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，可以大大地提高聚类速度。

**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：

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

对各节点上的非结构化文本进行格式化、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TF-IDF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，表示词语所在文档个数；然后得到词语的TF-IDF权值；所需数据计算方法包括：

采用并行化计算方式对各个分布式节点的分词结果进行处理，具体步骤为：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，然后对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，然后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），这样就得到了所有文本的词频向量；

采用并行化计算方式对各个分布式节点的词频向量进行处理，具体步骤为：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，遍历节点中的每个词频向量，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；然后将各个分布式节点的向量的对应成员值聚合相加得到总的文本频率向量；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，这样就得到了所有文本的逆文本频率向量，将该向量广播到各个分布式节点上；

在各个分布式节点上，将文本对应的词频向量和逆文本频率向量向量相乘就得到每个文本的向量，按照“key=文本编号，value=TF-IDF向量”的格式聚合所有节点上的向量，得到总的向量，再将其广播到各个节点上；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类别， 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一类， 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类。

**附图说明**

**具体实施方式**