(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 112364647 A
(43)申请公布日 2021.02.12

(21)申请号 202011326607.0

(22)申请日 2020.11.24

(71)申请人 南方电网海南数字电网研究院有限公司
地址 570100 海南省海口市美兰区海府路32号

(72)发明人 陈文博 胡微 王鹏 王保强 陈余

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 陈欢

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 40/216(2020.01)

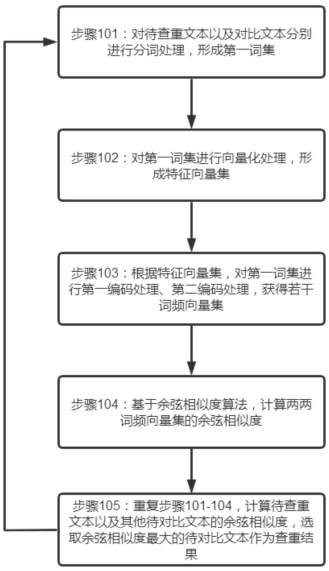
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种基于余弦相似度算法的查重方法

(57)摘要

本发明提供一种基于余弦相似度算法的查重方法，包括下列步骤：对待查重文本以及对比文本分别进行分词处理，形成第一词集；对第一词集进行向量化处理，形成特征向量集；根据特征向量集，对第一词集进行第一编码处理，形成包含若干编码子集的第二词集组，对第二词集组分别进行第二编码处理，形成包含若干词频向量集的第三词集组；基于余弦相似度算法，计算两两词频向量集的余弦相似度，若余弦相似度大于阈值，则将对比文本作为查重结果。



CN 112364647 A

1.一种基于余弦相似度算法的查重方法,其特征在于,包括下列步骤:

对待查重文本以及任一待对比文本分别进行分词处理,形成第一词集;

对第一词集进行向量化处理,形成特征向量集;

根据特征向量集,对第一词集进行第一编码处理,形成包含若干编码子集的第二词集组,对第二词集组分别进行第二编码处理,形成包含若干词频向量集的第三词集组;

基于余弦相似度算法,计算两两词频向量集的余弦相似度;

重复上述步骤,计算待查重文本以及其他待对比文本的余弦相似度,选取余弦相似度最大的待对比文本作为查重结果。

2.根据权利要求1所述的一种基于余弦相似度算法的查重方法,其特征在于,对待查重文本以及任一待对比文本分别进行分词处理,形成第一词集,包括:

对待查重文本分词处理,形成包含多个单词的第一分词集,对任一待对比文本进行分词处理,形成包含多个单词的第二分词集;

将第一分词集、第二分词集中的单词进行逐一对比,若存在相同单词,则选取一个相同单词放入第一词集中;

若存在不同单词,则将不同单词均放入第一词集中。

3.根据权利要求2所述的一种基于余弦相似度算法的查重方法,其特征在于,所述分词处理包括结巴分词法中的一种或多种。

4.根据权利要求1-3任一项所述的一种基于余弦相似度算法的查重方法,其特征在于,对第一词集进行向量化处理,形成特征向量集,包括:

对第一词集中的单词的出现顺序进行数字标号,形成包含单词以及数字的特征向量集。

5.根据权利要求4所述的一种基于余弦相似度算法的查重方法,其特征在于,所述第一编码处理包括:

根据特征向量集,将第一分词集转换为包含数字的第一编码子集,将第二分词集转换为包含数字的第二编码子集,所述第一编码子集、第二编码子集组成第二词集组。

6.根据权利要求5所述的一种基于余弦相似度算法的查重方法,其特征在于,第二编码处理,包括:

对第一编码子集进行oneHot编码处理,获得第一词频向量集;

对第二编码子集进行oneHot编码处理,获得第二词频向量集;

