(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 108829660 B
(45)授权公告日 2021.08.31

(21)申请号 201810437383.7

(22)申请日 2018.05.09

(65)同一申请的已公布的文献号
申请公布号 CN 108829660 A

(43)申请公布日 2018.11.16

(73)专利权人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 佘堃 廖贞林

(74)专利代理机构 成都弘毅天承知识产权代理有限公司 51230
代理人 徐金琼

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 16/33(2019.01)
G06F 21/10(2013.01)

(56)对比文件
CN 103441924 A,2013.12.11
CN 102682104 A,2012.09.19
CN 101453331 A,2009.06.10
CN 104636325 A,2015.05.20
CN 105653984 A,2016.06.08
CN 104715168 A,2015.06.17
CN 105376050 A,2016.03.02
US 8028039 B1,2011.09.27
VN 10008702 B,2010.10.25
刘兆丽等.数字签名研究的现状与发展.《电脑知识与技术》.2008,552-554,562.

审查员 蔡震震

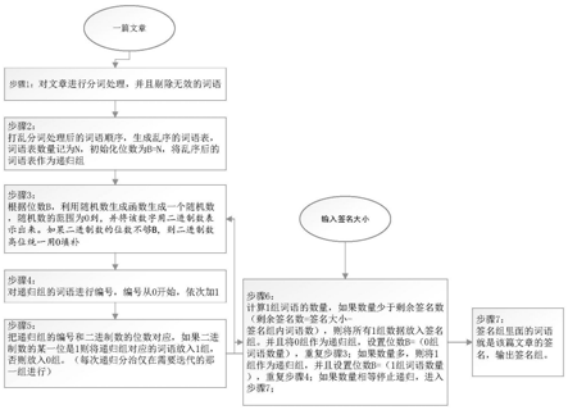
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种基于随机数分治递归的短文本签名生成方法

(57)摘要

本发明的提供一种基于随机数分治递归的短文本签名生成方法,属于信息技术领域中的文章签名方法领域,包括如下步骤:提出所有的无效词语得到有效词语;打乱所有有效词语的顺序生成乱序的词语表,将此词语表作为递归组;生成一个随机数,随机数的范围为0到2^B,再用二进制数表示随机数,如果二进制数的位数小于B,则二进制数高位统一用0填补;如果“1”组词语数量等于剩余签名数,则停止递归;根据剩余签名的数量,如果递归停止条件不满足,选择0组和1组中的其中一组作为递归组进行递归,直到递归条件满足;最后输出签名组里的所有词语作为输入文章的签名。本发明解决现有的文章签名生成方法速度慢、步骤复杂的问题。



CN 108829660 B

1.一种基于随机数分治递归的短文本签名生成方法,其特征在于,包括如下步骤:

步骤1:输入文章需要的签名大小;

步骤2:对文章进行分词处理,再剔除所有的无效词语得到有效词语;

步骤3:打乱所有有效词语的顺序生成乱序的词语表,将此词语表作为递归组,词语表数量记为N,初始化位数为B=N;

步骤4:根据位数B,生成一个随机数,随机数的范围为0到2^B,再用二进制数表示随机数,如果二进制数的位数小于B,则二进制数高位统一用0填补;

步骤5:将递归组的词语按顺序与二进制数对应,将对应的是二进制数1的词语放入“1”组,否则放入“0”组;

步骤6:如果“1”组词语数量小于剩余签名数,其中,剩余签名数=签名大小-签名组内词语数,则将所有“1”组数据放入签名组,并且将“0”组作为递归组进行递归;然后进入步骤7;

如果“1”组词语数量大于剩余签名数,则将所有“1”组词语在放入递归组进行递归;然后进入步骤7;

如果“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

步骤7:重新设定位数B为步骤7得到的递归组的的词语数,然后重复步骤4-6,直到“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

步骤8:输出签名组里的所有词语作为输入文章的签名。

2.根据权利要求1所述的一种基于随机数分治递归的短文本签名生成方法,其特征在于,所述步骤5的具体步骤为:

