(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 110245275 A
(43)申请公布日 2019.09.17

(21)申请号 201910521164.1

(22)申请日 2019.06.18

(71)申请人 中电科大数据研究院有限公司
地址 550000 贵州省贵阳市贵阳国家高新技术产业开发区金阳科技产业园黎阳大厦
申请人 贵州华云创谷科技有限公司
长沙军民先进技术研究有限公司

(72)发明人 鲍翊平 曹扬 王进 何春辉 张翀 葛斌 夏利锋 王绍丽

(74)专利代理机构 长沙市护航专利代理事务所（特殊普通合伙）43220
代理人 谢新苗

(51)Int.Cl.
G06F 16/9032(2019.01)
G06F 16/951(2019.01)
G06F 16/9535(2019.01)

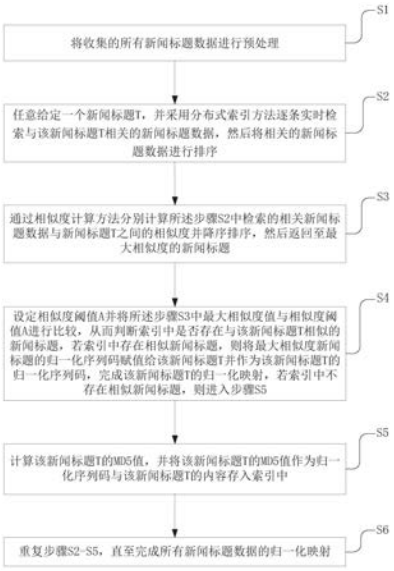
权利要求书3页 说明书7页 附图2页

(54)发明名称

一种大规模相似新闻标题快速归一化方法

(57)摘要

本发明公开了一种大规模相似新闻标题快速归一化方法，所述方法包括：S1、预处理新闻标题数据；S2、将与任意给定新闻标题相关的新闻标题数据进行排序；S3、计算相关新闻标题与该新闻标题之间相似度并排序；S4、比较相关新闻标题中最大相似度值与设定相似度阈值并判断索引中是否存在相似新闻标题，若存在则将最大相似度新闻标题MD5值作为该新闻标题归一化序列码，若不存在则进入步骤S5；S5、计算该新闻标题MD5值，作为其归一化序列码与其内容存入索引中；S6、重复步骤S2和S5，完成所有新闻标题的归一化映射。本发明可以快速找出相关的新闻标题并做出相似性判定，然后将相似新闻标题映射成唯一的归一化序列码，实现了大规模相似新闻标题的快速归一化。



CN 110245275 A

1.一种大规模相似新闻标题快速归一化方法,其特征在于,所述方法包括以下步骤:

S1、将收集的所有新闻标题数据进行预处理;

S2、任意给定一个新闻标题T,并采用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据,然后将相关的新闻标题数据进行排序;

S3、通过相似度计算方法分别计算所述步骤S2中检索的相关新闻标题数据与新闻标题T之间的相似度并降序排序,然后返回至最大相似度的新闻标题;

S4、设定相似度阈值A并将所述步骤S3中最大相似度值与相似度阈值A进行比较,从而判断索引中是否存在与该新闻标题T相似的新闻标题,若索引中存在相似新闻标题,则将最大相似度新闻标题的归一化序列码赋值给该新闻标题T并作为该新闻标题T的归一化序列码,完成该新闻标题T的归一化映射,若索引中不存在相似新闻标题,则进入步骤S5;

S5、计算该新闻标题T的MD5值,并将该新闻标题T的MD5值作为归一化序列码与该新闻标题T的内容存入索引中;

S6、重复步骤S2-S5,直至完成所有新闻标题数据的归一化映射。

2.如权利要求1所述的大规模相似新闻标题快速归一化方法,其特征在于,所述步骤S1中将收集的所有新闻标题数据进行预处理的具体实现方式为:去除新闻标题中的一些多余的空格、特殊字符和换行符等。

3.如权利要求2所述的大规模相似新闻标题快速归一化方法,其特征在于,所述步骤S2中分布式索引为Elasticsearch分布式索引。

4.如权利要求3所述的大规模相似新闻标题快速归一化方法,其特征在于,所述步骤S2中利用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据,然后将相关的新闻标题数据进行排序的具体实现方式包括:

S21、对输入的新闻标题T进行分词,并利用停用词典过滤相应的停用词,得到有效词语列表;

S22、将所述步骤S21中得到的有效词语列表作为真实的查询词语输入;

S23、计算所述步骤S22中查询词语与索引文档之间的检索得分Score(q,d),可用公式表示:

$$Score(q,d)=\sum_{k}^{n}W_k*R(q_k,d) \qquad (1)$$

式(1)中,q表示查询词语,d表示索引文档,W_k表示查询词语q中第k个有效词语的逆文档频率,q_k表示查询词语q中第k个有效词语,n表示查询词语q中有效词语的总数,$R(q_k,d)$表示查询词语q中第k个有效词语与索引文档d之间的相关性;

