(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 104123291 A
(43)申请公布日 2014.10.29

(21)申请号 201310148102.3

(22)申请日 2013.04.25

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 臧文阳 齐泉

(74)专利代理机构 深圳中一专利商标事务所 44237
代理人 张全文

(51)Int.Cl.
G06F 17/30(2006.01)

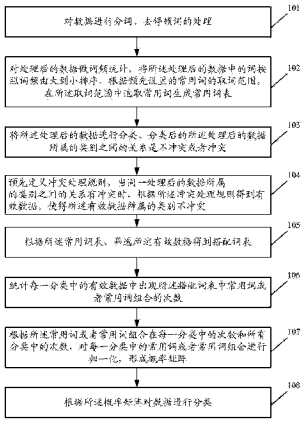
权利要求书4页 说明书17页 附图3页

(54)发明名称

一种数据分类的方法及装置

(57)摘要

本发明实施例公开了一种数据分类的方法及装置，所述方法通过预先定义分类的冲突处理规则，解决分类中有冲突的数据，得到有效数据，从而得到无歧义、无冲突的数据；并使用常用词表，筛选有效数据得到搭配词表，根据所述搭配词表形成概率矩阵，从而避免出现数据稀疏的情况。



CN 104123291 A

1. 一种数据分类的方法，其特征在于，所述方法包括：

对数据进行分词、去停顿词的处理；

对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词和常用词组合；

统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

根据所述概率矩阵对数据进行分类。

2. 根据权利要求 1 所述的方法，其特征在于，所述预先定义冲突处理规则，包括：

当对所述处理后的数据进行分类后，所述处理后的数据同时属于 A 类和 B 类，当 A 类和 B 类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于 A 类和 B 类；

当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 A 类；

或者当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 B 类；

当 A 类和 B 类互相冲突时，且没有冲突处理规则解决所述 A 类和 B 类的冲突时，则标识所述处理后的数据冲突。

3. 根据权利要求 2 所述的方法，其特征在于，所述根据所述冲突处理规则得到有效数据，包括：

当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

对化简后的分类集合，根据所述冲突处理规则中当 A 类和 B 类相互冲突时，且没有冲突处理规则解决 A 类和 B 类的冲突时，标识所述处理后的数据为冲突。

4. 根据权利要求 3 所述的方法，其特征在于，所述根据所述冲突处理规则化简所述两个以上的分类集合，包括：

当所述两个以上的分类集合中 A 类和 B 类不冲突时，将 A 类和 B 类化简为同一类；

当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 A 类为准，将 A 类和 B 类化简为 A 类；

当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 B 类为准，将 A 类和 B 类化简

为 B 类。

5. 根据权利要求 1 至 4 任意一项所述的方法，其特征在于，所述根据所述常用词表，筛选所述有效数据得到搭配词表，包括：

根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于 3 时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

6. 根据权利要求 5 所述的方法，其特征在于，所述统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数，包括：

统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数。

7. 根据权利要求 6 所述的方法，其特征在于，所述根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵，包括：

将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

8. 根据权利要求 1 至 7 所述的方法，其特征在于，所述根据所述概率矩阵对数据进行分类，包括：

在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

将概率最大的列对应的类别作为所述数据的类别。

9. 一种数据分类的装置，其特征在于，所述装置包括：

处理单元，用于对数据进行分词、去停顿词的处理；

第一统计单元，用于对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

第一分类单元，用于将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

解决冲突单元，用于预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

筛选单元，用于根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词或者常用词组合；

第二统计单元，用于统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

归一化单元，用于根据所述常用词或者常用词组合在每一分类中的次数和所有分类中

的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

第二分类单元，用于根据所述概率矩阵对数据进行分类。

10. 根据权利要求 9 所述的装置，其特征在于，所述解决冲突单元中执行步骤预先定义冲突处理规则，包括：

当对所述处理后的数据进行分类后，所述处理后的数据同时属于 A 类和 B 类时，当 A 类和 B 类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于 A 类和 B 类；

当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 A 类；

或者当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 B 类；

当 A 类和 B 类互相冲突时，且没有冲突处理规则解决所述 A 类和 B 类的冲突时，则标识所述处理后的数据冲突。

11. 根据权利要求 10 所述的装置，其特征在于，所述解决冲突单元中执行步骤根据所述冲突处理规则得到有效数据，包括：

当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

对化简后的分类集合，根据所述冲突处理规则中当 A 类和 B 类相互冲突时，且没有冲突处理规则解决 A 类和 B 类的冲突时，标识所述处理后的数据为冲突。

12. 根据权利要求 11 所述的装置，其特征在于，所述解决冲突单元中执行步骤根据所述冲突处理规则化简所述两个以上的分类集合，包括：

当所述两个以上的分类集合中 A 类和 B 类不冲突时，将 A 类和 B 类化简为同一类；

当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 A 类为准，将 A 类和 B 类化简为 A 类；

当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 B 类为准，将 A 类和 B 类化简为 B 类。

13. 根据权利要求 9 至 12 任意一项所述的装置，其特征在于，所述筛选单元具体用于：

根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于 3 时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

14. 根据权利要求 13 所述的装置，其特征在于，所述第二统计单元，具体用于：

统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次

数。

15. 根据权利要求 14 所述的装置，其特征在于，所述归一化单元，包括：

矩阵单元，用于将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

概率矩阵单元，用于根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

16. 根据权利要求 9 至 15 所述的装置，其特征在于，所述第二分类单元，包括：

筛选子单元，用于在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

第二分类子单元，用于将概率最大的列对应的类别作为所述数据的类别。

一种数据分类的方法及装置

技术领域

[0001] 本发明涉及数据分析处理领域，尤其涉及到一种数据分类的方法及装置。

背景技术

[0002] 实际工作中很多记录都是由人工记录的，属于超短文本，其中很多记录可能会出现描述前后不一致的情况。比如，在同一超短文本中，某些字段中写的故障原因是焊接，但是在某些字段又说明是雷击造成故障。如果在这种数据质量不好的情况下进行数据挖掘，会大大降低分析的准确度，所以有必要对数据进行预处理，对数据按照故障原因分为几类，通过分类方法解决问题。

