(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 104615728 B
(45)授权公告日 2018.02.23

(21)申请号 201510066697.7

(22)申请日 2015.02.09

(65)同一申请的已公布的文献号
申请公布号 CN 104615728 A

(43)申请公布日 2015.05.13

(73)专利权人 浪潮集团有限公司
地址 250100 山东省济南市高新区浪潮路1036号

(72)发明人 李克学 范莹 戴鸿君 王传国 刘永

(74)专利代理机构 济南信达专利事务所有限公司 37100
代理人 李世喆

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

(56)对比文件
CN 101937438 A,2011.01.05,
US 2012/0284616 A1,2012.11.08,
CN 102779170 A,2012.11.14,
CN 103309924 A,2013.09.18,
US 2014/0359413 A1,2014.12.04,

审查员 李迪

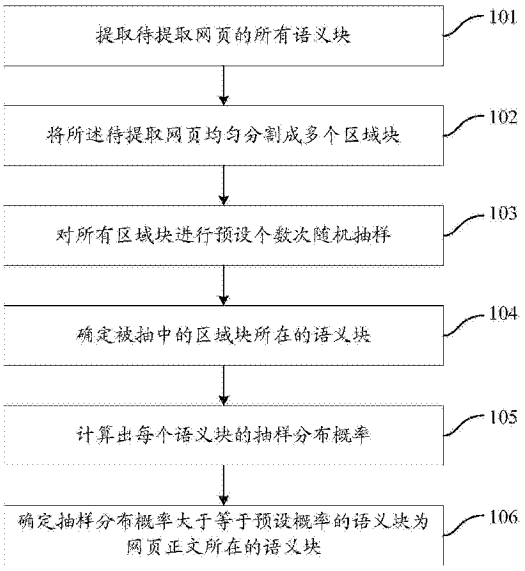
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种网页正文提取方法及装置

(57)摘要

本发明提供了一种网页正文提取方法及装置，该方法包括：提取待提取网页的所有语义块；将所述待提取网页均匀分割成多个区域块；对所有区域块进行预设个数次随机抽样；确定被抽中的区域块所在的语义块；计算出每个语义块的抽样分布概率；确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。本发明提供了一种网页正文提取方法及装置，能够提高提取网页正文的速度。



CN 104615728 B

1.一种网页正文提取方法,其特征在于,包括:

提取待提取网页的所有语义块;

将所述待提取网页均匀分割成多个区域块;

对所有区域块进行预设个数次随机抽样;

确定被抽中的区域块所在的语义块;

计算出每个语义块的抽样分布概率;

确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

2.根据权利要求1所述的方法,其特征在于,所述提取待提取网页的所有语义块,包括:

S1:对所述待提取网页的HTML源码建立文件对象模型DOM树;

S2:根据所述DOM树获得所述待提取网页的所有语义块。

3.根据权利要求2所述的方法,其特征在于,所述S2,包括:

S11:对所述DOM树进行页面分块,提取出所有的页面块;

S12:根据所述页面块,检测出页面块之间的所有分割条;

S13:根据所述分割条对所述页面块进行合并,得到合并后的页面块;

S14:获取合并后的页面块的内容相关度,判断当前页面块的内容相关度是否大于等于预设值,如果是,则确定当前页面块为语义块,否则,继续对这类语义块进行页面分块,返回步骤S11。

4.根据权利要求2所述的方法,其特征在于,在所述S1之前,还包括:

对所述待提取网页进行网页纠错,获得DOM树结构完整的待提取网页。

5.根据权利要求1-4任一所述的方法,其特征在于,还包括:记录每个语义块的坐标值和每个区域块的坐标值;

所述确定被抽中的区域块所在的语义块,包括:

根据所述每个语义块的坐标值和所述每个区域块的坐标值,确定被抽中的区域块所在的语义块。

6.一种网页正文提取装置,其特征在于,包括:

提取单元,用于提取待提取网页的所有语义块;

分割单元,用于将所述待提取网页均匀分割成多个区域块;

抽样单元,用于对所有区域块进行预设个数次随机抽样;

语义块确定单元,用于确定被抽中的区域块所在的语义块;

计算单元,用于计算出每个语义块的抽样分布概率;