其中,$W_k=log\left(\frac{N+1}{n(q_k)+1}\right)$,N表示索引文档的总数量,$n(q_i)$表示包含查询词语$q_k$的索引文档数量,1表示调节因子;$R(q_k,d)=\frac{f_k(h_1+1)}{f_k+H}*\frac{qf_k(h_2+1)}{qf_k+h_2}$,$h_1$和$h_2$表示调整系数,$f_k$表示查询词语$q_k$在索引文档d中的频率,$qf_k$表示查询词语$q_k$在查询词语q中的频率,H表示比例系数,$H=h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)$,其中b为调节系数,dl表示当前从索引中取出来与有效检索

词语相关的新闻标题的长度，avg(dl)表示从索引中检索出来与当前有效检索词语相关的全部新闻标题的平均长度，从而式(1)可表示为：

$$Score(q,d)=\sum_{k}^{n} log\left(\frac{N+1}{n(q_k)+1}\right)*\frac{f_k(h_1+1)}{f_k+h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)}*\frac{qf_k(h_2+1)}{qf_k+h_2};$$

S24、根据所述步骤S23计算出来的检索得分Score(q,d)对与新闻标题T相关的新闻标题数据进行排序。

5.如权利要求4所述的大规模相似新闻标题快速归一化方法，其特征在于，所述步骤S3中相似度计算方法为改进型Jaro-Winkler短文本相似度计算方法，所述方法的匹配窗口包括强匹配窗口和弱匹配窗口，所述强匹配窗口和弱匹配窗口的值可用公式表示：

$$SMW=\frac{max(L(s_1),\ L(s_2))}{2}-1 \qquad (2)$$

$$WMW=max(L(s_1),L(s_2))-index \qquad (3)$$

式(2)、(3)中，SMW表示强匹配窗口的值，WMW表示弱匹配窗口的值，s_1，s_2表示字符串，$L(s_1)$表示字符串s_1的长度，$L(s_2)$表示字符串s_2的长度，index表示当前强匹配窗口结束位置的值。

6.如权利要求5所述的大规模相似新闻标题快速归一化方法，其特征在于，所述步骤S3中相似度D_{jw}计算公式可表示为：

$$D_{jw}=\begin{cases}\frac{1}{3}\left(\frac{m}{L(s_1)}+\frac{m}{L(s_2)}+\frac{(m-t)}{m}\right), & m>0\\ 0 & m=0\end{cases} \qquad (4)$$

式(4)中，t表示字符串s_1或字符串s_2中的转置字符数，m表示字符串s_1或字符串s_2在强匹配窗口和弱匹配窗口中所有字符能够匹配的总次数。

7.如权利要求6所述的大规模相似新闻标题快速归一化方法，其特征在于，当所述字符串s_1和字符串s_2之间存在最长连续匹配字符时，所述相似度D_{jw}需要进行微调，可用公式表示：

$$D'_{jw}=D_{jw}+(L*p*(1-D_{jw})) \qquad (5)$$

式(5)中，p表示权重，p=min(0.1,1.0/max(L(s_1),L(s_2)))，L表示字符串s_1和字符串s_2中最长公共字符串的长度，L=max(C[i,j])，其中，C[i,j]表示字符串s_1和字符串s_2中公共字符串的长度，可用公式表示：

$$C[i,j]=\begin{cases}0 & i=0或j=0\\ C[i-1,j,j-1]+1 & x_i=y_j\\ 0 & x_i\neq y_j\end{cases} \qquad (6)$$

式(6)中，i表示字符串s_1中第i个字符索引，j表示字符串s_2中第个j字符索引，x_i表示字符串s_1的第i个字符，j表示字符串s_2的第j个字符。

8.如权利要求7所述的大规模相似新闻标题快速归一化方法，其特征在于，所述步骤S4中的相似度阈值A∈[0.6,1.0]。

9.如权利要求8所述的大规模相似新闻标题快速归一化方法，其特征在于，所述步骤S4

中的相似度阈值A＝0.8。

10.如权利要求9所述的大规模相似新闻标题快速归一化方法,其特征在于,所述强匹配窗口的权重值为1,弱匹配窗口的权重值为0.5。

一种大规模相似新闻标题快速归一化方法

技术领域

[0001] 本发明涉及计算机科学范围的归一化映射技术领域，尤其涉及一种大规模相似新闻标题快速归一化方法。

背景技术

[0002] 新闻标题是一篇新闻的“眼睛”，它能准确概括新闻的主题，随着互联网技术的发展，网页新闻已成为人们生活的一部分，并对人们的信息获取产生不可预估的影响。网络新闻由新闻标题、正文内容、发布时间、来源、作者、编辑等主体部分构成。随着网页新闻应用的推广与深化，大量的新闻数据得到积累，这些新闻数据不论是分析还是管理都面临着巨大的压力，需要借助计算机智能分析技术从中进行深度挖掘从而为相关决策提供有力的支持，这种分析具有重要的价值和意义。特别的，相似新闻的挖掘和分析具有很好地应用场景，例如事件的聚焦和相似新闻的快速聚合等热门应用。