所述第一词频向量集、第二词频向量集组成第三词集组。

一种基于余弦相似度算法的查重方法

技术领域

[0001] 本发明涉及数据查重技术领域，尤其涉及一种基于余弦相似度算法的查重方法。

背景技术

[0002] 随着计算机文本信息挖掘等各种自然语言处理应用的普及，当今社会对基于文本相似度的文档检索系统需求日益增加，同时人们对计算机文本处理也提出了更高的要求。在自然语言处理过程中，经常会涉及到如何度量两个文本之间的相似性，我们都知道文本是一种高维的语义空间，如何对其进行抽象分解，从而能够站在数学角度去量化其相似性，是此方法的重点。在相似度检索领域，现有的相似度检索方法，要么在检索效率上存在不足，要么在准确度方面不能令人满意。

发明内容

[0003] 本发明的目的在于提供一种基于余弦相似度算法的查重方法，以解决上述背景技术中提出的问题。

[0004] 本发明是通过以下技术方案实现的：一种基于余弦相似度算法的查重方法，包括下列步骤：

[0005] 对待查重文本以及对比文本分别进行分词处理，形成第一词集；

[0006] 对第一词集进行向量化处理，形成特征向量集；

[0007] 根据特征向量集，对第一词集进行第一编码处理，形成包含若干编码子集的第二词集组，对第二词集组分别进行第二编码处理，形成包含若干词频向量集的第三词集组；

[0008] 基于余弦相似度算法，计算两两词频向量集的余弦相似度；

[0009] 重复上述步骤，计算待查重文本以及其他待对比文本的余弦相似度，选取余弦相似度最大的待对比文本作为查重结果。

[0010] 优选的，对待查重文本以及对比文本分别进行分词处理，形成第一词集，包括：

[0011] 对待查重文本分词处理，形成包含多个单词的第一分词集，对对比文本进行分词处理，形成包含多个单词的第二分词集；

[0012] 将第一分词集、第二分词集中的单词进行逐一对比，若存在相同单词，则选取一个相同单词放入第一词集中；

[0013] 若存在不同单词，则将不同单词均放入第一词集中。

[0014] 优选的，所述分词处理包括结巴分词法中的一种或多种。

[0015] 优选的，对第一词集进行向量化处理，形成特征向量集，包括：

[0016] 对第一词集中的单词的出现顺序进行数字标号，形成包含单词以及数字的特征向量集。

[0017] 优选的，所述第一编码处理包括：

[0018] 根据特征向量集，将第一分词集转换为包含数字的第一编码子集，将第二分词集转换为包含数字的第二编码子集，所述第一编码子集、第二编码子集组成第二词集组。

[0019] 优选的，第二编码处理，包括：

[0020] 对第一编码子集进行oneHot编码处理，获得第一词频向量集；

[0021] 对第二编码子集进行oneHot编码处理，获得第二词频向量集；

[0022] 所述第一词频向量集、第二词频向量集组成第三词集组。

[0023] 与现有技术相比，本发明达到的有益效果如下：

[0024] 本发明提供的一种基于余弦相似度算法的查重方法，可提高系统查重效率及准确率，减少人力资源浪费。

附图说明

[0025] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的优选实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0026] 图1为本发明提供的一种基于余弦相似度算法的查重方法的流程图。

具体实施方式

[0027] 为了更好理解本发明技术内容，下面提供具体实施例，并结合附图对本发明做进一步的说明。

[0028] 参见图1，一种基于余弦相似度算法的查重方法，包括下列步骤：

[0029] 步骤101：对待查重文本以及对比文本分别进行分词处理，形成第一词集；

[0030] 具体的，对待查重文本分词处理，形成包含多个单词的第一分词集，对对比文本进行分词处理，形成包含多个单词的第二分词集；

[0031] 将第一分词集、第二分词集中的单词进行逐一对比，若存在相同单词，则选取一个相同单词放入第一词集中；

[0032] 若存在不同单词，则将不同单词均放入第一词集中。

[0033] 在本实施例的一种实施方式中，所述分词处理包括结巴分词法中的一种或多种。

[0034] 步骤102：对第一词集进行向量化处理，形成特征向量集；

[0035] 具体的，对第一词集中的单词的出现顺序进行数字标号，形成包含单词以及数字的特征向量集。

[0036] 步骤103：根据特征向量集，对第一词集进行第一编码处理，形成包含若干编码子集的第二词集组，对第二词集组分别进行第二编码处理，形成包含若干词频向量集的第三词集组；