[0003] IFIDF 分类方法的主要思想是如果某个词或者短语在同一超短文本中出现的频率 IF 高，并且在其他超短文本中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。IFIDF 实际是 IF*IDF，IF 是词频(Term Frequency)，IDF 是反文档频率(Inverse Document Frequency)，IF 表示词条在超短文本中出现的频率，IDF 表示词条在本超短文本和其他超短文本的对比结果，当词条在本超短文本出现频率越高，但在其他超短文本出现频率越低时，说明所述词条具有很好的类别区分能力，则所述词条在本超短文本的 IDF 值越大。IFIDF 分类方法的缺点是没有考虑分类和分类之间的关系；分类与分类之间存在交集，对于交集的超短文本没有做特别的处理；超短文本命中的准确率较低；只体现一个词和分类的关系，没有体现多个词搭配出现时和分类的关系。

[0004] N 元文法分类方法的主要思路是词条的概率是由一组特定的词构成的序列决定的，称为所述词条的历史(history)。N 元文法是大词汇连续出现时常用的一种语言模型，该模型基于这样一种假设，第 N 个词的出现只与前面 N-1 个词相关，而与其他任何词都不相关，整句的概率就是各个词出现的概率的乘积，而这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到，常用的是二元文法和三元文法。N 元文法的缺点是当由 4 个以上的词构成序列的情况下，超短文本中数据稀疏非常严重，基本上 N 元文法没办法使用；同时，序列需要重新训练语言模型，由人工标注，工作量比较大。

发明内容

[0005] 本发明提供了一种数据分类的方法及装置，所述方法旨在解决分类时数据冲突及数据稀疏的问题。

[0006] 第一方面，一种数据分类的方法，所述方法包括：

[0007] 对数据进行分词、去停顿词的处理；

[0008] 对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

[0009] 将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

[0010] 预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，

根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

[0011] 根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词和常用词组合；

[0012] 统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0013] 根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

[0014] 根据所述概率矩阵对数据进行分类。

[0015] 结合第一方面，在第一方面的第一种可能的实现方式中，所述预先定义冲突处理规则，包括：

[0016] 当对所述处理后的数据进行分类后，所述处理后的数据同时属于A类和B类时，当A类和B类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于A类和B类；

[0017] 当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于A类；

[0018] 或者当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于B类；

[0019] 当A类和B类互相冲突时，且没有冲突处理规则解决所述A类和B类的冲突时，则标识所述处理后的数据冲突。

[0020] 结合第一方面的第一种可能的实现方式，在第一方面的第二种可能的实现方式中，所述根据所述冲突处理规则得到有效数据，包括：

[0021] 当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

[0022] 当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

[0023] 当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

[0024] 对化简后的分类集合，根据所述冲突处理规则中当A类和B类相互冲突时，且没有冲突处理规则解决A类和B类的冲突时，标识所述处理后的数据为冲突。

[0025] 结合第一方面的第二种可能的实现方式，在第一方面的第三种可能的实现方式中，所述根据所述冲突处理规则化简所述两个以上的分类集合，包括：

[0026] 当所述两个以上的分类集合中A类和B类不冲突时，将A类和B类化简为同一类；

[0027] 当所述两个以上的分类集合中A类和B类互相冲突时以A类为准，将A类和B类化简为A类；

[0028] 当所述两个以上的分类集合中A类和B类互相冲突时以B类为准，将A类和B类化简为B类。

[0029] 结合第一方面或者第一方面的第一种可能的实现方式或者第一方面的第二种可能的实现方式或者第一方面的第三种可能的实现方式，在第一方面的第四种可能的实现方式中，所述根据所述常用词表，筛选所述有效数据得到搭配词表，包括：

[0030] 根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于 3 时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

[0031] 对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

[0032] 结合第一方面的第四种可能的实现方式，在第一方面的第五种可能的实现方式中，所述统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数，包括：

[0033] 统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0034] 统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数。

[0035] 结合第一方面的第五种可能的实现方式，在第一方面的第六种可能的实现方式中，所述根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵，包括：

[0036] 将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

[0037] 根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

[0038] 结合第一方面或者第一方面的第一种可能的实现方式或者第一方面的第二种可能的实现方式或者第一方面的第三种可能的实现方式或者第一方面的第四种可能的实现方式或者第一方面的第五种可能的实现方式或者第一方面的第六种可能的实现方式，在第一方面的第七种可能的实现方式中，所述根据所述概率矩阵对数据进行分类，包括：

[0039] 在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

[0040] 将概率最大的列对应的类别作为所述数据的类别。

[0041] 第二方面，一种数据分类的装置，所述装置包括：

[0042] 处理单元，用于对数据进行分词、去停顿词的处理；

[0043] 第一统计单元，用于对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

[0044] 第一分类单元，用于将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

[0045] 解决冲突单元，用于预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

[0046] 筛选单元，用于根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词或者常用词组合；

[0047] 第二统计单元，用于统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0048] 归一化单元，用于根据所述常用词或者常用词组合在每一分类中的次数和所有分

类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

[0049] 第二分类单元，用于根据所述概率矩阵对数据进行分类。

[0050] 结合第二方面，在第二方面的第一种可能的实现方式中，所述解决冲突单元中执行步骤预先定义分类之间的关系和冲突处理规则，包括：

[0051] 当对所述处理后的数据进行分类后，所述处理后的数据同时属于A类和B类时，当A类和B类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于A类和B类；

[0052] 当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于A类；

[0053] 或者当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于B类；

[0054] 当A类和B类互相冲突时，且没有冲突处理规则解决所述A类和B类的冲突时，则标识所述处理后的数据冲突。

[0055] 结合第二方面的第一种可能的实现方式，在第二方面的第二种可能的实现方式中，所述解决冲突单元中执行步骤根据所述冲突处理规则得到有效数据，包括：

[0056] 当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

[0057] 当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

[0058] 当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

[0059] 对化简后的分类集合，根据所述冲突处理规则中当A类和B类相互冲突时，且没有冲突处理规则解决A类和B类的冲突时，标识所述处理后的数据为冲突。

[0060] 结合第二方面的第二种可能的实现方式，在第二方面的第三种可能的实现方式中，所述解决冲突单元中执行步骤根据所述冲突处理规则化简所述两个以上的分类集合，包括：

[0061] 当所述两个以上的分类集合中A类和B类不冲突时，将A类和B类化简为同一类；

[0062] 当所述两个以上的分类集合中A类和B类互相冲突时以A类为准，将A类和B类化简为A类；

[0063] 当所述两个以上的分类集合中A类和B类互相冲突时以B类为准，将A类和B类化简为B类。

[0064] 结合第二方面或者第二方面的第一种可能的实现方式或者第二方面的第二种可能的实现方式或者第二方面的第三种可能的实现方式，在第二方面的第四种可能的实现方式中，所述筛选单元具体用于：

