本发明公开了一种并行化的文本聚类方法，其中，所述方法包括：将非结构化文本平均分发到各分布式节点上，对各节点上的文本进行预处理、分词、过滤停用词操作，采用并行化的计算方式对处理结果进行特征提取，获取处理后的文本特征向量；采用改进后的并行化聚类方法对上述文本特征向量进行聚类处理，增量式地获取多个文本簇；通过将聚类过程中的各个步骤并行化，在面对海量或高维数据时，提升了文本聚类的速度。



1.一种并行化的文本聚类方法，其特征在于，所述方法包括：

1.1将非结构化文本平均分发到各分布式节点上，对各节点上的文本进行预处理、分词、过滤停用词操作；

1.2采用并行化的计算方式对处理结果进行特征提取，获取处理后的文本特征向量；

1.3采用改进后的并行化聚类方法对上述文本特征向量进行聚类处理，增量式地获取多个文本簇。

2.如权利要求1所述的方法，其特征在于，所述将非结构化文本平均分发到各分布式节点上，对各节点上的文本进行预处理、分词、过滤停用词操作，包括：

2.1采用“key=文本编号，value=文本内容”的格式，将已有非结构化文本平均分发到各分布式节点上；

2.2对各分布式节点上的非结构化文本进行统一格式处理，去除文本首尾非文本部分，获取纯文本部分，若为空文本则跳过；

2.3将所述纯文本部分进行分词处理，针对词语词性，去除分词结果中的标点符号、拟声词、叹词、助词、连词、介词、副词、数词、量词；

3.如权利要求1所述的方法，其特征在于，所述采用并行化的计算方式对处理结果进行特征提取，获取处理后的文本特征向量，包括：

3.1采用并行化计算方式对各分布式节点的分词结果进行处理，获取所有文本的词频向量，具体步骤包括：

为每个文本构建一个维度足够大（如，该值可根据文本数量大小进行估计）的词频向量，向量长度为length，求词语的hash值，对length取模，得到该词语映射到该向量上的索引；对文本中的每个词语在上其对应索引位置进行加1操作，将该向量转化为一个稀疏的向量（记录非零元素的索引及其值），这样求出所有文本的词频向量；

3.2采用并行化计算方式对各分布式节点的词频向量进行处理，获取所有文本的逆文本频率向量，将该向量广播到各分布式节点上，具体步骤包括：

为该分布式节点构建词语出现的文本频率向量，维度与一致，向量的每个元素表示词语在该节点多少个文本中出现过，遍历节点中每个文本的词频向量，循环取出词频向量中非零元素（表示该词在该文本中出现了）的位置，在的对应索引位置进行加1操作；将各分布式节点的向量的对应成员值聚合相加得到总文本频率向量；进而获取所有文本的逆文本频率向量，将该向量广播到各分布式节点上；

3.3计算各分布式节点上词语的值，获取文本特征向量，具体步骤包括：

将文本对应的词频向量和逆文本频率向量相乘得到每个文本的向量，按照“key=文本编号，value=向量”的格式聚合所有节点上的向量，得到总向量；

4. 如权利要求1所述的方法，其特征在于，所述采用改进后的并行化聚类方法对上述文本特征向量进行聚类处理，增量式地获取多个文本簇，包括：

4.1将总向量广播到各分布式节点上，计算每个文本与总向量中该文本之前的所有文本的余弦相似度，具体步骤包括：

计算每个文本与总向量中该文本之前的所有文本的余弦相似度，从这个余弦相似度中取出最大值，即第个文本与前个文本的余弦相似度中的最大值；

4.2创建共享向量Data，维度与文本数量一致，以存放聚类结果；

4.3根据上述余弦相似度对所述文本进行改进后的并行化聚类，增量式地获取多个文本簇，具体步骤包括：

当或小于设定的阈值时，为第个文本新建一个文本簇， 以“key=文本编号，group=”的格式在Data索引为的位置存放数据；

当大于设定的阈值时，将第个文本与第个文本归为同一文本簇， 在Data获取文本的group值G，以“key=文本编号，group=G”的格式在Data索引为的位置存放数据；

