(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105808526 B
(45)授权公告日 2019.07.30

(21)申请号 201610191177.3

(22)申请日 2016.03.30

(65)同一申请的已公布的文献号
申请公布号 CN 105808526 A

(43)申请公布日 2016.07.27

(73)专利权人 北京京东尚科信息技术有限公司
地址 100080 北京市海淀区杏石口路65号西杉创意园西区11C楼东段1-4层西段1-4层
专利权人 北京京东世纪贸易有限公司

(72)发明人 高维国 陈海勇

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 许蓓

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 16/35(2019.01)

(56)对比文件
JP H11120183 A,1999.04.30,
孙兴东 等.一种基于聚类的微博关键词提取方法的研究与实现.《技术研究》.2014,(第12期),第29页.

审查员 薛双双

权利要求书5页 说明书11页 附图7页

(54)发明名称

商品短文本核心词提取方法和装置

(57)摘要

本发明公开了一种商品短文本核心词提取方法和装置，涉及大数据处理领域。其中的商品短文本核心词提取方法包括：获取数据集内的商品短文本；对商品短文本进行数据清理；根据商品短文本的上下文信息获得商品短文本的文档向量；根据文档向量对数据集内的商品短文本进行聚类；确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重；根据各个分词的聚类层级权重确定商品短文本的核心词。本发明参考了商品短文本中分词的上下文信息来获得该商品短文本的文档向量，可以弥补短文本信息量少的缺陷，使基于文档向量的聚类结果更准确，进而依据分词在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

CN 105808526 B

1.一种商品短文本核心词提取方法,其特征在于,包括:

获取数据集内的商品短文本;

对所述商品短文本进行分词处理;

根据所述商品短文本的上下文信息获得所述商品短文本的文档向量;

根据所述文档向量对所述数据集内的商品短文本进行聚类;

确定所述商品短文本中的各个分词在所述商品短文本所属的类别的聚类层级权重;

确定所述商品短文本中的各个分词的商品词性信息,根据商品词性和词性权重的对应关系确定所述商品短文本中的分词对应的词性权重,并根据所述商品短文本中的各个分词对应的词性权重确定所述各个分词的局部权重;

根据所述各个分词的局部权重以及聚类层级权重确定所述商品短文本的核心词。

2.根据权利要求1所述的方法,其特征在于,

还包括:确定所述商品短文本中的各个分词在所述数据集内的逆向文件频率,将所述逆向文件频率确定为所述各个分词的文档层级权重;

所述根据所述各个分词的局部权重以及聚类层级权重确定所述商品短文本的核心词包括:

根据所述各个分词的局部权重和文档层级权重以及聚类层级权重确定所述商品短文本的核心词。

3.根据权利要求1所述的方法,其特征在于,所述根据所述商品短文本的上下文信息获得所述商品短文本的文档向量包括:

根据所述数据集内商品短文本的平均长度确定窗口参数;

采用向量运算工具word2vec,将所述数据集作为输入语料、确定的窗口参数作为窗口大小进行计算,得到所述数据集内的商品短文本的文档向量。

4.根据权利要求1或2所述的方法,其特征在于,其中,所述聚类的数量根据商品品类的数量确定。

5.根据权利要求1或2所述的方法,其特征在于,所述确定所述商品短文本中的各个分词在所述商品短文本所属的类别的聚类层级权重包括:

采用卡方公式计算所述商品短文本中的各个分词在所述商品短文本所属的类别的卡方值,将所述卡方值作为所述商品短文本中的各个分词在所述商品短文本所属的类别的聚类层级权重。

6.根据权利要求1所述的方法,其特征在于,采用以下方法确定所述商品词性和词性权重的对应关系:

获取训练语料,所述训练语料包括若干训练商品短文本,并标注各个训练商品短文本中的核心词和核心词对应的商品词性信息,将具有相同商品词性信息的核心词的数量与所述训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重;

或者,

建立用户搜索的搜索词和点击的商品短文本的对应关系,将所述搜索词标注为对应的商品短文本的核心词,并标注所述核心词对应的商品词性信息,将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。

