(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 107403002 A
(43)申请公布日 2017.11.28

(21)申请号 201710601539.6

(22)申请日 2017.07.21

(71)申请人 山东师范大学
地址 250014 山东省济南市文化东路88号

(72)发明人 王红 刘锐

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 张勇

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

权利要求书1页 说明书5页 附图1页

(54)发明名称

一种基于词汇关键度的网络论坛正文提取方法、装置

(57)摘要

本发明公开了一种面向论坛主题帖的正文筛选方法，该方法涉及数据挖掘领域，是为解决从论坛主题帖中提取正文而提出的。本算法的实现方法是从网页总体中抽取部分样本，利用去除显著的非正文部分，对剩余内容进行分词，用TF-IDF方法评价所有词汇的关键度，定位关键度最高的若干词汇所在位置，记录出现最频繁的位置，利用该位置信息对数据集中剩余的主题帖页面进行正文提取。经实验验证，本方法具有较高的准确度和执行效率。

CN 107403002 A

1.一种基于词汇关键度的网络论坛正文提取方法,其特征在于,包括:抽取数据集中部分主题帖页面样本,去除非正文部分,对剩余内容进行分词,计算所有词汇的关键度,定位关键度最高的部分词汇所在位置,引导提取数据集中剩余主题帖页面的正文,若正文内容正确则输出正文,若不正确,则从抽取数据集中部分主题帖页面样本开始重新处理。

2.根据权利要求1所述的方法,其特征在于,所述去除非正文部分包括:

去除主题帖页面中显著的非正文内容;根据停用词库排除主题帖页面中无关内容;根据相似规则去除主题帖页面中不应被包含在正文中的内容。

3.根据权利要求2所述的方法,其特征在于,去除主题帖页面中显著的非正文内容包括:去除主题帖页面源码中的标签及其内容,所述标签至少包括:<head>、<script>和<a>。

4.根据权利要求2所述的方法,其特征在于,根据停用词库排除主题帖页面中无关内容包括:根据停用词库,将出现停用词的整行文本去除;或者根据停用词库与待测文本比对以决定是否保留该段待测文本。

5.根据权利要求2所述的方法,其特征在于,根据相似规则去除主题帖页面中不应被包含在正文中的内容包括:

比对两段待测文本的若干起始字符,判断是否保留这两段文本;或者根据相似规则去除由程序生成的不应包含在正文中的内容。

6.根据权利要求1所述的方法,其特征在于,采用TF-IDF方法计算所有自会的关键度。

7.根据权利要求1所述的方法,其特征在于,所述定位关键度最高的部分词汇所在位置,引导提取数据集中剩余主题帖页面的正文包括:

在主题帖页面内对关键度最高的词汇打分,选出正文出现概率最高的位置;

在不同页面中,对所述正文出现概率最高的位置再次打分,以确定正文位置;

根据所确定的正文位置,提取数据集中剩余主题帖页面的正文。

8.根据权利要求7所述的方法,其特征在于,根据所确定的正文位置,提取数据集中剩余主题帖页面的正文包括:

先根据所确定的正文位置构造解析规则,再根据所述解析规则对数据集中剩余主题帖页面进行正文提取。

9.一种计算机可读存储介质,其中存储有多条指令,其特征在于:所述指令适于由处理器加载并执行以下处理:

抽取数据集中部分主题帖页面样本,去除显著的非正文部分,对剩余内容进行分词,计算所有词汇的关键度,定位关键度最高的部分词汇所在位置,引导提取数据集中剩余主题帖页面的正文,若正文内容正确则输出正文,若不正确,则从抽取数据集中部分主题帖页面样本开始重新处理。

10.一种基于词汇关键度的网络论坛正文提取装置,其特征在于:包括处理器和计算机可读存储介质,处理器用于实现各指令;计算机可读存储介质用于存储多条指令,所述指令适于由处理器加载并执行以下处理:

抽取数据集中部分主题帖页面样本,去除显著的非正文部分,对剩余内容进行分词,计算所有词汇的关键度,定位关键度最高的部分词汇所在位置,引导提取数据集中剩余主题帖页面的正文,若正文内容正确则输出正文,若不正确,则从抽取数据集中部分主题帖页面样本开始重新处理。