最后得到的Data向量即聚类结果，group一致的文本被聚为同一文本簇。

**一种并行化的文本聚类方法**

**技术领域**

本文涉及计算机技术领域，尤其涉及一种并行化的文本聚类方法。

**背景技术**

随着信息网络技术的迅速发展和互联网的进一步普及，网络上的数据呈现几何式的增长，数据“爆炸”已成为当前网络时代的特征之一。面对如此庞大而且增长迅速的数据，高效地挖掘有用信息无论在商业、医疗还是科学研究方面，都有着非常巨大的价值。其中，大量信息都以文本形式存储，如新闻稿件、科技论文、书籍、数字图书馆、邮件、博客和网页等等。文本聚类技术可以将大量文本聚合为少数有意义的簇，从而在大量文本中导出高质量的信息，使得人们从数据中获取信息、知识和决策支持更加容易。

但是，传统串行式的文本聚类方法在处理海量或者高维数据时，聚类的速度不够快，在面对大规模数据时，受制于内存容量，往往不能有效地运行，因而传统串行式文本聚类方法已经难以满足当前实际应用的需求。

并行计算（Parallel Computing）是指同时使用多种计算资源解决计算问题的过程，是提高计算机系统计算速度和处理能力的一种有效手段。它的基本思想是用多个处理器来协同求解同一问题，即将被求解的问题分解成若干个部分，各部分均由一个独立的处理机来并行计算。并行计算系统既可以是专门设计的、含有多个处理器的超级计算机，也可以是以某种方式互连的若干台的独立计算机构成的集群。通过并行计算集群完成数据的处理，再将处理的结果返回给用户。

并行计算可以将大规模数据分发到多个分布式节点上并行地进行计算，最后将所有节点的计算结果归并为最终的结果，可以大大地提高计算速度。然而实际上，可能存在许多障碍使得特定类型的计算任务难以执行给定的并行化处理。通常来说，对于一项将进行并行处理的计算任务来说，需要将与该计算任务相关的数据复制到其使用的每一个处理器中，这里将产生一定的计算开销；若某项任务需要计算资源整合其它子任务的并行处理结果，以便得到一个统一的计算结果时，可能产生一定的整合资源计算开销，考虑到这些开销，许多类型的计算任务进行并行化是不现实的。为了将传统串行式的文本聚类方法并行化，需要对其进行方法上的改进，使其适合并行化计算的基本体系结构，这样才能高效地利用计算资源，在处理海量或者高维数据时，大大提高文本聚类的速度。

**发明内容**

本发明的目的在于克服现有技术的不足，提供一种并行化的文本聚类方法，充分利用并行计算的优点，提高文本聚类的速度，所述方法包括：

采用“key=文本编号，value=文本内容”的格式，将已有非结构化文本平均分发到各分布式节点上；

对各节点上的非结构化文本进行预处理、分词、过滤停用词操作，得到每个文本的分词结果；

基于向量空间模型对每个文本的分词结果进行特征提取，假设一个由n个文本构成的文本集合为，第个文本的词语构成为，m代表该文本中所有词语的个数，那么每一个文本表示为，其中为第个文本中词语的权重，该权重使用计算；其中词语在文本中的词频，表示第个文本中词语的个数；词语的逆文本频率指数，表示所有文本数目，表示词语所在文本个数；然后得到词语的权值；所需数据计算方法包括：

采用并行化计算方式对各分布式节点的分词结果进行处理，具体步骤包括：为每个文本构建一个维度足够大（如，该值可根据文本数量大小进行估计）的词频向量，向量长度为length，求词语的hash值，对length取模，得到该词语映射到该向量上的索引；对文本中的每个词语在上其对应索引位置进行加1操作，将该向量转化为一个稀疏的向量（记录非零元素的索引及其值），这样求出所有文本的词频向量；

采用并行化计算方式对各分布式节点的词频向量进行处理，具体步骤包括：为该分布式节点构建词语出现的文本频率向量，维度与一致，向量的每个元素表示词语在该节点多少个文本中出现过，遍历节点中每个文本的词频向量，循环取出词频向量中非零元素（表示该词在该文本中出现了）的位置，在的对应索引位置进行加1操作；将各分布式节点的向量的对应成员值聚合相加得到总文本频率向量；设总文档数为n，词语在中的文本频率指数为，则词语的逆文本频率指数，获取所有文本的逆文本频率向量，将该向量广播到各分布式节点上；