步骤5.1:对递归组的词语进行从0-N进行编号,编号从0开始,依次加1;

步骤5.2:将编号与二进制数进行对应,如果编号对应的是二进制数中的1,则将该编号所代表的词语放进“1”组;如果编号对应的是二进制数中的0,则将该编号所代表的词语放进“0”组。

一种基于随机数分治递归的短文本签名生成方法

技术领域

[0001] 本发明属于信息技术领域中的文章签名方法领域,具体为一种基于随机数分治递归的短文本签名生成方法。

背景技术

[0002] 当今社会,信息技术的快速发展在给人们提供便利的同时也带来了诸多挑战。在生活中,处处存在着抄袭的现象,网络文章的抄袭更是防不胜防,对此我们需要更多的算法来解决现在网络上存在的文章抄袭问题。该方法旨在发明一种新型的文章签名方法用于文本抄袭检测中。现有的文章签名方法于minhash是每次提前一个排最前面的特征,一次只能产生一个签名,并且minhash每生成一个签名需要一个函数,况且函数不能重复,且函数要事先设计,因此,这大大降低了文章签名的生成方法。

发明内容

[0003] 本发明的目的在于:为解决现有的文章签名生成方法速度慢、步骤复杂的问题,本发明提供一种基于随机数分治递归的短文本签名生成方法。

[0004] 本发明的技术方案如下:

[0005] 一种基于随机数分治递归的短文本签名生成方法,包括如下步骤:

[0006] 步骤1:输入文章需要的签名大小;

[0007] 步骤2:对文章进行分词处理,再剔除所有的无效词语得到有效词语;

[0008] 步骤3:打乱所有有效词语的顺序生成乱序的词语表,将此词语表作为递归组,词语表数量记为N,初始化位数为B=N;

[0009] 步骤4:根据位数B,生成一个随机数,随机数的范围为0到2^B,再用二进制数表示随机数,如果二进制数的位数小于B,则二进制数高位统一用0填补;

[0010] 步骤5:将递归组的词语按顺序与二进制数对应,将对应的是二进制数1的词语放入“1”组,否则放入“0”组。

[0011] 步骤6:如果“1”组词语数量小于剩余签名数,其中,剩余签名数=签名大小-签名组内词语数,则将所有“1”组数据放入签名组,并且将“0”组作为递归组进行递归;然后进入步骤7;

[0012] 如果“1”组词语数量大于剩余签名数,则将所有“1”组词语在放入递归组进行递归;然后进入步骤7;

[0013] 如果“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

[0014] 步骤7:重新设定位数B为步骤7得到的递归组的的词语数,然后重复步骤4-6,直到“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

[0015] 步骤8:输出签名组里的所有词语作为输入文章的签名。

[0016] 具体地,所述步骤5的具体步骤为:

[0017] 步骤5.1:对递归组的词语进行从0-N进行编号,编号从0开始,依次加1;

[0018] 步骤5.2:将编号与二进制数进行对应,如果编号对应的是二进制数中的1,则将该编号所代表的词语放进“1”组;如果编号对应的是二进制数中的0,则将该编号所代表的词语放进“0”组。

[0019] 采用上述方案后,本发明有益效果如下:

[0020] (1)本发明的方法通过产生一个随机数将词语表分成两部分,直接把随机数变成二进制数表示1表示签名,0表示非签名,进行分治处理,又通过递归准则,对相应的部分进行递归处理,然后不断重复直到取得要求数量的签名。随机数用二进制表现出来,出现0和1出现概率相差不大,一下子可以出现非常多签名然后可以按需要的签名数量通过分治递归的处理获得文章的签名组,以便输出文章的签名,就不需要像minhash一样一次只能产生一个签名,大大提高了提取的速度。

[0021] (2)本发明中随机数生成也比较简单,不需要像minhash方法一样每一个签名需要一个函数,更加无需事先设计不能重复的函数,提高了便捷性和实用性,将本发明用于快速比较文本的相似度方面效果更佳。