[0003] 从海量的网页新闻标题中分析出相似的新闻标题是一种亟待解决的热点需求，它可以有效的将相似新闻聚集到一起，从而达到对相似新闻数据进行关联分析和挖掘的目的。经过对相似新闻标题的归一化，可以快速将相似的新闻归档到同一个类别，这样可以有效的聚焦相似新闻。对于归档后的新闻数据，再根据新闻发布的时间轴进行升序拼接与组织，这样能让事件相关的新闻串联起来，从而清晰的揭示事件的详细发展轨迹。

[0004] 目前，对于大规模的相似新闻快速归一化处理求来说，业界缺乏成熟的技术支撑，尤其是面临大规模新闻数据分析任务时，问题尤为突出，而且，面对大规模相似新闻标题归一化任务，如果采用现有的字符串相似度算法直接计算任意两个新闻标题之间的相似度，这种计算效率非常低，根本就无法满足真实的需求。

[0005] 中国专利CN201110137785公开了一种分布式实时搜索引擎。本发明的分布式实时搜索引擎，其系统构建和运行至少包括以下步骤:A.设计系统的功能性结构，B.设计系统的数据索引结构，C.索引的创建，D.索引的更新，E.索引的检索。本发明的分布式实时搜索引擎能够在系统的内存中同时构建更新时索引和合并时索引，索引检索时通过同时访问更新时索引和合并时索引，当更新时索引的文档数量积累到阈值后，更新索引提交到磁盘索引并变更为合并时索引，原有的合并时索引变更为新的更新时索引，保证了正在更新中的数据也能够被检索到，提高了搜索引擎可检索数据的实时性。本发明是从数据整体出发，完成数据的索引与检索，而没有关注数据中是否存在相似的情况，并且所采用的检索方法一般都是普通的检索排序类，具有无法判定相似性的缺点。

[0006] 中国专利CN201410323334公开了一种文本相似度计算方法及装置。该文本相似度计算方法包括:通过比较两个文本的节点，计算两个文本的增删距离与替换距离，其中所述增删距离与所述替换距离的和为所述两个文本的编辑距离;根据增删距离与替换距离，计算两个文本的相似度。本发明能够采用一种不依赖于词典、切词以及模型训练的算法，来实现文本之间的相似度的计算，从而可以提高相似度的计算速度。但是本发明仅仅关注了文本的相似性计算，而不能完成归一化处理。

发明内容

[0007] 本发明的目的是提供一种能够给相似新闻标题快速映射成唯一归一化序列码的方法，所述方法主要是针对大规模相似新闻标题的快速归一化处理需求，通过集成使用索引检索技术和相似度计算方法以及MD5值计算方法实现了潜在相似新闻标题的快速查询与归一化，实现了大规模相似新闻标题的快速归一化任务。

[0008] 为解决上述技术问题，本发明提供一种大规模相似新闻标题快速归一化方法，所述方法包括以下步骤：

[0009] S1、将收集的所有新闻标题数据进行预处理；

[0010] S2、任意给定一个新闻标题T，并采用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据，然后将相关的新闻标题数据进行排序；

[0011] S3、通过相似度计算方法分别计算所述步骤S2中检索的相关新闻标题数据与新闻标题T之间的相似度并降序排序，然后返回至最大相似度的新闻标题；

[0012] S4、设定相似度阈值A并将所述步骤S3中最大相似度值与相似度阈值A进行比较，从而判断索引中是否存在与该新闻标题T相似的新闻标题，若索引中存在相似新闻标题，则将最大相似度新闻标题的归一化序列码赋值给该新闻标题T并作为该新闻标题T的归一化序列码，完成该新闻标题的归一化映射，若索引中不存在相似新闻标题，则进入步骤S5；

[0013] S5、计算该新闻标题T的MD5值，并将该新闻标题T的MD5值作为归一化序列码与该新闻标题T的内容存入索引中；

[0014] S6、重复步骤S2-S5，直至完成所有新闻标题数据的归一化映射。

[0015] 优选地，所述步骤S1中将收集的所有新闻标题数据进行预处理的具体实现方式为：去除新闻标题中的一些多余的空格、特殊字符和换行符等。

[0016] 优选地，所述步骤S2中分布式索引为Elasticsearch分布式索引。

[0017] 优选地，所述步骤S2中利用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据，然后将相关的新闻标题数据进行排序的具体实现方式包括：