7.根据权利要求1所述的方法，其特征在于，所述根据所述商品短文本中的各个分词对应的词性权重确定所述各个分词的局部权重包括：

对所述商品短文本中的各个分词对应的词性权重进行归一化处理，获得所述各个分词的局部权重。

8.根据权利要求2所述的方法，其特征在于，所述根据所述各个分词的局部权重以及聚类层级权重确定所述商品短文本的核心词包括：

将所述各个分词的局部权重以及聚类层级权重加权求和作为所述各个分词的核心权重，或者，将所述各个分词的局部权重和文档层级权重以及聚类层级权重加权求和作为所述各个分词的核心权重；

将所述商品短文本中核心权重的值最大的分词确定为所述商品短文本的核心词。

9.根据权利要求2所述的方法，其特征在于，

采用以下方法确定局部权重以及聚类层级权重的加权系数：

对训练数据集内的各个商品短文本进行分词处理；

标注所述训练数据集内各个商品短文本的核心词和非核心词；

计算训练数据集内各个分词的局部权重以及聚类层级权重；

将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重以及聚类层级权重，采用机器学习的方法计算局部权重以及聚类层级权重的加权系数；

或者，

采用以下方法确定局部权重和文档层级权重以及聚类层级权重的加权系数：

对训练数据集内的各个商品短文本进行分词处理；

标注所述训练数据集内各个商品短文本的核心词和非核心词；

计算训练数据集内各个分词的局部权重和文档层级权重以及聚类层级权重；

将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重以及聚类层级权重，采用机器学习的方法计算局部权重和文档层级权重以及聚类层级权重的加权系数。

10.根据权利要求1所述的方法，其特征在于，还包括：在对所述商品短文本进行分词处理之前，去掉商品短文本中的停用词和标点符号。

11.根据权利要求1所述的方法，其特征在于，还包括：在对所述商品短文本进行分词处理之后，统计所述数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词。

12.根据权利要求1所述的方法，其特征在于，其中，所述商品短文本包括商品标题、商品评论或商品信息页内容。

13.根据权利要求2所述的方法，其特征在于，其中，商品词性信息包括品牌、系列名称、品类、名词、属性词、款式、修饰词中的一个或多个。

14.一种商品短文本核心词提取装置，其特征在于，包括：

预处理模块，所述预处理模块包括：

短文本获取单元，用于获取数据集内的商品短文本；

分词单元，用于对所述商品短文本进行分词处理；

权重确定模块，包括：

聚类层级权重确定子模块，用于确定所述商品短文本中各个分词的权重；以及，局部权重确定子模块；

所述聚类层级权重确定子模块包括：

文档向量确定单元，用于根据所述商品短文本的上下文信息获得所述商品短文本的文档向量；

聚类单元，用于根据所述文档向量对所述数据集内的商品短文本进行聚类；

聚类层级权重确定单元，用于确定所述商品短文本中的各个分词在所述商品短文本所属的类别的聚类层级权重；

所述局部权重确定子模块，包括：

商品词性信息确定单元，用于确定所述商品短文本中的各个分词的商品词性信息；

词性权重确定单元，用于根据商品词性及其词性权重的对应关系确定所述商品短文本中的分词对应的词性权重；

局部权重确定单元，用于根据所述商品短文本中的各个分词对应的词性权重确定所述各个分词的局部权重；

核心词确定模块，用于根据所述各个分词的局部权重以及聚类层级权重确定所述商品短文本的核心词。

15.根据权利要求14所述的装置，其特征在于，

所述权重确定模块还包括文档层级权重确定子模块，所述核心词确定模块用于根据所述各个分词的局部权重和文档层级权重以及聚类层级权重确定所述商品短文本的核心词；

其中，所述文档层级权重确定子模块用于确定所述商品短文本中的各个分词在所述数据集内的逆向文件频率，将其确定为所述各个分词的文档层级权重。

16.根据权利要求14所述的装置，其特征在于，所述文档向量确定单元用于根据所述数据集内商品短文本的平均长度确定窗口参数，采用向量运算工具word2vec，将所述数据集作为输入语料、确定的窗口参数作为窗口大小进行计算，得到所述数据集内的商品短文本的文档向量。

17.根据权利要求14或15所述的装置，其特征在于，其中，所述聚类单元聚类的数量根据商品品类的数量确定。

18.根据权利要求14或15所述的装置，其特征在于，所述聚类层级权重确定单元用于采用卡方公式计算所述商品短文本中的各个分词在所述商品短文本所属的类别的卡方值，将所述卡方值作为所述商品短文本中的各个分词在所述商品短文本所属的类别的聚类层级权重。