附图说明

[0022] 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,下面描述中的附图仅仅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。通过附图所示,本发明的上述及其它目的、特征和优势将更加清晰。在全部附图中相同的附图标记指示相同的部分。并未刻意按实际尺寸等比例缩放绘制附图,重点在于示出本发明的主旨。

[0023] 图1为本发明的流程图。

具体实施方式

[0024] 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。

[0025] 本发明具体技术涉及有利用随机数生成函数生成一个随机数,分词的技术,二者均为现有技术,下面,将简要描述这两个技术的过程。

[0026] 随机数的生成:

[0027] 随机数的生成采用线性同余随机数生成方法。该方法代表了最好最朴素的伪随机数产生器算法,并且容易理解,容易实现,而且速度快。线性同余随机数生成算法数学上基于公式:

[0028] $X(n+1) = (a*X(n)+c) \% m$

[0029] 其中,各系数为:

[0030] 模m,$m>0$

[0031] 系数a,$0<a<m$

[0032] 增量c,$0<=c<m$

[0033] 原始值(种子)0<＝X(0)<m

[0034] 其中本方法中m＝2^32,a＝22695477,c＝1;

[0035] 当我们产生随机数之后,将随机数规范到指定范围,并且将随机数表示成二进制方式。

[0036] 分词方法:

[0037] 分词的方法采用了中科院的分词系统,当我们将一篇短文本进行分词之后,我们可以得到词语和词语对应的识别信息。根据识别信息,我们会去除一些无效的语义词,比如“的”,“你”等。

[0038] 本发明的一种基于随机数分治递归的短文本签名生成方法,包括如下步骤:

[0039] 步骤1:输入文章需要的签名大小;

[0040] 步骤2:对文章进行分词处理,再剔除所有的无效词语得到有效词语;

[0041] 步骤3:打乱所有有效词语的顺序生成乱序的词语表,将此词语表作为递归组,词语表数量记为N,初始化位数为B＝N;

[0042] 步骤4:根据位数B,生成一个随机数,随机数的范围为0到2^B,此处的范围包含了边缘值;再用二进制数表示随机数,如果二进制数的位数小于B,则二进制数高位统一用0填补;

[0043] 步骤5:将递归组的词语按顺序与二进制数对应,由于位数相同,所以对应的具体方式按照从高到低或者从低到高均可,将对应的是二进制数1的词语放入“1”组,否则放入“0”组;所述步骤5的具体步骤为:

[0044] 步骤5.1:对递归组的词语进行从0-N进行编号,编号从0开始,依次加1;

[0045] 步骤5.2:将编号与二进制数进行对应,如果编号对应的是二进制数中的1,则将该编号所代表的词语放进“1”组;如果编号对应的是二进制数中的0,则将该编号所代表的词语放进“0”组。

[0046] 步骤6:如果“1”组词语数量小于剩余签名数,其中,剩余签名数＝签名大小-签名组内词语数,则将所有“1”组数据放入签名组,并且将“0”组作为递归组进行递归;然后进入步骤7;

[0047] 如果“1”组词语数量大于剩余签名数,则将所有“1”组词语在放入递归组进行递归;然后进入步骤7;如果“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

[0048] 步骤7:重新设定位数B为步骤7得到的递归组的的词语数,然后重复步骤4-6,直到“1”组词语数量等于剩余签名数,则停止递归,进入步骤8;

[0049] 步骤8:输出签名组里的所有词语作为输入文章的签名。

[0050] 本发明中所称的短文本为500个字左右的文本,由于对象是短文本,即使是文章有一些随机的因素,也不会对精确度有太大的影响,在对比两篇文章的相似度方面,本发明具有重大意义。