[0065] 根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于3时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

[0066] 对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

[0067] 结合第二方面的四种可能的实现方式，在第二方面的第五种可能的实现方式中，

所述第二统计单元，具体用于：

[0068] 统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0069] 统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数。

[0070] 结合第二方面的五种可能的实现方式，在第二方面的第六种可能的实现方式中，所述归一化单元，包括：

[0071] 矩阵单元，用于将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

[0072] 概率矩阵单元，用于根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

[0073] 结合第二方面或者第二方面的第一种可能的实现方式或者第二方面的第二种可能的实现方式或者第二方面的第三种可能的实现方式或者第二方面的第四种可能的实现方式或者第二方面的第五种可能的实现方式或者第二方面的第六种可能的实现方式，在第二方面的第七种可能的实现方式中，所述第二分类单元，包括：

[0074] 筛选子单元，用于在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

[0075] 第二分类子单元，用于将概率最大的列对应的类别作为所述数据的类别。与现有技术相比，本发明实施例提供一种数据分类的方法，所述方法通过预先定义分类的冲突处理规则，解决分类中有冲突的数据，得到有效数据，从而得到无歧义、无冲突的数据；并使用常用词表，筛选有效数据得到搭配词表，根据所述搭配词表形成概率矩阵，从而避免出现数据稀疏的情况。

附图说明

[0076] 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他的附图。

[0077] 图 1 是本发明实施例提供的一种数据分类的方法流程图；

[0078] 图 2 是本发明实施例提供的一种数据分类装置的装置结构图；

[0079] 图 3 是本发明实施例提供的一种数据分类装置中归一化单元的装置结构图；

[0080] 图 4 是本发明实施例提供的一种数据分类装置中第二分类单元的装置结构图；

[0081] 图 5 是本发明实施例提供的一种数据分类装置的装置结构图。

具体实施方式

[0082] 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他实施例，都属于本发明保护的范围。

[0083] 参考图 1，图 1 是本发明实施例提供的一种数据分类的方法流程图。如图 1 所示，所述方法包括以下步骤：

[0084] 步骤 101，对数据进行分词、去停顿词的处理；

[0085] 具体的，可以用自然语言处理工具进行处理，完成分词、去停顿词等工作。

[0086] 步骤 102，对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

[0087] 可选地，所述常用词的取值范围为前 20%。

[0088] 具体的，对所有数据做完处理后，统计处理后的所有数据中的词的频率，并按照频率的大小对所有数据中的词进行排序，选取排序前 20% 的词作为常用词，生成常用词表。

[0089] 步骤 103，将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

[0090] 具体的，使用传统的方法对所有数据进行分类或者使用朴素贝叶斯方法对所有数据进行分类。假设同一个数据分类后既是 A 类，又是 B 类，A 类是雷击，B 类是进水，则 A 类和 B 类是同时出现的，则该数据分类后所属的类别是不冲突的。

[0091] 步骤 104，预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

[0092] 可选地，所述预先定义冲突处理规则，包括：

[0093] 当对所述处理后的数据进行分类后，所述处理后的数据同时属于 A 类和 B 类时，当 A 类和 B 类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于 A 类和 B 类；

[0094] 当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 A 类；

[0095] 或者当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 B 类；

[0096] 当 A 类和 B 类互相冲突时，且没有冲突处理规则解决所述 A 类和 B 类的冲突时，则标识所述处理后的数据冲突。

[0097] 具体的，冲突处理规则是由连个分类和一个操作符组成。可以定义一些符号表示这些规则：

[0098] A ∞ B 表示 A、B 不冲突，不冲突。比如，A 类是进水，B 类是雷击，进水和雷击可能是因为同一个原因下雨而同时出现的，则 A 类和 B 类是不冲突的；

[0099] A ≺ B 表示 A、B 两个类互相冲突，而且冲突时以分类 B 的为准，假设 A 类是焊接，B 类是雷击，根据外部经验，A 类和 B 类是无关的，既无法找到一个原因是既能造成焊接又能造成雷击，因此，A 类和 B 类是冲突的；

[0100] A ≻ B 表示 A、B 两个类互相冲突，而且冲突时以分类 A 的为准；

[0101] 比如：焊接≺雷击，或者雷击≻焊接；

[0102] A ≠ B 表示 A、B 两个类互相冲突，但是没有规则解决冲突，比如：进水≠焊接。

[0103] 可选地，所述根据所述冲突处理规则得到有效数据，包括：

[0104] 当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标

识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

[0105] 当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

[0106] 当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

[0107] 对化简后的分类集合，根据所述冲突处理规则中当A类和B类相互冲突时，且没有冲突处理规则解决A类和B类的冲突时，标识所述处理后的数据为冲突。

[0108] 其中，所述有效数据是标识为A类和/或B类的数据。

[0109] 具体的，参考如下的伪代码：

[0110]

```
for 一条数据 in 数据集
{
        获得数据的初始分类;
        if 存在多个分类
        {
                if(分类数量==2)
                {
                        if(规则1)
                          标志不冲突;
                        if(规则2 or 规则3)
                          标志不冲突, 按照规则修改分类状态
                        if(规则4 )
                          标志冲突;
                }
                else
                {
                        规则1化简集合;
                        规则2化简集合;
                        for(分类1, 分类2 in 化简后的分类集合)
                        {
                                if(规则4 )
                                        {
                                                标志冲突;
                                                return;
                                        }

                        }
                        标志为不冲突;
                }
        }
        else
        {
                标志不冲突;
        }
}
```

可选地，

[0111] 所述根据所述冲突处理规则化简所述两个以上的分类集合，包括：

[0112] 当所述两个以上的分类集合中A类和B类不冲突时，将A类和B类化简为同一类；

[0113] 当所述两个以上的分类集合中A类和B类互相冲突时以A类为准，将A类和B类化简为A类；

[0114] 当所述两个以上的分类集合中A类和B类互相冲突时以B类为准，将A类和B类

化简为 B 类。

[0115] 步骤 105，根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词和常用词组合；

[0116] 可选地，所述根据所述常用词表，筛选所述有效数据得到搭配词表，包括：

[0117] 根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于 3 时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