正文确定单元,用于确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

7.根据权利要求6所述的装置,其特征在于,所述提取单元,包括:

建立子单元,用于对所述待提取网页的HTML源码建立文件对象模型DOM树;

提取子单元,用于根据所述建立子单元建立的所述DOM树获得所述待提取网页的所有语义块。

8.根据权利要求7所述的装置,其特征在于,所述提取子单元,包括:

页面块提取子单元,用于对所述建立子单元建立的所述DOM树进行页面分块,提取出所有的页面块,并对判断子单元建立的所述DOM数进行页面分块,提取出所有页面块;

检测子单元，用于根据所述页面块提取子单元提取出的所述页面块，检测出页面块之间的所有分割条；

合并子单元，用于根据检测子单元检测出的所述分割条对所述页面块进行合并，得到合并后的页面块；

判断子单元，用于获取所述合并子单元得到的合并后的页面块的内容相关度，判断当前页面块的内容相关度是否大于等于预设值，当判断结果为是时，确定当前页面块为语义块，当判断结果为否时，建立当前页面的DOM树，通知所述页面块提取子单元。

9.根据权利要求7所述的装置，其特征在于，还包括：

纠错子单元，用于对所述待提取网页进行网页纠错，获得DOM树结构完整的待提取网页。

10.根据权利要求6-9任一所述的装置，其特征在于，还包括：记录单元，用于记录每个语义块的坐标值和每个区域块的坐标值；

所述语义块确定单元，用于根据所述每个语义块的坐标值和所述每个区域块的坐标值，确定被抽中的区域块所在的语义块。

一种网页正文提取方法及装置

技术领域

[0001] 本发明涉及数据处理技术领域，特别涉及一种网页正文提取方法及装置。

背景技术

[0002] 随着网页信息资源快速的发展，每天都会产生很多网页。网页中可以包括正文信息和一些广告信息。如何从网页中提取出正文，变得十分重要。

[0003] 现有技术中，通过网页中标签之间的嵌套关系先从HTML(Hyper Text Mark-up Language，超文本标记语言文件)网页中解析出DOM(Document Object Model，文件对象模型)树，然后遍历所有DOM树，依据正文信息在DOM树中的分布规律确定正文的位置。

[0004] 通过上述描述可见，现有技术中提取网页正文的方法需要遍历所有DOM树，提取网页正文的速度较慢。

发明内容

[0005] 有鉴于此，本发明提供了一种网页正文提取方法及装置，能够提高提取网页正文的速度。

[0006] 本发明提供了一种网页正文提取方法，包括：

[0007] 提取待提取网页的所有语义块；

[0008] 将所述待提取网页均匀分割成多个区域块；

[0009] 对所有区域块进行预设个数次随机抽样；

[0010] 确定被抽中的区域块所在的语义块；

[0011] 计算出每个语义块的抽样分布概率；

[0012] 确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

[0013] 进一步地，所述提取待提取网页的所有语义块，包括：

[0014] S1：对所述待提取网页的HTML源码建立文件对象模型DOM树；

[0015] S2：根据所述DOM树获得所述待提取网页的所有语义块。

[0016] 进一步地，所述S2，包括：

[0017] S11：对所述DOM树进行页面分块，提取出所有的页面块；

[0018] S12：根据所述页面块，检测出页面块之间的所有分割条；

[0019] S13：根据所述分割条对所述页面块进行合并，得到合并后的页面块；

[0020] S14：获取合并后的页面块的内容相关度，判断当前页面块的内容相关度是否大于等于预设值，如果是，则确定当前页面块为语义块，否则，继续对这类语义块进行页面分块，返回步骤S11。