[0018] S21、对输入的新闻标题T进行分词，并利用停用词典过滤相应的停用词，得到有效词语列表；

[0019] S22、将所述步骤S21中得到的有效词语列表作为真实的查询词语输入；

[0020] S23、计算所述步骤S22中查询词语与索引文档之间的检索得分Score(q,d)，可用公式表示：

[0021] $$Score(q,d)=\sum_{k}^{n} W_k * R(q_k,d) \qquad (1)$$

[0022] 式(1)中，q表示查询词语，d表示索引文档，W_k表示查询词语q中第k个有效词语的逆文档频率，q_k表示查询词语q中第k个有效词语，n表示查询词语q中有效词语的总数，$R(q_k,d)$表示查询词语q中第k个有效词语与索引文档d之间的相关性；

[0023] 其中，$W_k=log\left(\frac{N+1}{n(q_k)+1}\right)$，N表示索引文档的总数量，$n(q_i)$表示包含查询词语$q_k$的索引文档数量，1表示调节因子；$R(q_k,d)=\frac{f_k(h_1+1)}{f_k+H}*\frac{qf_k(h_2+1)}{qf_k+h_2}$，$h_1$和$h_2$表示调整系数，$f_k$表示查询词语$q_k$在索引文档d中的频率，$qf_k$表示查询词语$q_k$在查询词语q中的频率，H表示比例

系数，$H=h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)$，其中b为调节系数，dl表示当前从索引中取出来与有效检索词语相关的新闻标题的长度，avg(dl)表示从索引中检索出来与当前有效检索词语相关的全部新闻标题的平均长度，从而式(1)可表示为：

[0024] $$Score(q,d)=\sum_k^n log\left(\frac{N+1}{n(q_k)+1}\right)*\frac{f_k(h_1+1)}{f_k+h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)}*\frac{qf_k(h_2+1)}{qf_k+h_2};$$

[0025] S24、根据所述步骤S23计算出来的检索得分Score(q,d)对与新闻标题T相关的新闻标题数据进行排序。

[0026] 优选地，所述步骤S3中相似度计算方法为改进型Jaro-Winkler短文本相似度计算方法，所述方法的匹配窗口包括强匹配窗口和弱匹配窗口，所述强匹配窗口和弱匹配窗口的值可用公式表示：

[0027] $$SMW=\frac{max(L(s_1),\ L(s_2))}{2}-1 \qquad (2)$$

[0028] WMW＝max(L(s_1),L(s_2))-index (3)

[0029] 式(2)、(3)中，SMW表示强匹配窗口的值，WMW表示弱匹配窗口的值，s_1，s_2表示字符串，L(s_1)表示字符串s_1的长度，L(s_2)表示字符串s_2的长度，index表示当前强匹配窗口结束位置的值。

[0030] 优选地，所述步骤S3中相似度D_{jw}计算公式可表示为：

[0031] $$D_{jw}=\begin{cases}\frac{1}{3}\left(\frac{m}{L(s_1)}+\frac{m}{L(s_2)}+\frac{(m-t)}{m}\right), & m>0\\ 0 & m=0\end{cases} \qquad (4)$$

[0032] 式(4)中，t表示字符串s_1或字符串s_2中的转置字符数，m表示字符串s_1或字符串s_2在强匹配窗口和弱匹配窗口中所有字符能够匹配的总次数；

[0033] 优选地，当所述字符串s_1和字符串s_2之间存在最长连续匹配字符时，所述相似度D_{jw}需要进行微调，可用公式表示：

[0034] $D'_{jw}=D_{jw}+(L*p*(1-D_{jw}))$ (5)

[0035] 式(5)中，p表示权重，p＝min(0.1,1.0/max(L(s_1),L(s_2)))，L表示字符串s_1和字符串s_2中最长公共字符串的长度，L＝max(C[i,j])，其中，C[i,j]表示字符串s_1和字符串s_2中公共字符串的长度，可用公式表示：

[0036] $$C[i,j]=\begin{cases}0 & i=0或j=0\\ C[i-1,j,j-1]+1 & x_i=y_j\\ 0 & x_i\neq y_j\end{cases} \qquad (6)$$

[0037] 式(6)中，i表示字符串s_1中第i个字符索引，j表示字符串s_2中第个j字符索引，x_i表示字符串s_1的第i个字符，j表示字符串s_2的第j个字符。

[0038] 优选地，所述步骤S4中的相似度阈值A∈[0.6,1.0]。

[0039] 优选地，所述相似度阈值A＝0.8。

[0040] 优选地，所述强匹配窗口的权重值为1，弱匹配窗口的权重值为0.5。

[0041] 与现有技术比较，本发明一种大规模相似新闻标题快速归一化方法，采用分布式索引方法建立待分析新闻标题数据的索引结构并快速检索出相关的新闻标题，为潜在相似新闻标题的快速查找提供了解决方案，然后利用改进型相似度计算方法和新闻标题数据MD5值的计算方法实现了潜在相似新闻标题的快速查询与归一化方法，通过本发明的归一化方法可以找出相关的新闻标题数据并做出相似性判定，然后将相似新闻标题数据映射成唯一的归一化序列码，突破了现有技术的瓶颈，实现了大规模相似新闻标题的快速归一化。