在各分布式节点上，将文本对应的词频向量和逆文本频率向量相乘得到每个文本的向量，按照“key=文本编号，value=向量”的格式聚合所有节点上的向量，得到总向量，再将其广播到各节点上；现进行对各文本的聚类，方法如下：

计算每个文本与总向量中该文本之前的所有文本的余弦相似度，即，其中，表示第个文本的向量，从这个余弦相似度中取出最大值，即第个文本与前个文本的余弦相似度中的最大值；创建共享向量Data，维度与文本数量一致，以存放聚类结果；

当或小于设定的阈值时，为第个文本新建一个文本簇， 以“key=文本编号，group=”的格式在Data索引为的位置存放数据；

当大于设定的阈值时，将第个文本与第个文本归为同一文本簇， 在Data获取文本的group值G，以“key=文本编号，group=G”的格式在Data索引为的位置存放数据；

最后得到的Data向量即聚类结果，group一致的文本被聚为同一文本簇。

**附图说明**

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它的附图。

图1是本发明实施例的一种并行化的文本聚类方法的系统框图；

图2是本发明实施例的文本预处理、分词、过滤停用词操作步骤的流程示意图；

图3是本发明实施例的获取文本特征向量步骤的流程示意图；

图4是本发明实施例的改进的并行化文本聚类方法步骤的流程示意图。

**具体实施方式**

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，所描述的实施例仅仅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作出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

图1是本发明实施例的一种并行化的文本聚类方法的系统框图，如图1所示，该方法包括：

首先将主节点的文本平均分发到各个从节点上以进行并行化处理，接下来讲述的大部分操作步骤都为某单节点上数据的操作步骤，但各节点上操作相同，可以完全复制如下操作步骤；主节点可以聚合分节点的数据，同时也可以向分节点分发数据，以下步骤中S27、S28和S3为主节点操作，其余步骤皆在各从节点上并行化操作；

步骤S1：对各节点上的文本进行预处理、分词、过滤停用词操作，得到分词数据；

步骤S2：对各节点上的分词数据进行特征提取，获取处理后的文本特征向量；

步骤S3：将各节点上的文本特征向量进行聚合得到所有文本的特征向量，再将其广播到各节点上；

步骤S4：在各节点上，采用改进后的并行化聚类方法对上述文本特征向量进行聚类处理，增量式地获取多个文本簇；

对步骤S1作进一步说明：

采用“key=文本编号，value=文本内容”的格式，预先就将非结构化文本平均分发到各节点，之后的大部分操作都将在各节点完成，以提升操作完成速度；数据库中文本可能存在首尾部分有冗余内容或者文本自身为空的情况，需要先进行一步预处理，再对纯文本进行分词操作，获取分词结果，其中包括词语词性，过滤掉部分词性的词语。

进一步地，图2是本发明实施例的文本预处理、分词、过滤停用词操作步骤的流程示意图，如图2所示，该步骤包括：

步骤S11：对各节点上的非结构化文本进行统一格式处理，去除文本首尾非文本部分，获取纯文本部分，若为空文本则跳过；

步骤S12：将所述纯文本部分进行分词处理；

步骤S13：针对词语词性，去除分词结果中的标点符号、拟声词、叹词、助词、连词、介词、副词、数词、量词。

对步骤S2作进一步说明：

采用并行化计算方式对各节点的分词结果进行处理，获取所有文本的词频向量；采用并行化计算方式对各节点的词频向量进行处理，获取所有文本的逆文本频率向量。将文本对应的词频向量和逆文本频率向量相乘得到每个文本的向量， 即各文本的特征向量。