一种基于词汇关键度的网络论坛正文提取方法、装置

技术领域

[0001] 本发明设计网络数据挖掘领域，具体为根据论坛主题帖内词汇的关键度，提取主题帖正文的方法、装置。

背景技术

[0002] 正文是一个论坛主题帖最重要的部分。因而提取出主题帖正文是对页面进行后续处理前最重要的准备工作。目前，对网页正文提取的方法主要有根据网页结构和文本特征进行正文提取的方法；使用布局相似性作为一句进行正文提取的方法；使用DOM节点的文本密度作为标准的正文提取方法等等。但是，在实际中，由于论坛正文的特征和论坛自身的主题紧密相关，人为指定特征缺乏客观性，又难以找到具有通用性的页面特征，上述方法均难以满足通用正文提取的需求。目前，基于词汇关键度的网络论坛正文提取方法尚未出现。

发明内容

[0003] 为了解决现有技术的不足，本发明提供了一种基于词汇关键度的网络论坛正文提取方法，根据页面内有意义文本中各个词汇的关键度，选择关键词频繁出现的区域，以此指导正文提取，具有高准确度和执行效率。

[0004] 本发明采用的技术方案为：

[0005] 一种基于词汇关键度的网络论坛正文提取方法，包括：抽取数据集中部分主题帖页面样本，去除非正文部分，对剩余内容进行分词，计算所有词汇的关键度，定位关键度最高的部分词汇所在位置，引导提取数据集中剩余主题帖页面的正文，若正文内容正确则输出正文，若不正确，则从抽取数据集中部分主题帖页面样本开始重新处理。

[0006] 进一步的，所述去除非正文部分包括：

[0007] 去除主题帖页面中显著的非正文内容；根据停用词库排除主题帖页面中无关内容；根据相似规则去除主题帖页面中不应被包含在正文中的内容。

[0008] 进一步的，去除主题帖页面中显著的非正文内容包括：去除主题帖页面源码中的标签及其内容，所述标签至少包括：<head>、<script>和<a>。

[0009] 进一步的，根据停用词库排除主题帖页面中无关内容包括：根据停用词库，将出现停用词的整行文本去除；或者根据停用词库与待测文本比对以决定是否保留该段待测文本。

[0010] 进一步的，根据相似规则去除主题帖页面中不应被包含在正文中的内容包括：

[0011] 比对两段待测文本的若干起始字符，判断是否保留这两段文本；或者根据相似规则去除由程序生成的不应包含在正文中的内容。

[0012] 进一步的，采用TF-IDF方法计算所有自会的关键度。

[0013] 进一步的，所述定位关键度最高的部分词汇所在位置，引导提取数据集中剩余主题帖页面的正文包括：

[0014] 在主题帖页面内对关键度最高的词汇打分，选出正文出现概率最高的位置；

[0015] 在不同页面中，对所述正文出现概率最高的位置再次打分，以确定正文位置；

[0016] 根据所确定的正文位置，提取数据集中剩余主题帖页面的正文。

[0017] 进一步的，根据所确定的正文位置，提取数据集中剩余主题帖页面的正文包括：

[0018] 先根据所确定的正文位置构造解析规则，再根据所述解析规则对数据集中剩余主题帖页面进行正文提取。

[0019] 本发明还提出了一种计算机可读存储介质，其中存储有多条指令，所述指令适于由处理器加载并执行以下处理：

[0020] 抽取数据集中部分主题帖页面样本，去除显著的非正文部分，对剩余内容进行分词，计算所有词汇的关键度，定位关键度最高的部分词汇所在位置，引导提取数据集中剩余主题帖页面的正文，若正文内容正确则输出正文，若不正确，则从抽取数据集中部分主题帖页面样本开始重新处理。

[0021] 本发明还提出了一种基于词汇关键度的网络论坛正文提取装置，包括处理器和计算机可读存储介质，处理器用于实现各指令；计算机可读存储介质用于存储多条指令，所述指令适于由处理器加载并执行以下处理：

[0022] 抽取数据集中部分主题帖页面样本，去除显著的非正文部分，对剩余内容进行分词，计算所有词汇的关键度，定位关键度最高的部分词汇所在位置，引导提取数据集中剩余主题帖页面的正文，若正文内容正确则输出正文，若不正确，则从抽取数据集中部分主题帖页面样本开始重新处理。

[0023] 本发明的有益效果：

[0024] 本发明是一种基于词汇关键度的网络论坛正文提取方法，在去除无关内容后，根据词汇的关键度确定正文的位置，使得可以将该位置信息用于对同论坛的大规模正文提取中，具有高准确度和执行效率。

附图说明

[0025] 图1为本发明完整流程的流程图；

具体实施方式：

[0026] 下面结合附图与实施例对本发明作进一步说明：

[0027] 应该指出，以下详细说明都是例示性的，旨在对本申请提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本申请所属技术领域的普通技术人员通常理解的相同含义。