附图说明

[0042] 图1是本发明一种大规模相似新闻标题快速归一化方法流程图，

[0043] 图2是本发明中所述新闻标题的分布式索引方法流程图。

具体实施方式

[0044] 为了使本技术领域的人员更好地理解本发明的技术方案，下面结合附图对本发明作进一步的详细说明。

[0045] 参见图1，图1是本发明提供的一种大规模相似新闻标题快速归一化方法流程图。

[0046] 一种大规模相似新闻标题快速归一化方法，所述方法包括以下步骤：

[0047] S1、将收集的所有新闻标题数据进行预处理；

[0048] S2、任意给定一个新闻标题T，并采用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据，然后将相关的新闻标题数据进行排序；

[0049] S3、通过相似度计算方法分别计算所述步骤S2中检索的相关新闻标题数据与新闻标题T之间的相似度并降序排序，然后返回至最大相似度的新闻标题；

[0050] S4、设定相似度阈值A并将所述步骤S3中最大相似度值与相似度阈值A进行比较，从而判断索引中是否存在与该新闻标题T相似的新闻标题，若索引中存在相似新闻标题，则将最大相似度新闻标题的归一化序列码赋值给该新闻标题T并作为该新闻标题T的归一化序列码，完成该相似新闻标题的归一化映射，若索引中不存在相似新闻标题，则进入步骤S5；

[0051] S5、计算该新闻标题T的MD5(信息摘要，Message Digest)值，并将该新闻标题T的MD5值作为归一化序列码与新闻标题T的内容存入索引中；

[0052] S6、重复步骤S2-S5，直至完成所有新闻标题数据的归一化映射。

[0053] 本实施例中，通过采用分布式索引方法建立待分析新闻标题数据的索引结构并快速检索出相关的新闻标题，为潜在相似新闻标题的快速查找提供了解决方案，然后利用改进型相似度计算方法和新闻标题数据MD5值的计算方法实现了潜在相似新闻标题的快速查询与归一化方法，所述归一化方法可以准确找出相关的新闻标题并做出相似性判定，然后将相似新闻标题数据映射成唯一的归一化序列码，突破了现有技术的瓶颈，实现了大规模相似新闻标题的快速归一化

[0054] 本实施例中，当所述步骤S2中任意给定的新闻标题为第一条分析的新闻标题时，此时索引数据为空且该新闻标题没有相应的归一化序列码，则索引数据中检索出来的结果为空，即不存在与该条新闻标题相似的索引记录，直接进入步骤S5计算该新闻标题的MD5值，并将该新闻标题T的MD5值作为归一化序列码与新闻标题T的内容存入索引中，然后进行

下一条新闻标题的分析，随着新闻标题数据的实时检索不断进行，该索引数据也会实时发生变化，同时也使得索引数据中所有新闻标题均不相似。

[0055] 如图1所示，所述步骤S1中将收集的所有新闻标题数据进行预处理的具体实现方式为：去除新闻标题中的一些多余的空格、特殊字符和换行符等。本实施例中，通过对新闻标题进行分析前的预处理，从而有效提高了新闻标题数据后期的处理效率。

[0056] 如图1所示，所述步骤S2中分布式索引为Elasticsearch分布式索引。本实施例中，采用Elasticsearch(弹性检索)分布式索引来处理新闻标题数据。在其他实施例中，也可以采用Lucene或者solr的倒排索引机制来构建所需要的分布式索引。

[0057] 如图2所示，所述步骤S2中利用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据，然后将相关的新闻标题数据进行排序的具体实现方式包括：

[0058] S21、对输入的新闻标题T进行分词，并利用停用词典过滤相应的停用词，得到有效词语列表；

[0059] S22、将所述步骤S21中得到的有效词语列表作为真实的查询词语输入；

[0060] S23、计算所述步骤S22中查询词语与索引文档之间的检索得分Score(q,d)，可用公式表示：

[0061] $$Score(q,d)=\sum_{k}^{n}W_k*R(q_k,d) \qquad (1)$$

[0062] 式(1)中，q表示查询词语，d表示索引文档，W_k表示查询词语q中第k个有效词语的逆文档频率，q_k表示查询词语q中第k个有效词语，n表示查询词语q中有效词语的总数，$R(q_k,d)$表示查询词语q中第k个有效词语与索引文档d之间的相关性；

[0063] 其中，$W_k=log\left(\frac{N+1}{n(q_k)+1}\right)$，N表示索引文档的总数量，$n(q_i)$表示包含查询词语$q_k$的索引文档数量，1表示调节因子；$R(q_k,d)=\frac{f_k(h_1+1)}{f_k+H}*\frac{qf_k(h_2+1)}{qf_k+h_2}$，$h_1$和$h_2$表示调整系数，$f_k$表示查询词语$q_k$在索引文档d中的频率，$qf_k$表示查询词语$q_k$在查询词语q中的频率，H表示比例系数，$H=h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)$，其中b为调节系数，dl表示当前从索引中取出来与有效检索词语相关的新闻标题的长度，avg(dl)表示从索引中检索出来与当前有效检索词语相关的全部新闻标题的平均长度，从而式(1)可表示为：