19.根据权利要求14所述的装置，其特征在于，所述局部权重确定子模块还包括第一词性权重对应单元或者第二词性权重对应单元；

其中，所述第一词性权重对应单元用于获取包括若干训练商品短文本的训练语料，并标注各个训练商品短文本中的核心词和核心词对应的商品词性信息，将具有相同商品词性信息的核心词的数量与所述训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重；

其中，所述第二词性权重对应单元用于建立用户搜索的搜索词和点击的商品短文本的对应关系，将所述搜索词标注为对应的商品短文本的核心词，并标注所述核心词对应的商

品词性信息，再将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。

20.根据权利要求14所述的装置，其特征在于，所述局部权重确定单元用于对所述商品短文本中的各个分词对应的词性权重进行归一化处理，获得所述各个分词的局部权重。

21.根据权利要求15所述的装置，其特征在于，所述核心词确定模块包括：

核心权重计算单元，用于将所述各个分词的局部权重以及聚类层级权重、或者局部权重和文档层级权重以及聚类层级权重加权求和作为所述各个分词的核心权重；

核心词确定单元，用于将所述商品短文本中核心权重的值最大的分词确定为所述商品短文本的核心词。

22.根据权利要求21所述的装置，其特征在于，所述核心词确定模块还包括加权系数确定单元；

所述加权系数确定单元用于确定所述各个分词的局部权重以及聚类层级权重的加权系数并发送给所述核心权重计算单元，所述加权系数确定单元包括：

训练数据分词子单元，用于对训练数据集内的各个商品短文本进行分词处理；

训练数据标注子单元，用于标注所述训练数据集内各个商品短文本的核心词和非核心词；

分词权重计算子单元，用于计算训练数据集内各个分词的局部权重以及聚类层级权重；

机器学习子单元，用于将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重以及聚类层级权重，采用机器学习的方法计算局部权重以及聚类层级权重的加权系数；

或者，

所述加权系数确定单元用于确定所述各个分词的局部权重和文档层级权重以及聚类层级权重的加权系数并发送给所述核心权重计算单元，所述加权系数确定单元包括：

训练数据分词子单元，用于对训练数据集内的各个商品短文本进行分词处理；

训练数据标注子单元，用于标注所述训练数据集内各个商品短文本的核心词和非核心词；

分词权重计算子单元，用于计算训练数据集内各个分词的局部权重和文档层级权重以及聚类层级权重；

机器学习子单元，用于将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重以及聚类层级权重，采用机器学习的方法计算局部权重和文档层级权重以及聚类层级权重的加权系数。

23.根据权利要求14所述的装置，其特征在于，所述预处理模块还包括数据清理单元，和/或，分词过滤单元；

其中，所述数据清理单元，用于去掉商品短文本中的停用词和标点符号，并将处理后的所述商品短文本发送给所述分词单元；

其中，所述分词过滤单元，用于在对所述商品短文本进行分词处理之后，统计所述数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词。

24.根据权利要求15所述的装置，其特征在于，其中，商品词性信息包括品牌、系列名

称、品类、名词、属性词、款式、修饰词中的一个或多个;或者,所述商品短文本包括商品标题、商品评论或商品信息页内容。

25.一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现权利要求1～13中任一项所述的商品短文本核心词提取方法。

商品短文本核心词提取方法和装置

技术领域

[0001] 本发明涉及大数据处理领域，尤其涉及一种商品短文本核心词提取方法和装置。

背景技术

[0002] 随着电子商务网站的快速发展，电子商务网站提供了品类繁多的商品以及非常详细的商品信息，越来越多的顾客在网上购买商品、评论商品。因此，电子商务网站产生了大量的商品标题数据和商品评论数据。早期的通过人工标注挖掘核心词的方法已经不能适应从海量商品数据挖掘核心词的需要。

[0003] 目前一般采用词袋模型(BOW，Bag of Words)自动提取文本的核心词。在词袋模型中，文本被看作是无序的词汇集合，忽略语法甚至是单词的顺序。因此这种方法在提取长文本的核心词时效果较好，用于短文本时则效果不佳。相较于长文本，短文本包含的词语较少，因此具有特征稀疏、主题不明确的特点，准确提取核心词的难度更大。