[0118] 对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

[0119] 具体的，当数据 i1 根据所述常用词表筛选后，得到“失效”，则搭配词表中会出现“失效”；当数据 i2 根据所述常用词表筛选后，得到“短路，脱落”，则搭配词表中会出现“短路，脱落”。

[0120] 同时，当数据 i1 根据常用词表筛选后，“失效”出现两次，则仅按照一次做处理，即搭配词表中“失效”仅出现一次；当数据 i3 根据所述常用词表筛选后，得到“脱落，短路”，则搭配词表中可用“短路，脱落”替代，不考虑词的顺序；当数据 i4 根据常用词表筛选后，出现“脱落，短路，失效，雷击，进水”等四个以上的词时，可不考虑将筛选后的数据 i4 作为搭配词表中的常用词组合。

[0121] 搭配词表的格式可如表 1 所示：

[0122]

常用词组合	词的个数
失效	1
短路，脱落	2
…	…

[0123] 表 1

[0124] 步骤 106，统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0125] 可选地，所述统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数，形成矩阵，包括：

[0126] 统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0127] 统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数。

[0128] 具体的，统计方法可参考如下的伪代码：

[0129]

```
for 一条数据  in 数据集
{
        使用通用词表过滤，保留常用词
        保存记录对应的词或者词的搭配；
}
得到搭配词表

for 词条 in 搭配词表
{
        for 分类 in 所有类别
        {
                词条在当前分类出现的次数；
                保存数据；
        }
}
```

①
②
出有两个原则：
1一个词在短句中出现了多次，只按一次处理。
2每种搭配中词语的顺序不做限制。

[0130] 统计后的格式可参考表 2：

[0131]

	进水	雷击	焊接	失效
单板、烧毁、短路	32	980	50	20
烧焦	15	90	70	5
…	…	…	. . .	

[0132] 表 2

[0133] 如表 2 所示，对有效数据根据搭配词表进行筛选，例如，对进水下的所有有效数据根据搭配词表筛选，假设 i1 数据属于进水，且 i1 数据根据搭配词表筛选后得到“烧焦”，则统计进水类别下“烧焦”为 1，依此类推。假设 i1 数据筛选后统计到“烧焦”出现次数大于一次，仅以统计一次。

[0134] 步骤 107，根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

[0135] 可选地，所述根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵，包括：

[0136] 将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

[0137] 根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

[0138] 具体的，参考表 3，

[0139]

	进水	雷击	焊接	失效
单板、烧毁、短路	0.295	0.906	0.046	0.018
烧焦	0.083	0.5	0.389	0.028
…	…	…	. . .	

[0140] 表 3

[0141] 具体的，以计算“烧焦”为例，在“进水”分类下的概率为 15/(15+90+70+5)=0.083，在“雷击”分类下的概率为 90/(15+90+70+5)=0.5，在“焊接”分类下的概率为 70/(15+90+70+5)=0.389，在“失效”分类下的概率为 5/(15+90+70+5)=0.028。

[0142] 步骤 108，根据所述概率矩阵对数据进行分类。

[0143] 所述根据所述概率矩阵对数据进行分类，包括：

[0144] 在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

[0145] 将概率最大的列对应的类别作为所述数据的类别。

[0146] 具体的，当新数据 i5 出现时，对所述 i5 数据进行分词、去停顿词的处理；根据所述搭配词表，对处理后的 i5 数据进行筛选，当筛选后得到“进水，短路，腐蚀”这组搭配属于“进水”分类的概率是 0.7，而“进水，短路”这组搭配属于“进水”分类的概率是 0.8，则以最长常用词组合“进水，短路，腐蚀”这组搭配的概率为准，即数据 i5 的进水的概率是 0.7。

[0147] 假设筛选后得到“烧焦”、“进水，短路”，则分别计算“烧焦”和“进水，短路”在“进水”、“雷击”、“焊接”、“失效”下的概率，即计算“烧焦”和“进水，短路”在“进水”类别下的概率为 A1+B1，A1 是“烧焦”在“进水”类别下的概率，B1 是“进水，短路”在“进水”类别下的概率；计算“烧焦”和“进水，短路”在“雷击”类别下的概率为 A2+B2，A2 是“烧焦”在“雷击”类别下的概率，B2 是“进水，短路”在“雷击”类别下的概率；计算“烧焦”和“进水，短路”在“焊接”类别下的概率为 A3+B3，A3 是“烧焦”在“焊接”类别下的概率，B3 是“进水，短路”在“焊接”类别下的概率；计算“烧焦”和“进水，短路”在“失效”类别下的概率为 A4+B4，A4 是“烧焦”在“失效”类别下的概率，B4 是“进水，短路”在“失效”类别下的概率，归一化后得到数据在所有类别下的概率，即 i5 数据在“进水”类别下的概率为(A1+B1)/(A1+B1+A2+B2+A3+B3+A4+B4)。

[0148] 本发明实施例提供一种数据分类的方法，所述方法通过预先定义分类的冲突处理规则，解决分类中有冲突的数据，得到有效数据，从而得到无歧义、无冲突的数据；并使用常用词表，筛选有效数据得到搭配词表，根据所述搭配词表形成概率矩阵，从而避免出现数据稀疏的情况。

[0149] 参考图 2，图 2 是本发明实施例提供的一种数据分类装置的装置结构图。如图 2 所示，所述装置包括以下单元：

[0150] 处理单元 201，用于对数据进行分词、去停顿词的处理；

[0151] 具体的，可以用自然语言处理工具进行处理，完成分词、去停顿词等工作。

[0152] 第一统计单元 202，用于对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

[0153] 可选地，所述常用词的取值范围为前 20%。

[0154] 具体的，对所有数据做完处理后，统计处理后的所有数据中的词的频率，并按照频率的大小对所有数据中的词进行排序，选取排序前 20% 的词作为常用词，生成常用词表。

[0155] 第一分类单元 203，用于将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

[0156] 具体的，使用传统的方法对所有数据进行分类或者使用朴素贝叶斯方法对所有数

据进行分类。假设同一个数据分类后既是 A 类，又是 B 类，A 类是雷击，B 类是进水，则 A 类和 B 类是同时出现的，则该数据分类后所属的类别是不冲突的。

[0157] 解决冲突单元 204，用于预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

[0158] 可选地，所述解决冲突单元中执行步骤预先定义冲突处理规则，包括：

[0159] 当对所述处理后的数据进行分类后，所述处理后的数据同时属于 A 类和 B 类时，当 A 类和 B 类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于 A 类和 B 类；

