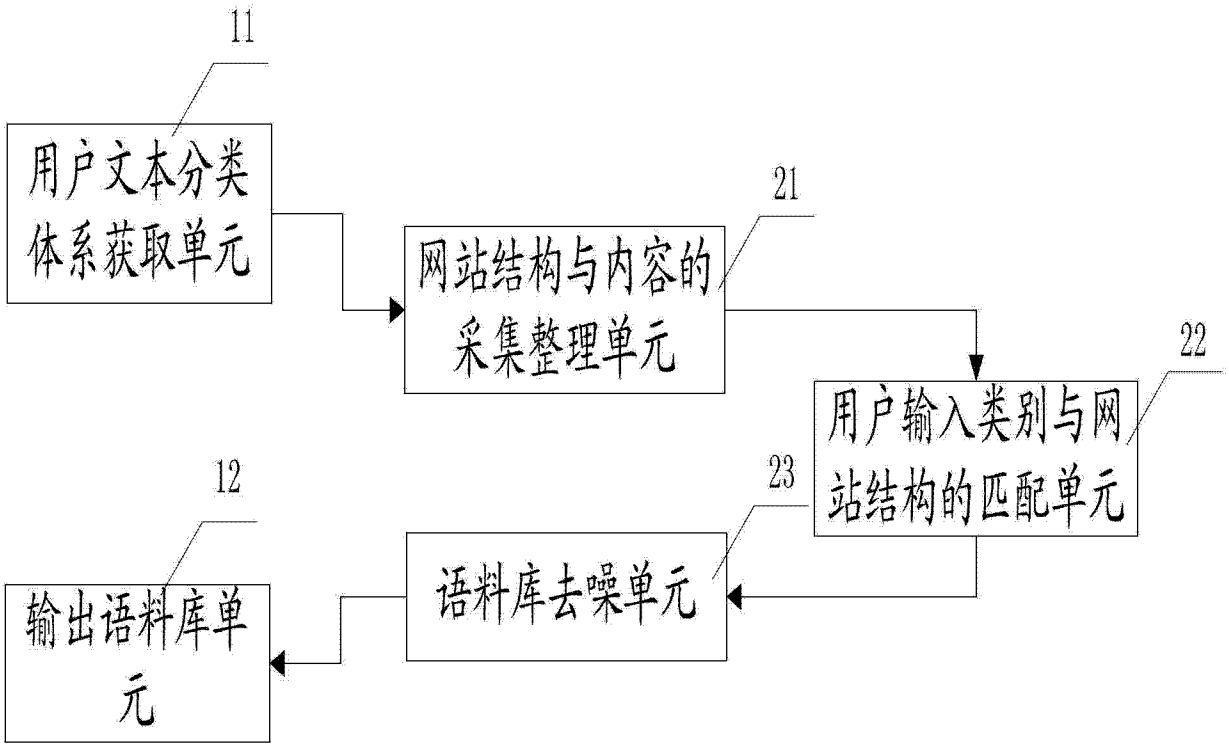

11
用户文本分类体系获取单元
21
网站结构与内容的采集整理单元
22
用户输入类别与网站结构的匹配单元
23
语料库去噪单元
12
输出语料库单元

图 14