[0021] 进一步地，在所述S1之前，还包括：

[0022] 对所述待提取网页进行网页纠错，获得DOM树结构完整的待提取网页。

[0023] 进一步地，还包括：记录每个语义块的坐标值和每个区域块的坐标值；

[0024] 所述确定被抽中的区域块所在的语义块，包括：

[0025] 根据所述每个语义块的坐标值和所述每个区域块的坐标值，确定被抽中的区域块所在的语义块。

[0026] 另一方面，本发明提供了一种网页正文提取装置，包括：

[0027] 提取单元，用于提取待提取网页的所有语义块；

[0028] 分割单元，用于将所述待提取网页均匀分割成多个区域块；

[0029] 抽样单元，用于对所有区域块进行预设个数次随机抽样；

[0030] 语义块确定单元，用于确定被抽中的区域块所在的语义块；

[0031] 计算单元，用于计算出每个语义块的抽样分布概率；

[0032] 正文确定单元，用于确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

[0033] 进一步地，所述提取单元，包括：

[0034] 建立子单元，用于对所述待提取网页的HTML源码建立文件对象模型DOM树；

[0035] 提取子单元，用于根据所述建立子单元建立的所述DOM树获得所述待提取网页的所有语义块。

[0036] 进一步地，所述提取子单元，包括：

[0037] 页面块提取子单元，用于对所述建立子单元建立的所述DOM树进行页面分块，提取出所有的页面块，并对判断子单元建立的所述DOM数进行页面分块，提取出所有页面块；

[0038] 检测子单元，用于根据所述页面块提取子单元提取出的所述页面块，检测出页面块之间的所有分割条；

[0039] 合并子单元，用于根据检测子单元检测出的所述分割条对所述页面块进行合并，得到合并后的页面块；

[0040] 判断子单元，用于获取所述合并子单元得到的合并后的页面块的内容相关度，判断当前页面块的内容相关度是否大于等于预设值，当判断结果为是时，确定当前页面块为语义块，当判断结果为否时，建立当前页面块的DOM树，通知所述页面块提取子单元。

[0041] 进一步地，还包括：

[0042] 纠错子单元，用于对所述待提取网页进行网页纠错，获得DOM树结构完整的待提取网页。

[0043] 进一步地，还包括：记录单元，用于记录每个语义块的坐标值和每个区域块的坐标值；

[0044] 所述语义块确定单元，用于根据所述每个语义块的坐标值和所述每个区域块的坐标值，确定被抽中的区域块所在的语义块。

[0045] 本发明提供了一种网页正文提取方法及装置，提取待提取网页的所有语义块，将所述待提取网页均匀分割成多个区域块，对区域块进行随机抽样，确定被抽中的区域块所在的语义块，计算出每个语义块的抽样分布概率，通过每个语义块的抽样分布概率来表征每个语义块的面积大小，当语义块的抽样分布概率大时，说明该语义块的面积较大，该语义块为网页正文所在的语义块的概率也大，当语义块的抽样分布概率大于等于预设概率时，则确定该语义块为网页正文所在的语义块，这种方法只需进行简单的随机抽样计算即可，无需对待提取网页的DOM树进行遍历，能够提高提取网页正文的速度。

附图说明

[0046] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0047] 图1是本发明一实施例提供的一种网页正文提取方法的流程图；

[0048] 图2是本发明一实施例提供的一种网页正文提取装置的示意图。

具体实施方式

[0049] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例，基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0050] 如图1所示，本发明实施例提供了一种网页正文提取方法，该方法可以包括以下步骤：

[0051] 步骤101：提取待提取网页的所有语义块；

[0052] 步骤102：将所述待提取网页均匀分割成多个区域块；

[0053] 步骤103：对所有区域块进行预设个数次随机抽样；

[0054] 步骤104：确定被抽中的区域块所在的语义块；

[0055] 步骤105：计算出每个语义块的抽样分布概率；

[0056] 步骤106：确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

[0057] 通过本发明实施例提供的一种网页正文提取方法，提取待提取网页的所有语义块，将所述待提取网页均匀分割成多个区域块，对区域块进行随机抽样，确定被抽中的区域块所在的语义块，计算出每个语义块的抽样分布概率，通过每个语义块的抽样分布概率来表征每个语义块的面积大小，当语义块的抽样分布概率大时，说明该语义块的面积较大，该语义块为网页正文所在的语义块的概率也大，当语义块的抽样分布概率大于等于预设概率时，则确定该语义块为网页正文所在的语义块，这种方法只需进行简单的随机抽样计算即可，无需对待提取网页的DOM树进行遍历，能够提高提取网页正文的速度。