[0160] 当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 A 类；

[0161] 或者当 A 类和 B 类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于 B 类；

[0162] 当 A 类和 B 类互相冲突时，且没有冲突处理规则解决所述 A 类和 B 类的冲突时，则标识所述处理后的数据冲突。

[0163] 具体的，冲突处理规则是由连个分类和一个操作符组成。可以定义一些符号表示这些规则：

[0164] $A \infty B$ 表示 A、B 不冲突，不冲突。比如，A 类是进水，B 类是雷击，进水和雷击可能是因为同一个原因下雨而同时出现的，则 A 类和 B 类是不冲突的；

[0165] $A \prec B$ 表示 A、B 两个类互相冲突，而且冲突时以分类 B 的为准，假设 A 类是焊接，B 类是雷击，根据外部经验，A 类和 B 类是无关的，既无法找到一个原因是既能造成焊接又能造成雷击，因此，A 类和 B 类是冲突的；

[0166] $A \succ B$ 表示 A、B 两个类互相冲突，而且冲突时以分类 A 的为准；

[0167] 比如：焊接≺雷击，或者雷击≻焊接；

[0168] $A \neq B$ 表示 A、B 两个类互相冲突，但是没有规则解决冲突，比如：进水≠焊接。

[0169] 可选地，所述解决冲突单元中执行步骤根据所述冲突处理规则得到有效数据，包括：

[0170] 当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

[0171] 当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

[0172] 当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

[0173] 对化简后的分类集合，根据所述冲突处理规则中当 A 类和 B 类相互冲突时，且没有冲突处理规则解决 A 类和 B 类的冲突时，标识所述处理后的数据为冲突。

[0174] 其中，所述有效数据是标识为 A 类和 / 或 B 类的数据。

[0175] 具体的，参考如下：

[0176]

```
for 一条数据 in 数据集
{
        获得数据的初始分类；
        if 存在多个分类
        {
                if(分类数量==2)
                {
                        if(规则1)
                          标志不冲突；
                        if(规则2 or 规则3)
                          标志不冲突，按照规则修改分类状态
                        if(规则4 )
                          标志冲突；
                }
                else
                {
                        规则1化简集合；
                        规则2化简集合；
                        for(分类1，分类2 in 化简后的分类集合)
                        {
                                if(规则4 )
                                        {
                                                标志冲突；
                                                return；
                                        }

                        }
                        标志为不冲突；
                }
        }
        else
        {
                标志不冲突；
        }
}
```

[0177] 可选地，所述解决冲突单元中执行步骤根据所述冲突处理规则化简所述两个以上的分类集合，包括：

[0178] 当所述两个以上的分类集合中 A 类和 B 类不冲突时，将 A 类和 B 类化简为同一类；

[0179] 当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 A 类为准，将 A 类和 B 类化简为 A 类；

[0180] 当所述两个以上的分类集合中 A 类和 B 类互相冲突时以 B 类为准，将 A 类和 B 类化简为 B 类。

[0181] 筛选单元 205，用于根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词或者常用词组合；

[0182] 可选地，所述筛选单元具体用于：

[0183] 根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于 3 时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

[0184] 对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

[0185] 具体的，当数据 i1 根据所述常用词表筛选后，得到“失效”，则搭配词表中会出现

“失效”;当数据 i2 根据所述常用词表筛选后,得到“短路,脱落”,则搭配词表中会出现“短路,脱落”。

[0186] 同时,当数据 i1 根据常用词表筛选后,“失效”出现两次,则仅按照一次做处理,即搭配词表中“失效”仅出现一次 ;当数据 i3 根据所述常用词表筛选后,得到“脱落,短路”,则搭配词表中可用“短路,脱落”替代,不考虑词的顺序 ;当数据 i4 根据常用词表筛选后,出现“脱落,短路,失效,雷击,进水”等四个以上的词时,可不考虑将筛选后的数据 i4 作为搭配词表中的常用词组合。

[0187] 搭配词表的格式可如表 1 所示 :

[0188]

常用词组合	词的个数
失效	1
短路,脱落	2
…	…

[0189] 表 1

[0190] 第二统计单元 206,用于统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数 ;

[0191] 可选地,所述第二统计单元 206,具体用于 :

[0192] 统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数 ;

[0193] 统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数。

[0194] 具体的,统计方法可参考如下 :

[0195]

```
for 一条数据  in 数据集
{
        使用通用词表过滤，保留常用词
        保存记录对应的词或者词的搭配;
}
得到搭配词表

for 词条 in 搭配词表
{
        for 分类 in 所有类别
        {
                词条在当前分类出现的次数;
                保存数据;
        }
}
```

①
②
出有两个原则:
1一个词在短句中出现了多次，只按一次处理。
2每种搭配中词语的顺序不做限制。

[0196] 统计后的表格形式参考表 2,

[0197]

	进水	雷击	焊接	失效
单板、烧毁、短路	32	980	50	20
烧焦	15	90	70	5
…	…	…	. . .	

[0198] 表 2

[0199] 如表 2 所示，对有效数据根据搭配词表进行筛选，例如，对进水下的所有有效数据根据搭配词表筛选，假设 i1 数据属于进水，且 i1 数据根据搭配词表筛选后得到“烧焦”，则统计进水类别下“烧焦”为 1，依此类推。假设 i1 数据筛选后统计到“烧焦”出现次数大于一次，仅以统计一次。

[0200] 归一化单元 207，用于根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

[0201] 可选地，所述归一化单元 207，包括：

[0202] 矩阵单元 301，用于将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

[0203] 概率矩阵单元 302，用于根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

[0204] 具体的，参考表 3，

[0205]

	进水	雷击	焊接	失效
单板、烧毁、短路	0.295	0.906	0.046	0.018
烧焦	0.083	0.5	0.389	0.028
…	…	…	. . .	