[0064] $$Score(q,d)=\sum_{k}^{n}log\left(\frac{N+1}{n(q_k)+1}\right)*\frac{f_k(h_1+1)}{f_k+h_1*\left(1-b+b*\frac{dl}{avg(dl)}\right)}*\frac{qf_k(h_2+1)}{qf_k+h_2};$$

[0065] S24、根据所述步骤S23计算出来的检索得分Score(q,d)对与新闻标题T相关的新闻标题数据进行排序。

[0066] 本实施例中，调节因子1的作用是为了防止分母为零，而计算查询词语q中第k个有效词语与索引文档d之间的相关性公式中的调整系数h_1和h_2设为1，查询词语q中第k个有效词语的逆文档频率计算公式中的调节系数b取值为0.75。若在其他实施例中，对于输入的新闻标题没有进过前期的预处理，则在计算查询词语与与索引文档之间的检索得分时，需要增加一个用来表示当前检索词语有效性的指示函数，若当前检索词语为有效检索词，则该

有效性指示函数取值为1,若当前检索词语为无效检索词,则该有效性指示函数取值为0,从而起到调节作用;由于本实施例中对输入的新闻标题均做了预处理,故该有效性指示函数取值均为1,即在计算查询词语与与索引文档之间的检索得分时去掉了该有效性指示函数。

[0067] 如图1所示,所述步骤S3中相似度计算方法为改进型Jaro-Winkler短文本相似度计算方法,所述方法的匹配窗口包括强匹配窗口和弱匹配窗口,所述强匹配窗口和弱匹配窗口的值可用公式表示:

[0068] $$SMW=\frac{max(L(s_1),\ L(s_2))}{2}-1 \qquad (2)$$

[0069] WMW=max(L(s_1),L(s_2))-index (3)

[0070] 式(2)、(3)中,SMW表示强匹配窗口的值,WMW表示弱匹配窗口的值,s_1,s_2表示字符串,L(s_1)表示字符串s_1的长度,L(s_2)表示字符串s_2的长度,index表示当前强匹配窗口结束位置的值。

[0071] 本实施例中,通过基于字符改进的Jaro-Winkler短文本相似度计算方法来作为新闻标题之间的相似度判定方法,同时考虑到该方法中匹配窗口大小对相似度判断的影响,为保证待分析的新闻标题数据之间相似度判断的准确性,通过设置强匹配窗口和弱匹配窗口的分层匹配方法来计算并判定其相似度。在其他实施例中,也可以使用基于字符特征相似度的余弦相似度计算方法来实现。

[0072] 如图1所示,所述步骤S3中相似度D_{jw}计算公式可表示为:

[0073] $$D_{jw}=\begin{cases}\frac{1}{3}\left(\frac{m}{L(s_1)}+\frac{m}{L(s_2)}+\frac{(m-t)}{m}\right), & m>0\\ 0 & m=0\end{cases} \qquad (4)$$

[0074] 式(4)中,t表示字符串s_1或字符串s_2中的转置字符数,m表示字符串s_1或字符串s_2在强匹配窗口和弱匹配窗口中所有字符能够匹配的总次数;

[0075] 如图1所示,当所述字符串s_1和字符串s_2之间存在最长连续匹配字符时,所述相似度D_{jw}需要进行微调,可用公式表示:

[0076] $D'_{jw}=D_{jw}+(L*p*(1-D_{jw}))$ (5)

[0077] 式(5)中,p表示权重,p=min(0.1,1.0/max(L(s_1),L(s_2))),L表示字符串s_1和字符串s_2中最长公共字符串的长度,L=max(C[i,j]),其中,C[i,j]表示字符串s_1和字符串s_2中公共字符串的长度,可用公式表示:

[0078] $$C[i,j]=\begin{cases}0 & i=0\text{或}j=0\\ C[i-1,j,j-1]+1 & x_i=y_j\\ 0 & x_i\neq y_j\end{cases} \qquad (6)$$

[0079] 式(6)中,i表示字符串s_1中第i个字符索引,j表示字符串s_2中第个j字符索引,x_i表示字符串s_1的第i个字符,j表示字符串s_2的第j个字符。

[0080] 本实施例中,当所述字符串s_1和字符串s_2之间存在最长连续匹配字符时,通过对Jaro-Winkler方法所计算出来的相似度值进行微调,从而提高新闻标题数据之间相似度判定的准确度。