[0058] 为了能够更加准确的提取出待提取网页的所有语义块，所述提取待提取网页的所有语义块，包括：

[0059] S1：对所述待提取网页的HTML源码建立DOM树；

[0060] S2：根据所述DOM树获得所述待提取网页的所有语义块。

[0061] 在一种可能的实现方式中，所述S2，包括：

[0062] S11：对所述DOM树进行页面分块，提取出所有的页面块；

[0063] S12：根据所述页面块，检测出页面块之间的所有分割条；

[0064] S13：根据所述分割条对所述页面块进行合并，得到合并后的页面块；

[0065] S14：获取合并后的页面块的内容相关度，判断当前页面块的内容相关度是否大于

等于预设值，如果是，则确定当前页面块为语义块，否则，继续对这类语义块进行页面分块，返回步骤S11。

[0066] 在该实现方式中，为了使得每个语义块中内容的相关性较高，需要保证每个输出的语义块的内容相关度较高。在输出语义块之前对每个合并后的页面块进行内容相关度的判断，当页面块的内容相关度大于等于预设值时，该页面块满足要求，确定该页面块为语义块；如果有合并后的页面块的内容相关度小于预设值，则继续对这类语义块进行页面分块，直到所有的页面块的内容相关度都大于等于预设值。另外，分割条包括横向和纵向的分割条。

[0067] 由于HTML在编写过程中存在不规范现象，为了能够获得准确的DOM树，需要对待提取网页进行网页纠错，使待提取网页规范。在所述S1之前，还包括：

[0068] 对所述待提取网页进行网页纠错，获得DOM树结构完整的待提取网页。

[0069] 举例来说，所述对所述待提取网页进行网页纠错，包括：对所述待提取网页进行HTML标签补全、错误标签去除、脚本、代码注释去除。

[0070] 为了能够准确的确定被抽中的区域块所在的语义块。该方法还包括：记录每个语义块的坐标值和每个区域块的坐标值；

[0071] 所述确定被抽中的区域块所在的语义块，包括：

[0072] 根据所述每个语义块的坐标值和所述每个区域块的坐标值，确定被抽中的区域块所在的语义块。

[0073] 举例来说，待提取网页对应的多个区域块和语义块均为矩形，每个区域块和每个语义块均可以通过一条对角线上的两个顶点来确定。

[0074] 另外，在步骤102中，可以根据预设的精确度将所述待提取网页均匀分割成多个区域块，从网页的横向和纵向对所述待提取网页进行分割。其中，为了提高每个语义块的抽样分布概率的准确度，区域块的大小越小越好。

[0075] 本发明实施例提供了一种网页正文提取方法，该方法可以包括图中未示出的以下步骤：

[0076] 步骤A1：提取待提取网页的所有4个语义块，分别是第一语义块、第二语义块、第三语义块、第四语义块；

[0077] 步骤A2：将所述待提取网页均匀分割成多个区域块；

[0078] 步骤A3：对所有区域块进行100次随机抽样；

[0079] 步骤A4：确定被抽中的区域块所在的语义块；

[0080] 步骤A5：计算出每个语义块的抽样分布概率；

[0081] 举例来说，100次抽样中有70个区域块位于第一语义块中，则计算出第一语义块的抽样分布概率为0.7。

[0082] 步骤A6：确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

[0083] 举例来说，预设概率为0.6，其中，第一语义块的抽样分布概率为0.7，大于预设概率0.6，则确定第一语义块为网页正文所在的语义块。

[0084] 本发明实施例还提供了一种网页正文提取装置，参见图2，该装置包括：

[0085] 提取单元201，用于提取待提取网页的所有语义块；

[0086] 分割单元202，用于将所述待提取网页均匀分割成多个区域块；

[0087] 抽样单元203，用于对所有区域块进行预设个数次随机抽样；

[0088] 语义块确定单元204，用于确定被抽中的区域块所在的语义块；

[0089] 计算单元205，用于计算出每个语义块的抽样分布概率；

[0090] 正文确定单元206，用于确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块。

[0091] 为了能够更加准确的提取出待提取网页的所有语义块，所述提取单元201，包括：

[0092] 建立子单元，用于对所述待提取网页的HTML源码建立文件对象模型DOM树；

[0093] 提取子单元，用于根据所述建立子单元建立的所述DOM树获得所述待提取网页的所有语义块。

[0094] 在一种可能的实现方式中，所述提取子单元，包括：

[0095] 页面块提取子单元，用于对所述建立子单元建立的所述DOM树进行页面分块，提取出所有的页面块，并对判断子单元建立的所述DOM数进行页面分块，提取出所有页面块；