[0206] 表 3

[0207] 具体的，以计算“烧焦”为例，在“进水”分类下的概率为 15/(15+90+70+5)=0.083，在“雷击”分类下的概率为 90/(15+90+70+5)=0.5，在“焊接”分类下的概率为 70/(15+90+70+5)=0.389，在“失效”分类下的概率为 5/(15+90+70+5)=0.028。

[0208] 第二分类单元 208，用于根据所述概率矩阵对数据进行分类。

[0209] 可选地，所述第二分类单元 208，包括：

[0210] 筛选子单元 401，用于在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

[0211] 第二分类子单元 402，用于将概率最大的列对应的类别作为所述数据的类别。

[0212] 具体的，当新数据 i5 出现时，对所述 i5 数据进行分词、去停顿词的处理；根据所述搭配词表，对处理后的 i5 数据进行筛选，当筛选后得到“进水，短路，腐蚀”这组搭配属于“进水”分类的概率是 0.7，而“进水，短路”这组搭配属于“进水”分类的概率是 0.8，则以最

长常用词组合“进水，短路，腐蚀”这组搭配的概率为准，即数据 i5 的进水的概率是 0.7。

[0213] 假设筛选后得到“烧焦”、“进水，短路”，则分别计算“烧焦”和“进水，短路”在“进水”、“雷击”、“焊接”、“失效”下的概率，即计算“烧焦”和“进水，短路”在“进水”类别下的概率为 A1+B1，A1 是“烧焦”在“进水”类别下的概率，B1 是“进水，短路”在“进水”类别下的概率；计算“烧焦”和“进水，短路”在“雷击”类别下的概率为 A2+B2，A2 是“烧焦”在“雷击”类别下的概率，B2 是“进水，短路”在“雷击”类别下的概率；计算“烧焦”和“进水，短路”在“焊接”类别下的概率为 A3+B3，A3 是“烧焦”在“焊接”类别下的概率，B3 是“进水，短路”在“焊接”类别下的概率；计算“烧焦”和“进水，短路”在“失效”类别下的概率为 A4+B4，A4 是“烧焦”在“失效”类别下的概率，B4 是“进水，短路”在“失效”类别下的概率，归一化后得到数据在所有类别下的概率，即 i5 数据在“进水”类别下的概率为（A1+B1）/(A1+B1+A2+B2+A3+B3+A4+B4)。

[0214] 本发明实施例提供一种数据分类的装置，所述装置通过预先定义分类的冲突处理规则，解决分类中有冲突的数据，得到有效数据，从而得到无歧义、无冲突的数据；并使用常用词表，筛选有效数据得到搭配词表，根据所述搭配词表形成概率矩阵，从而避免出现数据稀疏的情况。

[0215] 参考图 5，图 5 是本发明实施例提供的一种数据分类装置的装置结构图。参考图 5，图 5 是本发明实施例提供的一种数据分类装置 500，本发明具体实施例并不对所述数据分类装置的具体实现做限定。所述数据分类装置 500 包括：

[0216] 处理器 (processor)501，通信接口 (Communications Interface)502，存储器 (memory)503，总线 504。

[0217] 处理器 501，通信接口 502，存储器 503 通过总线 504 完成相互间的通信。

[0218] 通信接口 502，用于与其他数据分类装置进行通信；

[0219] 处理器 501，用于执行程序。

[0220] 具体地，程序可以包括程序代码，所述程序代码包括计算机操作指令。

[0221] 处理器 501 可能是一个中央处理器 CPU，或者是特定集成电路 ASIC（Application Specific Integrated Circuit），或者是被配置成实施本发明实施例的一个或多个集成电路。

[0222] 存储器 503，用于存放程序。存储器 503 可能包含高速 RAM 存储器，也可能还包括非易失性存储器（non-volatile memory）。程序具体用于：

[0223] 对数据进行分词、去停顿词的处理；

[0224] 对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表；

[0225] 将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突；

[0226] 预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突；

[0227] 根据所述常用词表，筛选所述有效数据得到搭配词表，所述搭配词表包括常用词或者常用词组合；

[0228] 统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次

数；

[0229] 根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵；

[0230] 根据所述概率矩阵对数据进行分类。

[0231] 所述预先定义冲突处理规则，包括：

[0232] 当对所述处理后的数据进行分类后，所述处理后的数据同时属于A类和B类时，当A类和B类不冲突时，则标识所述处理后的数据不冲突，所述处理后的数据同时属于A类和B类；

[0233] 当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于A类；

[0234] 或者当A类和B类互相冲突时，则标识所述处理后的数据不冲突，所述处理后的数据属于B类；

[0235] 当A类和B类互相冲突时，且没有冲突处理规则解决所述A类和B类的冲突时，则标识所述处理后的数据冲突。

[0236] 所述根据所述冲突处理规则得到有效数据，包括：

[0237] 当对所述处理后的数据进行分类后，所述处理后的数据属于某一个分类时，则标识所述处理后的数据不冲突，所述处理后的数据属于所述某一个分类；

[0238] 当所述处理后的数据同时属于两个冲突的分类时，则根据所述冲突处理规则，得到有效数据；

[0239] 当所述处理后的数据同时属于两个以上的分类时，则根据所述冲突处理规则化简所述两个以上的分类集合；

[0240] 对化简后的分类集合，根据所述冲突处理规则中当A类和B类相互冲突时，且没有冲突处理规则解决A类和B类的冲突时，标识所述处理后的数据为冲突。

[0241] 所述根据所述冲突处理规则化简所述两个以上的分类集合，包括：

[0242] 当所述两个以上的分类集合中A类和B类不冲突时，将A类和B类化简为同一类；

[0243] 当所述两个以上的分类集合中A类和B类互相冲突时以A类为准，将A类和B类化简为A类；

[0244] 当所述两个以上的分类集合中A类和B类互相冲突时以B类为准，将A类和B类化简为B类。

[0245] 所述根据所述常用词表，筛选所述有效数据得到搭配词表，包括：

[0246] 根据所述常用词表，筛选所述有效数据得到所述常用词表中的词，当所述常用词表中的同一个词出现多次时，仅按照一次处理，当筛选后的词少于等于3时，得到所述有效数据的常用词组合，所述常用词组合中对有效数据中的词的顺序不做限制；