发明内容

[0004] 本发明提供了一种商品短文本的核心词提取方法和装置，用来改善从商品短文本中提取核心词的准确度问题。

[0005] 根据本发明的第一个方面，提供了一种商品短文本核心词提取方法，包括：获取数据集内的商品短文本；对商品短文本进行分词处理；根据商品短文本的上下文信息获得商品短文本的文档向量；根据文档向量对数据集内的商品短文本进行聚类；确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重；根据各个分词的聚类层级权重确定商品短文本的核心词。

[0006] 在一个实施例中，商品短文本核心词提取方法还包括：确定商品短文本中的各个分词的商品词性信息，根据商品词性及其词性权重的对应关系确定商品短文本中的分词对应的词性权重，并根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重；和/或，确定商品短文本中的各个分词在数据集内的逆向文件频率，将其确定为各个分词的文档层级权重；根据各个分词的聚类层级权重确定商品短文本的核心词包括：根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重确定商品短文本的核心词。

[0007] 在一个实施例中，根据商品短文本的上下文信息获得商品短文本的文档向量包括：根据数据集内商品短文本的平均长度确定窗口参数；采用向量运算工具word2vec，将数据集作为输入语料、确定的窗口参数作为窗口大小进行计算，得到数据集内的商品短文本的文档向量。

[0008] 在一个实施例中，聚类的数量根据商品品类的数量确定。

[0009] 在一个实施例中，确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重包括：采用卡方公式计算商品短文本中的各个分词在商品短文本所属的类别的卡方值，将卡方值作为商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。

[0010] 在一个实施例中，采用以下方法确定商品词性及其词性权重的对应关系：获取训练语料，训练语料包括若干训练商品短文本，并标注各个训练商品短文本中的核心词和核心词对应的商品词性信息，将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重；或者，建立用户搜索的搜索词和点击的商品短文本的对应关系，将搜索词标注为对应的商品短文本的核心词，并标注核心词对应的商品词性信息，将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。

[0011] 在一个实施例中，根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重包括：对商品短文本中的各个分词对应的词性权重进行归一化处理，获得各个分词的局部权重。

[0012] 在一个实施例中，根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重确定商品短文本的核心词包括：将各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重加权求和作为各个分词的核心权重；将商品短文本中核心权重的值最大的分词确定为商品短文本的核心词。

[0013] 在一个实施例中，采用以下方法确定局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数：对训练数据集内的各个商品短文本进行分词处理；标注训练数据集内各个商品短文本的核心词和非核心词；计算训练数据集内各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重；将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重，采用机器学习的方法计算局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数。

[0014] 在一个实施例中，商品短文本核心词提取方法还包括：在对商品短文本进行分词处理之前，去掉商品短文本中的停用词和标点符号。

[0015] 在一个实施例中，商品短文本核心词提取方法还包括：在对商品短文本进行分词处理之后，统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词。

[0016] 在一个实施例中，商品短文本包括商品标题、商品评论或商品信息页内容。

[0017] 在一个实施例中，商品词性信息包括品牌、系列名称、品类、名词、属性词、款式、修饰词中的一个或多个。

[0018] 根据本发明的第二个方面，提供一种商品短文本核心词提取装置，包括：预处理模块，预处理模块包括：短文本获取单元，用于获取数据集内的商品短文本；分词单元，用于对商品短文本进行分词处理；权重确定模块，包括聚类层级权重确定子模块，用于确定商品短文本中各个分词的权重；聚类层级权重确定子模块包括：文档向量确定单元，用于根据商品短文本的上下文信息获得商品短文本的文档向量；聚类单元，用于根据文档向量对数据集内的商品短文本进行聚类；聚类层级权重确定单元，用于确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重；核心词确定模块，用于根据各个分词的聚类层级权重确定商品短文本的核心词。

[0019] 在一个实施例中，权重确定模块还包括局部权重确定子模块和/或文档层级权重确定子模块；其中，局部权重确定子模块包括：商品词性信息确定单元，用于确定商品短文本中的各个分词的商品词性信息；词性权重确定单元，用于根据商品词性及其词性权重的