[0096] 检测子单元，用于根据所述页面块提取子单元提取出的所述页面块，检测出页面块之间的所有分割条；

[0097] 合并子单元，用于根据检测子单元检测出的所述分割条对所述页面块进行合并，得到合并后的页面块；

[0098] 判断子单元，用于获取所述合并子单元得到的合并后的页面块的内容相关度，判断当前页面块的内容相关度是否大于等于预设值，当判断结果为是时，确定当前页面块为语义块，当判断结果为否时，建立当前页面块的DOM树，通知所述页面块提取子单元。

[0099] 由于HTML在编写过程中存在不规范现象，为了能够获得准确的DOM树，需要对待提取网页进行网页纠错，使待提取网页规范。该装置还包括：

[0100] 纠错子单元，用于对所述待提取网页进行网页纠错，获得DOM树结构完整的待提取网页。

[0101] 为了能够准确的确定被抽中的区域块所在的语义块。该装置还包括：记录单元，用于记录每个语义块的坐标值和每个区域块的坐标值；

[0102] 所述语义块确定单元，用于根据所述每个语义块的坐标值和所述每个区域块的坐标值，确定被抽中的区域块所在的语义块。

[0103] 上述装置内的各单元之间的信息交互、执行过程等内容，由于与本发明方法实施例基于同一构思，具体内容可参见本发明方法实施例中的叙述，此处不再赘述。

[0104] 本发明实施例提供的一种网页正文提取方法及装置，具有如下有益效果：

[0105] 1、通过本发明实施例提供的一种网页正文提取方法及装置，提取待提取网页的所有语义块，将所述待提取网页均匀分割成多个区域块，对区域块进行随机抽样，确定被抽中的区域块所在的语义块，计算出每个语义块的抽样分布概率，通过每个语义块的抽样分布概率来表征每个语义块的面积大小，当语义块的抽样分布概率大时，说明该语义块的面积较大，该语义块为网页正文所在的语义块的概率也大，当语义块的抽样分布概率大于等于预设概率时，则确定该语义块为网页正文所在的语义块，这种方法只需进行简单的随机抽样计算即可，无需对待提取网页的DOM树进行遍历，能够提高提取网页正文的速度。

[0106] 2、通过本发明实施例提供的一种网页正文提取方法及装置，能够精确抽取网页正

文信息，这种方法只需进行简单的随机抽样计算即可，无需对待提取网页的DOM树进行遍历，降低了提取网页正文的复杂度。

[0107] 需要说明的是，在本文中，诸如第一和第二之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语"包括"、"包含"或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句"包括一个……"限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同因素。

[0108] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储在计算机可读取的存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质中。

[0109] 最后需要说明的是：以上所述仅为本发明的较佳实施例，仅用于说明本发明的技术方案，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内所做的任何修改、等同替换、改进等，均包含在本发明的保护范围内。