[0247] 对所有有效数据做筛选后，将所述筛选后的词和所述常用词组合形成搭配词表。

[0248] 所述统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数，包括：

[0249] 统计每一分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合的次数；

[0250] 统计所有分类中的所有有效数据中出现所述搭配词表中常用词或者常用词组合

的次数。

[0251] 所述根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵，包括：

[0252] 将所有分类作为列，将所述搭配词表中常用词或者常用词组合在每一列下出现的次数作为行，形成矩阵；

[0253] 根据所述矩阵，将所述矩阵中每一行在每一列的次数除以所述每一行在所有列的总次数，得到每一行在每一列的概率，形成概率矩阵。

[0254] 所述根据所述概率矩阵对数据进行分类，包括：

[0255] 在所述概率矩阵中找到数据筛选后得到的最长的常用词组合在每一列的概率；

[0256] 将概率最大的列对应的类别作为所述数据的类别。

[0257] 以上所揭露的仅为本发明较佳实施例而已，当然不能以此来限定本发明之权利范围，因此依本发明权利要求所作的等同变化，仍属本发明所涵盖的范围。

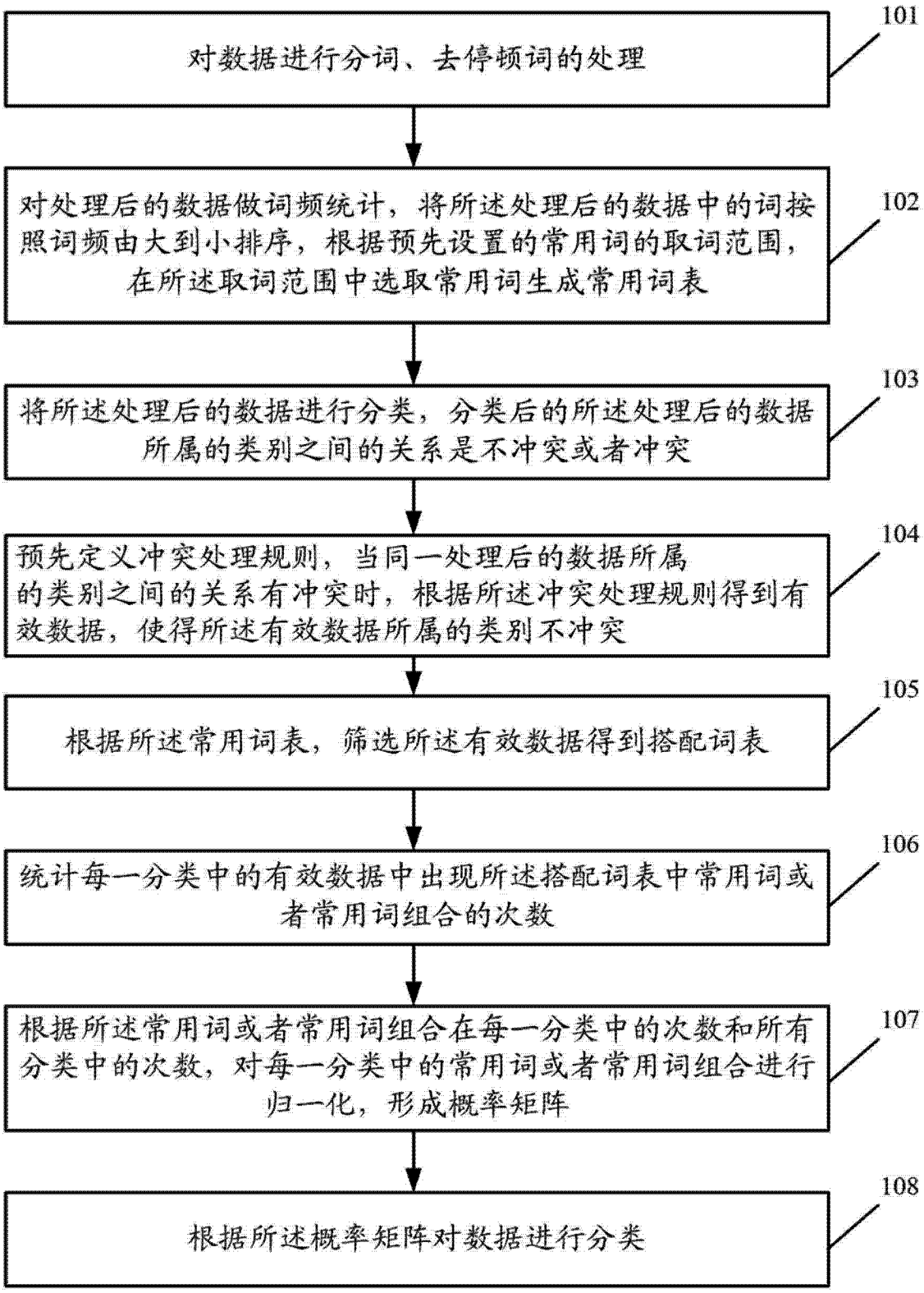

101
对数据进行分词、去停顿词的处理
102
对处理后的数据做词频统计，将所述处理后的数据中的词按照词频由大到小排序，根据预先设置的常用词的取词范围，在所述取词范围中选取常用词生成常用词表
103
将所述处理后的数据进行分类，分类后的所述处理后的数据所属的类别之间的关系是不冲突或者冲突
104
预先定义冲突处理规则，当同一处理后的数据所属的类别之间的关系有冲突时，根据所述冲突处理规则得到有效数据，使得所述有效数据所属的类别不冲突
105
根据所述常用词表，筛选所述有效数据得到搭配词表
106
统计每一分类中的有效数据中出现所述搭配词表中常用词或者常用词组合的次数
107
根据所述常用词或者常用词组合在每一分类中的次数和所有分类中的次数，对每一分类中的常用词或者常用词组合进行归一化，形成概率矩阵
108
根据所述概率矩阵对数据进行分类