对应关系确定商品短文本中的分词对应的词性权重;局部权重确定单元,用于根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重;其中,文档层级权重确定子模块用于确定商品短文本中的各个分词在数据集内的逆向文件频率,将其确定为各个分词的文档层级权重;核心词确定模块用于根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重确定商品短文本的核心词。

[0020] 在一个实施例中,文档向量确定单元用于根据数据集内商品短文本的平均长度确定窗口参数,采用向量运算工具word2vec,将数据集作为输入语料、确定的窗口参数作为窗口大小进行计算,得到数据集内的商品短文本的文档向量。

[0021] 在一个实施例中,聚类单元聚类的数量根据商品品类的数量确定。

[0022] 在一个实施例中,聚类层级权重确定单元用于采用卡方公式计算商品短文本中的各个分词在商品短文本所属的类别的卡方值,将卡方值作为商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。

[0023] 在一个实施例中,局部权重确定模块还包括第一词性权重对应单元或者第二词性权重对应单元。其中,第一词性权重对应单元用于获取包括若干训练商品短文本的训练语料,并标注各个训练商品短文本中的核心词和核心词对应的商品词性信息,将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。其中,第二词性权重对应单元用于建立用户搜索的搜索词和点击的商品短文本的对应关系,将搜索词标注为对应的商品短文本的核心词,并标注核心词对应的商品词性信息,再将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。

[0024] 在一个实施例中,局部权重确定单元用于对商品短文本中的各个分词对应的词性权重进行归一化处理,获得各个分词的局部权重。

[0025] 在一个实施例中,核心词确定模块包括:核心权重计算单元,用于将各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重加权求和作为各个分词的核心权重;核心词确定单元,用于将商品短文本中核心权重的值最大的分词确定为商品短文本的核心词。

[0026] 在一个实施例中,核心词确定模块还包括加权系数确定单元,用于确定各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数并发送给核心权重计算单元,加权系数确定单元包括:训练数据分词子单元,用于对训练数据集内的各个商品短文本进行分词处理;训练数据标注子单元,用于标注训练数据集内各个商品短文本的核心词和非核心词;分词权重计算子单元,用于计算训练数据集内各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重;机器学习子单元,用于将核心词作为正样本,非核心词作为负样本,根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重,采用机器学习的方法计算局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数。

[0027] 在一个实施例中,预处理模块还包括数据清理单元,用于去掉商品短文本中的停用词和标点符号,并将处理后的商品短文本发送给分词单元。

[0028] 在一个实施例中,预处理模块还包括分词过滤单元,用于在对商品短文本进行分词处理之后,统计数据集内所有分词的出现频率,去掉商品短文本中出现频率低于过滤阈

值的分词。

[0029] 在一个实施例中，商品短文本包括商品标题、商品评论或商品信息页内容。

[0030] 在一个实施例中，商品词性信息包括品牌、系列名称、品类、名词、属性词、款式、修饰词中的一个或多个。

[0031] 本发明参考了商品短文本中分词的上下文信息来获得该商品短文本的文档向量，可以弥补短文本信息量少的缺陷，使基于文档向量的聚类结果更准确，进而依据分词在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

[0032] 此外，在聚类层级权重的基础上，通过结合分词的局部权重，可以进一步针对商品短文本的特点，从其商品词性信息的角度对结果进行优化，提高了核心词确定的准确性。

[0033] 此外，在聚类层级权重的基础上，通过结合分词的文档层级权重，能够进一步从数据集层面的重要性对核心词提取过程进行优化，从而提升了核心词确定的准确性。

[0034] 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

附图说明

[0035] 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。

[0036] 图1为本发明商品短文本核心词提取方法的一个实施例的流程图。

[0037] 图2为本发明商品短文本核心词提取方法的另一个实施例的流程图。

[0038] 图3A为本发明商品短文本核心词提取方法又一个实施例的流程图。

[0039] 图3B为本发明商品短文本核心词提取方法再一个实施例的流程图。

[0040] 图4为本发明确定各个权重的加权系数的流程图。

[0041] 图5为本发明商品短文本核心词提取装置的一个实施例的结构图。

[0042] 图6为本发明商品短文本核心词提取装置的另一个实施例的结构图。

[0043] 图7为本发明商品短文本核心词提取装置的又一个实施例的结构图。

具体实施方式

[0044] 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。以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。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。