进一步地，图3是本发明实施例的获取文本特征向量步骤的流程示意图，如图3所示，该步骤包括：

步骤S21：为每个文本构建一个维度足够大的词频向量，维度可根据文本数量大小进行估计，向量长度为length；

步骤S22：求文本中每个词语的hash值，对length取模，得到该词语映射到该向量上的索引；

步骤S23：对文本中的每个词语在其对应索引位置进行加1操作；

步骤S24：记录非零元素的索引及其值，将该向量转化为一个稀疏的向量，获取所有文本的词频向量；

步骤S25：为各节点构建词语出现的文本频率向量，维度与词频向量一致；

步骤S26：遍历节点中每个文本的词频向量，循环取出词频向量中非零元素的位置，在文本频率向量对应索引位置进行加1操作；

步骤S27：将各节点的文本频率向量的对应成员值聚合相加得到总文本频率向量；

步骤S28：对总文本频率向量进行计算得到所有文本的逆文本频率向量，将总文本频率向量分发到各节点上；

步骤S29：将文本对应的词频向量和逆文本频率向量相乘得到每个文本的向量，即文本特征向量。

对步骤S21作进一步说明：

词频向量的维度应足够大，来保证步骤S22中词语的索引不会频繁出现冲突，如，该值可根据文本数量进行设置。

对步骤S23作进一步说明：

该步骤统计了该文本中各词语出现的次数，便于接下来计算词频。

对步骤S24作进一步说明：

将向量转化为稀疏向量可以降低向量维度，数据格式为“index : value”；词频是指一个词语在文本中出现的频率，j文本中词语i的词频为，其中分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和；文本中所有词语的词频构成该文本的词频向量；通过该公式计算得到所有文本的词频向量。

对步骤S26作进一步说明：

该步骤统计了各词语在多少个文本中出现过，便于接下来结算逆文本频率；S27是对各节点S26操作的统计。

对步骤S28作进一步说明：

逆文本频率是一个词语普遍重要性的度量，某一特定词语的逆文本频率，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到，其中表示所有文本数目，表示词语所在文本个数；如果该词语不在语料库中，就会导致分母为零，因此一般情况下会在分母分子上分别加1得到；文本中所有词语的逆文本频率构成该文本的逆文本频率；通过该公式计算得到所有文本的逆文本频率向量。

对步骤S29作进一步说明：

权值，通过该公式计算得到所有文本的向量。

对步骤S3作进一步说明：

聚合各节点上所有文本的向量得到总向量，将总向量广播到各节点上。

对步骤S4作进一步说明：

计算节点上每个文本与总向量中该文本之前的所有文本的余弦相似度，从这个余弦相似度中取出最大值，即第个文本与前个文本的余弦相似度中的最大值；例如文本集合为，那么对于文本就只需求文本与文本集中各文本的余弦相似度，并从中取出最大值。

创建共享向量Data，维度与文本数量一致，以存放聚类结果；根据上述余弦相似度对所述文本进行改进后的并行化聚类，增量式地获取多个文本簇。

进一步地，图4是本发明实施例的改进的并行化文本聚类方法步骤的流程示意图，如图4所示，该步骤包括：

（1）将总向量广播到各节点上，遍历各节点每个文本；

（2）判断若该文本为总向量的首个文本则跳到步骤（7），若不是则往下执行；

（3）计算每个文本与总向量中该文本之前的所有文本的余弦相似度；

（4）从中取出最大值max；

（5）判断是否Max > Threshold，若是到达步骤（6），若不是则到达步骤（7）；

（6）将得到最大余弦相似度的两个文本归为同一文本簇；

（7）为该文本新建一个文本簇。

对步骤（3）作进一步说明：

余弦相似度计算公式，其中，表示第个文本的向量，进行余弦相似度计算时都应按照总向量的文本排列顺序进行计算。

对步骤（5）作进一步说明：

本实施例中将Threshold设为0.3有较好的聚类结果，该阈值可根据实际情况进行设置。

上述步骤中S27、S28和S3是主节点对从节点数据的聚合和聚合后数据的分发操作，此外其它步骤都将在所有从节点上进行并行化操作。

在本发明实施例中，采用并行化的方式对各节点上的文本进行预处理、分词、过滤停用词、特征提取、聚类处理，增量式地获取多个文本簇；通过将聚类过程中的各个步骤并行化，在面对海量或高维数据时，提升了文本聚类的速度。

上述实施例的各步骤可以由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于随机存取存储器RAM、只读存储器ROM、硬盘、光盘或者本领域熟知的任何其它形式的存储介质中。

尽管上面对本发明说明性的具体实施例进行了描述，以便于本技术领域的 技术人员理解本发明，但应该清楚，本发明不限于具体实施方式的范围，对本技术领域的普通技术人员来讲，只要各种变化在所附的权利要求限定和确定的本发明的精神和范围内，这些变化是显而易见的，一切利用本发明构思的发明创造均在保护之列。



图1



图2



图3



图4