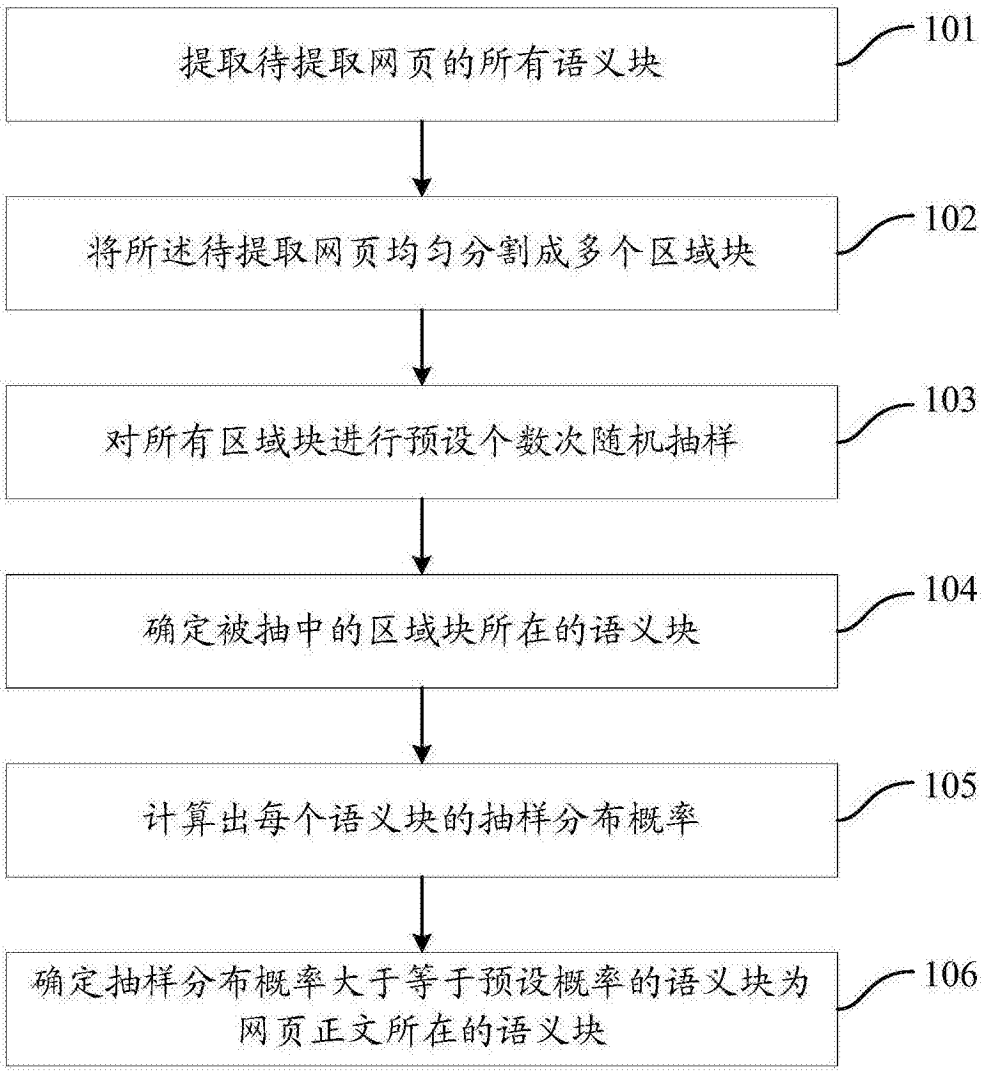

提取待提取网页的所有语义块
101
将所述待提取网页均匀分割成多个区域块
102
对所有区域块进行预设个数次随机抽样
103
确定被抽中的区域块所在的语义块
104
计算出每个语义块的抽样分布概率
105
确定抽样分布概率大于等于预设概率的语义块为网页正文所在的语义块
106

图1

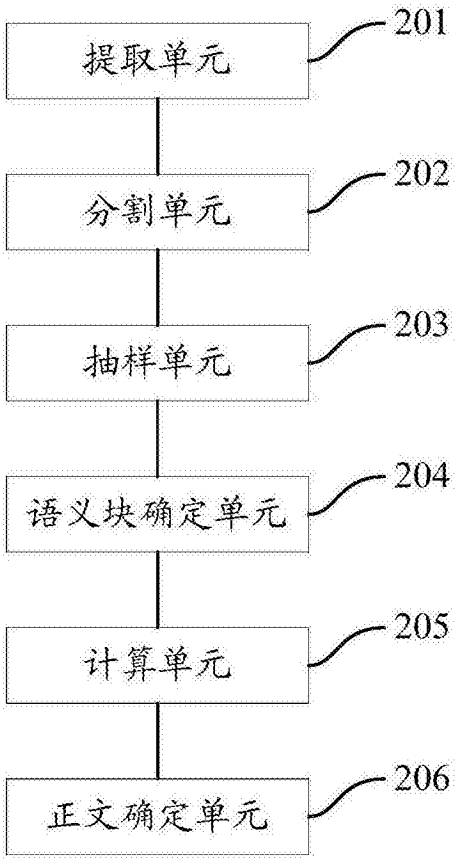


图2