[0081] 如图1所示,所述步骤S4中的相似度阈值A∈[0.6,1.0]。

[0082] 如图1所示,所述步骤S4中的相似度阈值A=0.8。

[0083] 本实施例中，所述相似度阈值A可以根据不同需求进行自行设置，其取值范围设为[0.6,1.0]，更进一步的，相似度阈值A设为0.8，当完成相关新闻标题数据与新闻标题T之间的相似度计算并降序排序后，通过判断排序中最大相似度值是否小于0.8，如果小于0.8，则认为现有索引数据中不存在与当前新闻标题相似的新闻标题，此时进入步骤S5计算当前输入的新闻标题的MD5值，并将该新闻标题的MD5值作为归一化序列码与该新闻标题的内容存入到索引数据中，作为下一个输入新闻标题的比对目标；如果最大相似度值大于等于0.8，则认为当前输入的新闻标题在索引中存在与其相似的新闻标题，此时直接将索引中最大相似度标题所对应的归一化序列码赋值给当前输入的新闻标题，并作为当前输入的新闻标题的归一化序列码，继续输入下一条新闻标题的分析，从而实现了大规模相似新闻标题的快速归一化。

[0084] 优选地，所述强匹配窗口的权重值为1，弱匹配窗口的权重值为0.5。本实施例中，当两个字符距离小于匹配窗口值时，则认为两个字符匹配，若字符位于强匹配窗口，当两个字符距离小于强匹配窗口值1时，可认为这两个字符匹配；若字符位于弱匹配窗口，当两个字符距离小于弱匹配窗口值0.5时，即也可认为这两个字符匹配。

[0085] 以上对本发明所提供的一种大规模相似新闻标题快速归一化方法进行了详细介绍。本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的核心思想。应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以对本发明进行若干改进和修饰，这些改进和修饰也落入本发明权利要求的保护范围内。