[0028] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制根据本申请的示例性实施方式。如在这里所使用的，除非上下文另外明确指出，否则单数形式也意图包括复数形式，此外，还应当理解的是，当在本说明书中使用术语“包含”和/或“包括”时，其指明存在特征、步骤、操作、器件、组件和/或它们的组合。

[0029] 本发明的典型实施例是一种基于词汇关键度的网络论坛正文提取方法，包括：抽取数据集中部分主题帖页面样本，去除显著的非正文部分，对剩余内容进行分词，计算所有词汇的关键度，在主题帖页面内对关键度最高的词汇打分，选出正文出现概率最高的位置；在不同页面中，对所述正文出现概率最高的位置再次打分，以确定正文位置；根据所确定的

正文位置，提取数据集中剩余主题帖页面的正文，若正文内容正确则输出正文，若不正确，则从抽取数据集中部分主题帖页面样本开始重新处理。

[0030] 去除非正文部分包括：

[0031] 去除主题帖页面中显著的非正文内容；根据停用词库排除主题帖页面中无关内容；根据相似规则去除主题帖页面中不应被包含在正文中的内容。

[0032] 对应的虚拟模块是无关内容去除模块，用于去除页面中显著的非正文内容；停用规则模块，用于根据停用词库排除无关内容；相似规则模块，用于去除由程序生成的，不应被包含在正文中的内容；分词模块，用于将大段文本拆分为若干词汇；关键度评价模块，用于评价所有词汇的关键度；打分模块，用于避免网页噪声造成的正文定位错误；提取模块，用于根据打分定位结果从网页总体中大规模提取正文；

[0033] 去除主题帖页面中显著的非正文内容包括：去除主题帖页面源码中的标签及其内容，所述标签至少包括：<head>、<script>和<a>。

[0034] 根据停用词库排除主题帖页面中无关内容包括：根据停用词库，将出现停用词的整行文本去除；或者根据停用词库与待测文本比对以决定是否保留该段待测文本。

[0035] 根据相似规则去除主题帖页面中不应被包含在正文中的内容包括：

[0036] 比对两段待测文本的若干起始字符，判断是否保留这两段文本；或者根据相似规则去除由程序生成的不应包含在正文中的内容。

[0037] 本实施例中采用TF-IDF方法计算所有自会的关键度。

[0038] 本发明还提出了一种计算机可读存储介质和一种基于词汇关键度的网络论坛正文提取装置，存储介质中存储有多条指令，所述指令适于由处理器加载并执行以下处理：

[0039] 抽取数据集中部分主题帖页面样本，去除显著的非正文部分，对剩余内容进行分词，计算所有词汇的关键度，定位关键度最高的部分词汇所在位置，引导提取数据集中剩余主题帖页面的正文，若正文内容正确则输出正文，若不正确，则从抽取数据集中部分主题帖页面样本开始重新处理。

[0040] 下面给出一个实际应用例：

[0041] 我们利用网络爬虫在互联网上爬取了若干论坛页面，去除非主题帖页面后，所有论坛剩余主题帖页面总计3417，这些页面来自13个不同的网络论坛。本例旨在从主题帖中提取出正文信息。

[0042] 步骤一：抽样。从每个论坛中随机抽取5个页面进行训练，以下将针对单一论坛的处理过程进行描述。

[0043] 步骤二：无关内容去除。将网页源码中的<head>,<a>,<script>等标签及其内容去除。主题帖页面的正文应当是在<body>部分的纯文本，因而<head>标签所包含的网页元数据和<a>标签所包含的超链接等均显然不是正文，同样，<script>标签包含的脚本代码也非正文，应当去除。

[0044] 步骤三：标签去除。将网页源码中所有标签去除，保留其包含的内容。至此，剩余的源码应均为能够在网页中显示为纯文本的内容，源码中的每一行对应于网页中的一段文字。

[0045] 步骤四：停用词去除。根据停用词库，将出现停用词的整行文本去除。如某行中出现“版权所有”，由于此文本常出现于版权声明中，有理由认为其非正文，应将其所在的整行

文本去除。

[0046] 步骤五:相似词去除。以每行起始的若干字符为键,若键相同,则将这两行都去除。如某行中出现“发表于”,由于此文本常出现于描述发帖时间的部分,由后台程序自动生成,而非用户所撰写,故应将其去除。同时由于不同论坛对发帖时间的描述方法存在差异,但该类文本出现频率较高,故不将其添加到停用词库中,而采用此法去除。