[0037] 具体的，第一编码处理包括：根据特征向量集，将第一分词集转换为包含数字的第一编码子集，将第二分词集转换为包含数字的第二编码子集，所述第一编码子集、第二编码子集组成第二词集组。

[0038] 第二编码处理，包括：

[0039] 对第一编码子集进行oneHot编码处理，获得第一词频向量集；

[0040] 对第二编码子集进行oneHot编码处理，获得第二词频向量集；

[0041] 所述第一词频向量集、第二词频向量集组成第三词集组。

[0042] 步骤104:基于余弦相似度算法,计算两两词频向量集的余弦相似度;

[0043] 步骤105:重复步骤101-104,计算待查重文本以及其他待对比文本的余弦相似度,选取余弦相似度最大的待对比文本作为查重结果。

[0044] 若余弦相似度大于阈值,则将对比文本作为查重结果。

[0045] 下面以句子A以及句子B为例进行说明。

[0046] 句子A为待查重文本:这只皮靴号码大了。那只号码合适。

[0047] 句子B为对比文本:这只皮靴号码不小,那只更合适。

[0048] 对句子A以结巴分词的方式进行分词处理,得到第一分词集:

[0049] 第一分词集=[‘这‘,‘只‘,‘皮靴‘,‘号码‘,‘大‘,‘了‘,‘那‘,‘只‘,‘号码‘,‘合适‘];

[0050] 对句子B以结巴分词的方式进行分词处理,得到第二分词集:

[0051] 第二分词集=[‘这‘,‘只‘,‘皮靴‘,‘号码‘,‘不小‘,‘那‘,‘只‘,‘更合‘,‘合适‘]。

[0052] 对比第一分词集、第二分词集,将第一分词集、第二分词集中的单词进行逐一对比,若存在相同单词,则选取一个相同单词放入第一词集中,最终获得如下第一词集:

[0053] 第一词集={’不小’,’了’,’合适’,’那’,’只’,’皮靴’,’更合’,’号码’,’这’,’大’}。

[0054] 按照第一词集中各个单词出现的顺序进行标号,用以实现第一词集的向量化处理,最终结果如下:

[0055] 特征向量集={’不小’:0,’了’:1,’合适’:2,’那’:3,’只’:4,’皮靴’:5,’更合’:6,’号码’:7,’这’:8,’大’:9}

[0056] 根据特征向量集对第一分词集、第二分词集进行第一编码处理,获得如下结果:

[0057] 第一编码子集=[8,4,5,7,9,1,3,4,7,2];

[0058] 第二编码子集=[8,4,5,7,0,3,4,6,2];

[0059] 对第一编码子集、第二编码子集进行oneHot编码处理,就是计算每个分词出现的次数,其结果如下:

[0060] 第一词频向量集=[0,1,1,1,2,1,0,2,1,1];

[0061] 第二词频向量集[1,0,1,1,2,1,1,1,1,0];

[0062] 得出两个句子的词频向量之后,就变成了计算第一词频向量集、第二词频向量集之间夹角的余弦值,值越大相似度越高,其具体的计算公式如下:

[0063]
$$\cos(\theta) = \frac{0*1+1*0+1*1+1*1+2*2+1*1+0*1+2*1+1*1+1*0}{\sqrt{0^2+1^2+1^2+2^2+1^2+0^2+2^2+1^2+1^2}*\sqrt{1^2+0^2+1^2+1^2+2^2+1^2+1^2+}} = 0.81$$

[0064] 同理,重新计算句子A与句子C之间的余弦相似度,计算句子A与句子D之间的余弦相似度,对三个余弦相似度进行比较,选取余弦相似度最大的待对比文本作为查重结果。

[0065] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,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,均应包含在本发明保护的范围之内。