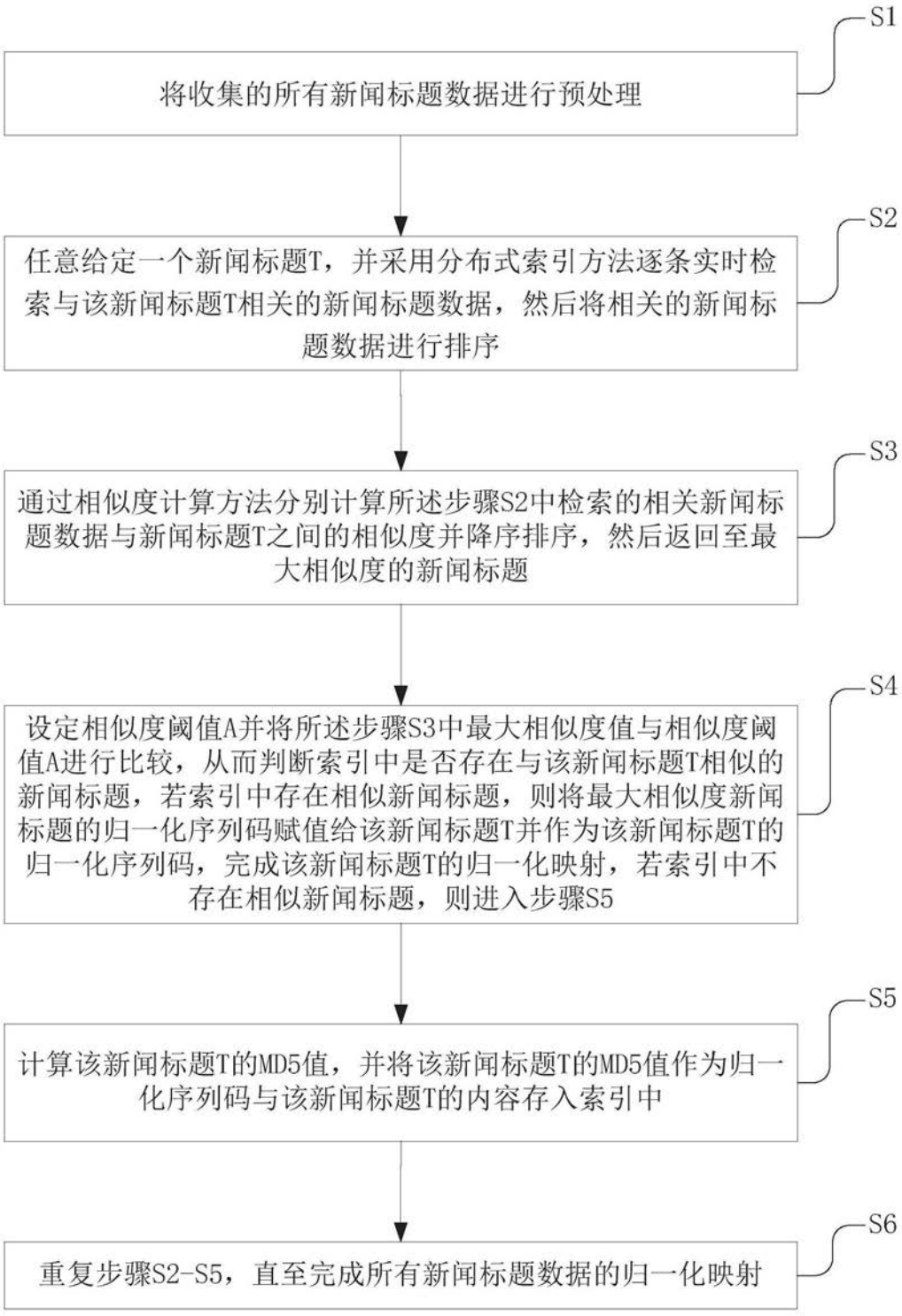

S1
将收集的所有新闻标题数据进行预处理
S2
任意给定一个新闻标题T，并采用分布式索引方法逐条实时检索与该新闻标题T相关的新闻标题数据，然后将相关的新闻标题数据进行排序
S3
通过相似度计算方法分别计算所述步骤S2中检索的相关新闻标题数据与新闻标题T之间的相似度并降序排序，然后返回至最大相似度的新闻标题
S4
设定相似度阈值A并将所述步骤S3中最大相似度值与相似度阈值A进行比较，从而判断索引中是否存在与该新闻标题T相似的新闻标题，若索引中存在相似新闻标题，则将最大相似度新闻标题的归一化序列码赋值给该新闻标题T并作为该新闻标题T的归一化序列码，完成该新闻标题T的归一化映射，若索引中不存在相似新闻标题，则进入步骤S5
S5
计算该新闻标题T的MD5值，并将该新闻标题T的MD5值作为归一化序列码与该新闻标题T的内容存入索引中
S6
重复步骤S2-S5，直至完成所有新闻标题数据的归一化映射

图1

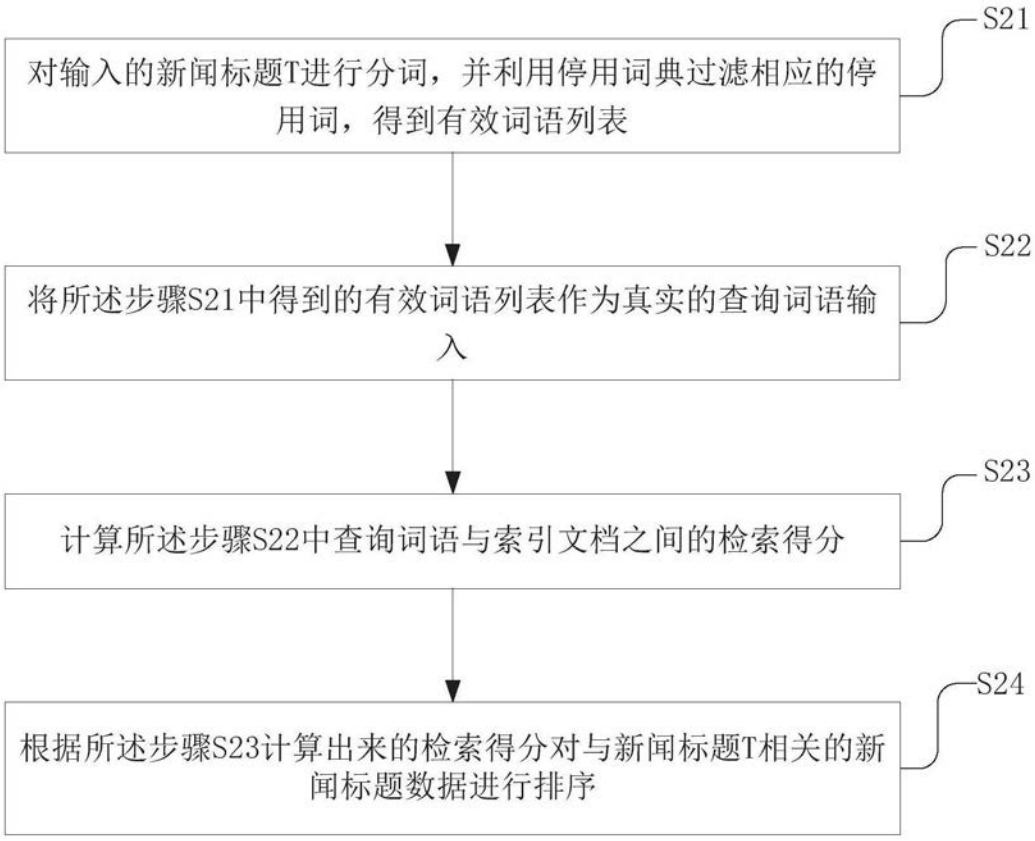

S21
对输入的新闻标题T进行分词，并利用停用词典过滤相应的停用词，得到有效词语列表
S22
将所述步骤S21中得到的有效词语列表作为真实的查询词语输入
S23
计算所述步骤S22中查询词语与索引文档之间的检索得分
S24
根据所述步骤S23计算出来的检索得分对与新闻标题T相关的新闻标题数据进行排序

图2