[0047] 步骤六:分词。将每行中的文本划分为若干词汇。对于英文论坛,可以简单按照空格和标点进行分词,对于中文论坛,需使用专业软件分词,本例中使用开源分词软件jieba。

[0048] 步骤七:计算关键度。对每个词汇,计算其关键度。本例中使用较为通用的词汇关键度评价方法TF-IDF方法,关键度较高的词汇有更高的概率出现在正文部分。其中TF-IDF方法的公式为

[0049] $$TF(f,w) = N(f,w) / \sum_{i \in F} N(i,w) \quad (1)$$

[0050] 其中,F为语料库中所有文件的集合,f∈F为文件,w为词汇,N(f,w)为词汇w在文件f中出现的次数。

[0051] $$IDF(f,w) = \log\left[\sum_{i \in F} 1 / \left(1 + \sum_{i \in F} P(i,w)\right)\right] \quad (2)$$

[0052] 其中,P(f,w)为词汇出现次数,当词汇w出现在文件f中,P(f,w)为1,否则为0,分母部分+1是为了防止w在语料库中未出现,导致分母为0的情况。

[0053] W(f,w)=TF(f,w)×IDF(f,w) (3)

[0054] 其中,W为词汇w的重要度。

[0055] 步骤八:定位关键词。记录关键度最高的若干词汇所在的位置。本例中使用源码中标签的class属性作为位置信息记录。

[0056] 步骤九:页面内打分。打分选出正文出现概率最高的位置。其实质是通过比较同一class属性值的出现次数判断正文的位置,在进行过步骤二到步骤五的预处理步骤后,正文应当占据剩余内容的主要部分。

[0057] 步骤十:页面间打分。将不同页面判断的正文位置再次打分,以确定正文位置。单一页面可能由于诸如回帖过少,图片过多等异常导致正文位置判断错误,对多个页面进行打分可以降低最终结果出现错误的概率。

[0058] 步骤十一:构造解析规则。根据步骤十的结果构造解析规则,用于大规模正文提取。本例中由于使用的是class属性值,因而无需额外处理即可直接使用。

[0059] 步骤十二:应用解析规则。利用步骤十一得到的解析规则对所有数据集中所有主题帖页面进行正文提取。具体而言,从各网页源码中提取具有相同class属性值的标签,将其中包含的标签去除,余下的内容即是正文。

[0060] 在本例中,我们每次从同一论坛的页面中随机抽取5个页面进行训练,得到解析规则,如此对每个论坛重复进行10次,分析解析规则的正确性。我们对前十一个步骤进行了计时,以体现本方法的效率。详细结果如表1所示。

[0061] 表1正文提取方法结果

[0062]

序号	正确数/次	平均用时/秒
1	10	28.61
2	10	44.02
3	7	38.80
4	9	9.62
5	8	37.31
6	9	25.52
7	10	21.53
8	10	12.14
9	9	18.29
10	8	11.98
11	10	4.61
12	10	18.25
13	8	96.46

[0063] 从表中可以看出，本发明提出的正文提取方法具有较高的准确率，在用时上，单网页的解析速度视内容量在1s到9s内浮动，效率较高。

[0064] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

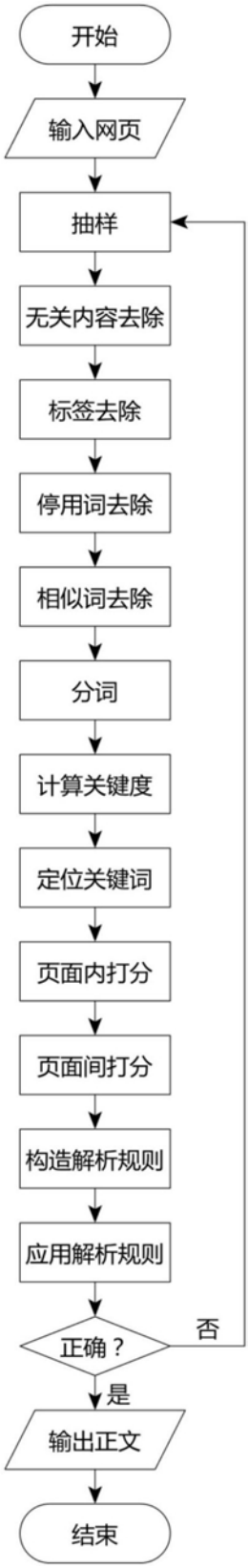
开始
输入网页
抽样
无关内容去除
标签去除
停用词去除
相似词去除
分词
计算关键度
定位关键词
页面内打分
页面间打分
构造解析规则
应用解析规则
正确？
否
是
输出正文
结束

图1