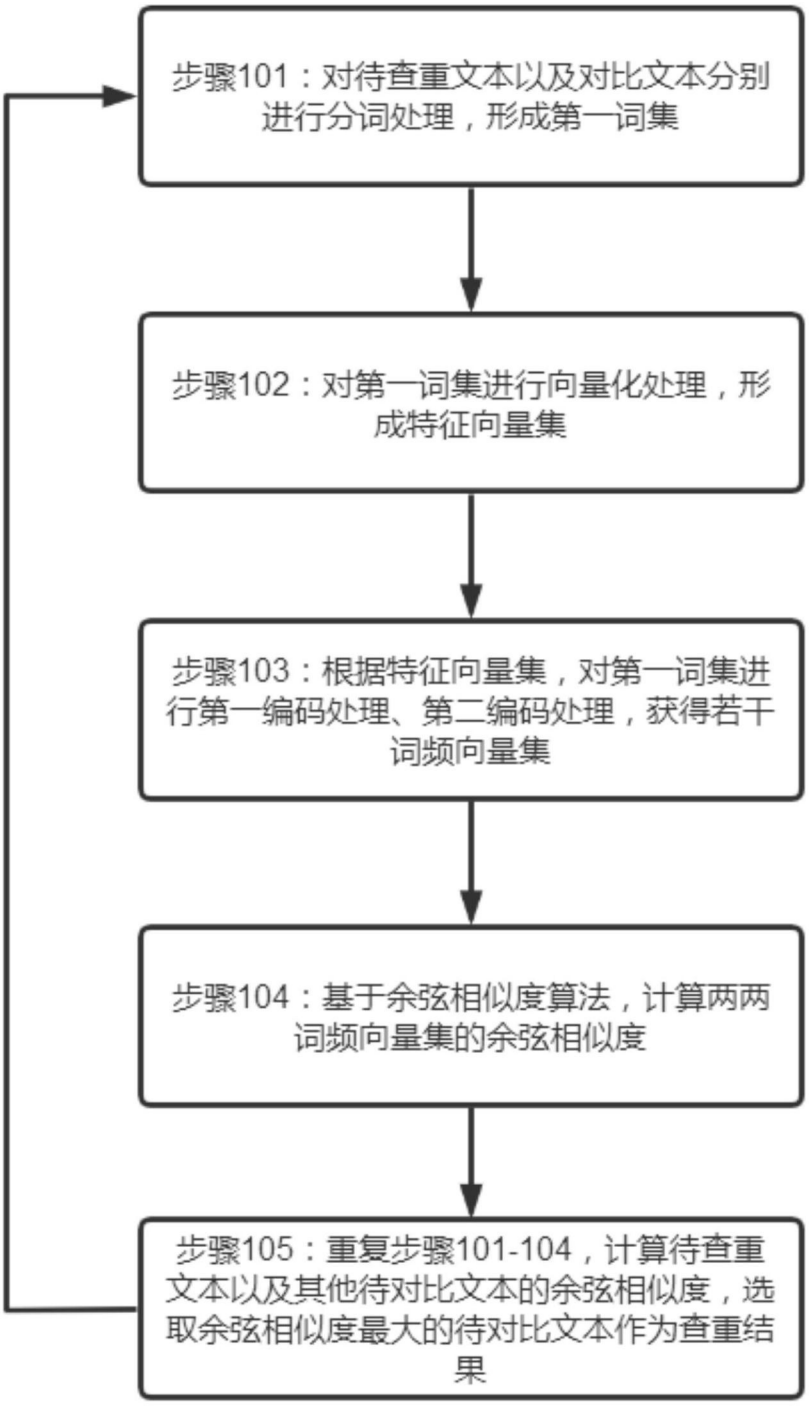

步骤101：对待查重文本以及对比文本分别进行分词处理，形成第一词集
步骤102：对第一词集进行向量化处理，形成特征向量集
步骤103：根据特征向量集，对第一词集进行第一编码处理、第二编码处理，获得若干词频向量集
步骤104：基于余弦相似度算法，计算两两词频向量集的余弦相似度
步骤105：重复步骤101-104，计算待查重文本以及其他待对比文本的余弦相似度，选取余弦相似度最大的待对比文本作为查重结果

图1