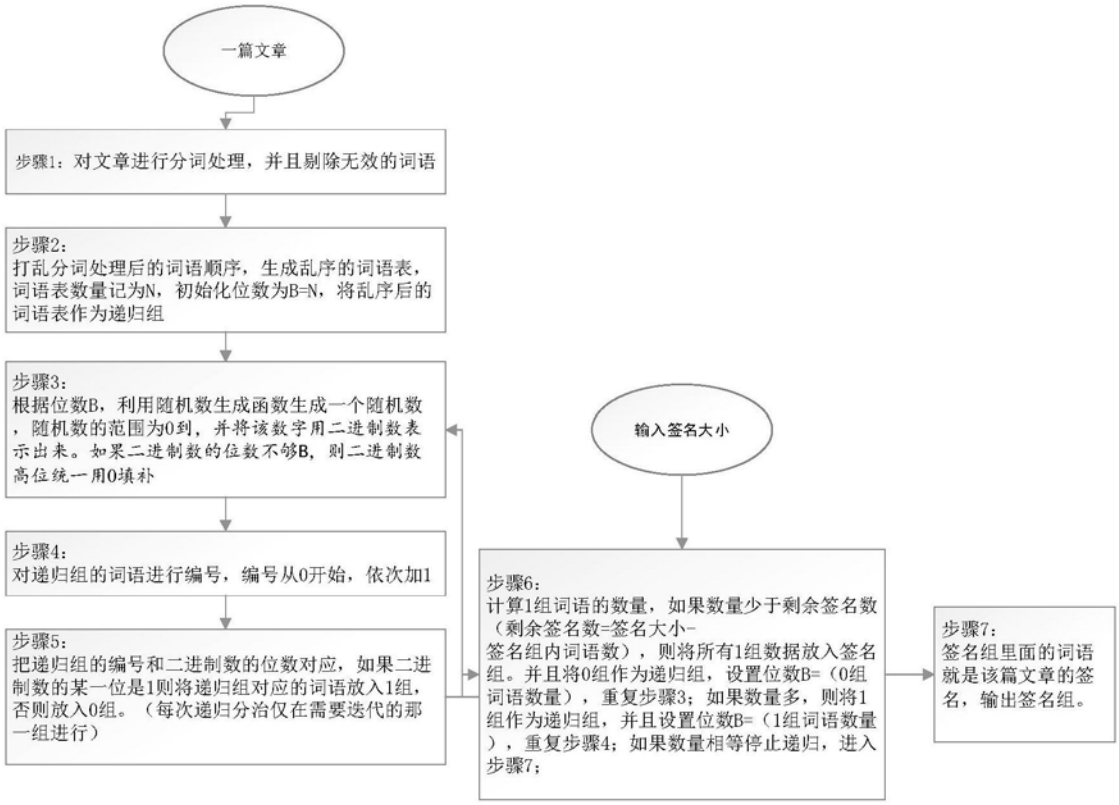

一篇文章
步骤1：对文章进行分词处理，并且剔除无效的词语
步骤2：
打乱分词处理后的词语顺序，生成乱序的词语表，词语表数量记为N，初始化位数为B=N，将乱序后的词语表作为递归组
步骤3：
根据位数B，利用随机数生成函数生成一个随机数，随机数的范围为0到，并将该数字用二进制数表示出来。如果二进制数的位数不够B，则二进制数高位统一用0填补
步骤4：
对递归组的词语进行编号，编号从0开始，依次加1
步骤5：
把递归组的编号和二进制数的位数对应，如果二进制数的某一位是1则将递归组对应的词语放入1组，否则放入0组。（每次递归分治仅在需要迭代的那一组进行）
输入签名大小
步骤6：
计算1组词语的数量，如果数量少于剩余签名数（剩余签名数=签名大小-
签名组内词语数），则将所有1组数据放入签名组。并且将0组作为递归组，设置位数B=（0组词语数量），重复步骤3；如果数量多，则将1组作为递归组，并且设置位数B=（1组词语数量），重复步骤4；如果数量相等停止递归，进入步骤7；
步骤7：
签名组里面的词语就是该篇文章的签名，输出签名组。

图1