图 1

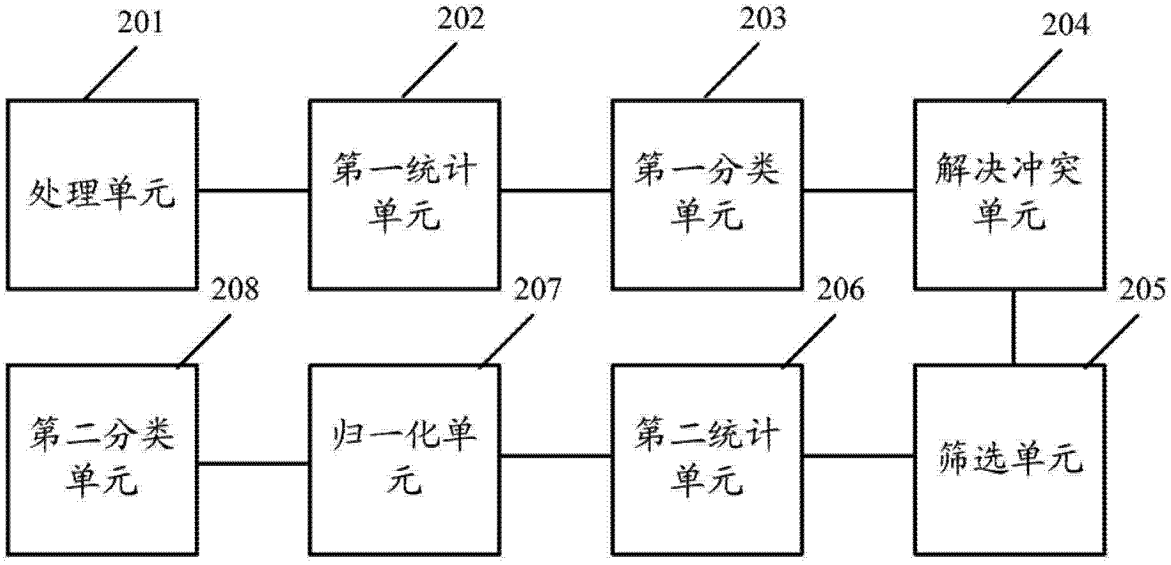

201
202
203
204
处理单元
第一统计单元
第一分类单元
解决冲突单元
208
207
206
205
第二分类单元
归一化单元
第二统计单元
筛选单元

图 2

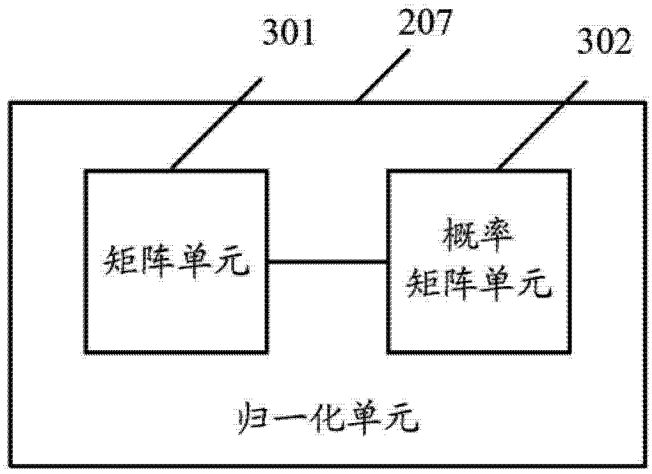

301
207
302
矩阵单元
概率矩阵单元
归一化单元

图 3

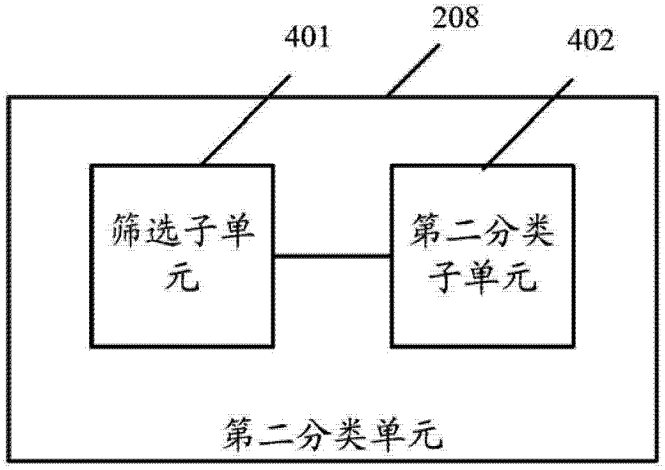

401
208
402
筛选子单元
第二分类子单元
第二分类单元

图 4

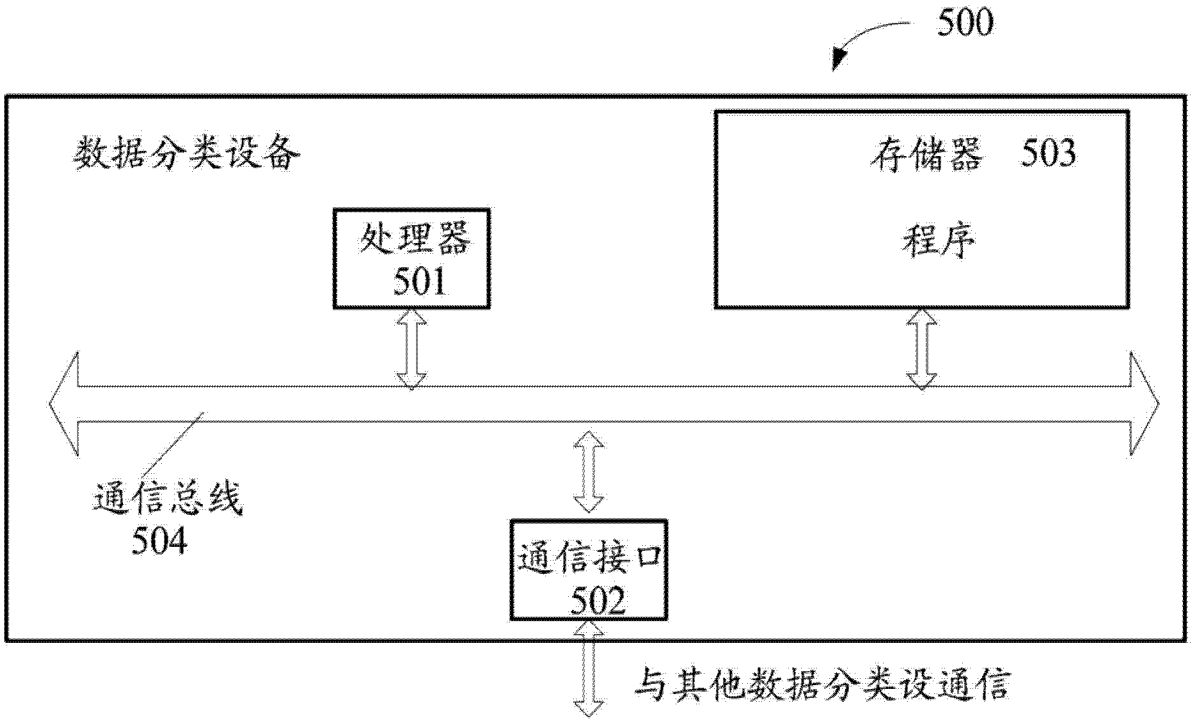

500
数据分类设备
存储器 503
程序
处理器 501
通信总线 504
通信接口 502
与其他数据分类设通信

图 5