[0045] 为了改善从商品短文本中提取核心词的准确度，提出本发明。在本发明中，一条商品短文本可以视为一个文档，其往往具有特征稀疏、主题不明确等特点。

[0046] 下面参考图1描述本发明一个实施例的商品短文本核心词提取方法。

[0047] 图1为本发明商品短文本核心词提取方法的一个实施例的流程图。如图1所示，该实施例的方法包括：

[0048] 首先，对数据集内的商品短文本进行预处理，得到各个商品短文本的分词。例如，可以采用步骤S102～S105实现预处理：

[0049] 步骤S102，获取数据集内的商品短文本。

[0050] 其中，数据集为待测商品短文本的集合，可以从存储商品短文本的数据库中获取。例如，可以采用Hadoop提供的数据仓库工具hive进行商品短文本数据的查询和获取，也可以根据业务需求和系统设置，采用例如网页爬虫等其他方式获取。商品短文本例如可以为商品标题、商品评论或商品信息页内容等，商品信息页内容例如可以为有关商品的材质、型号等短文本信息。

[0051] 步骤S103，对商品短文本进行数据清理。

[0052] 其中，可以根据业务需要选择性地进行步骤S103。数据清理例如包括去掉商品短文本中的停用词和标点符号，以便提升后续数据处理的效率。其中，停用词可以进行预先设定，例如可以设置包含助词、感叹词等无意义词汇或非目标词汇的停用词表，根据停用词表清理数据集。

[0053] 步骤S104，对商品短文本进行分词处理得到各个分词。

[0054] 步骤S105，统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词。

[0055] 其中，可以根据业务需要，在进行分词处理之后选择性地进行步骤S105。通过过滤出现频率低于过滤阈值的分词，可以提升后续核心词的提取效率。

[0056] 在预处理之后，可以确定分词的聚类层级权重，以衡量一个分词在某一类中的重要程度。例如，可以采用步骤S106～S110确定分词的聚类层级权重：

[0057] 步骤S106，根据商品短文本的上下文信息获得商品短文本的文档向量。

[0058] 其中，一种实现方法为，采用基于分词上下文语境的向量运算工具word2vec获得商品短文本的文档向量，从而将每条商品短文本表征为一个k维的向量，具体的维度可以设定。

[0059] 在采用word2vec工具进行文档向量的计算时，还可以调整输入参数以便对结果进行优化。例如，可以根据数据集内商品短文本的平均长度确定窗口参数，将数据集作为输入语料、确定的窗口参数作为窗口大小进行计算得到数据集内的商品短文本的文档向量。合理地确定窗口大小参数可以提高文档向量计算的准确性和效率。例如，还可以对size参数进行优化。size参数用于表示向量的大小，当size参数值较大时，计算的精度较高，但是对计算设备的性能要求也相应提高。使用时，可以根据需要对size参数进行调整。

[0060] 步骤S108，根据文档向量对数据集内的商品短文本进行聚类。

[0061] 由于文档向量可以将商品短文本抽象为向量空间中的若干点，因此可以采用现有的基于距离的聚类算法进行聚类，例如K-MEANS算法、K-MEDOIDS算法、BIRCH算法等等。将文档向量输入特定的聚类算法进行聚类即可得到各个类。

[0062] 在进行聚类时，还需要合理地确定聚类的数量，以使得具有共性的某类商品短文本能够尽可能划分到一类，尽量避免差异较大的商品短文本被划分为一类或者很相似的一些商品短文本却被聚到不同的类。

[0063] 对于商品短文本，其聚类数量可以根据商品品类的数量确定，例如可以设置聚类数量大致等于商品品类的数量。品类是指电商网站中根据商品特性对商品进行分类后形成

的类别名称，例如女装、箱包、日化、数码产品等，具体的品类数量、名称、细分程度等可以根据业务需求和产品情况进行设置。从而，使基于文档向量的聚类结果更准确，进而依据分词在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

[0064] 步骤S110，确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。

[0065] 其中，可以通过计算商品短文本中的各个分词与商品短文本所属类别的相关性来确定聚类层级权重。例如，采用卡方公式计算商品短文本中的各个分词在商品短文本所属的类别的卡方值后，可以将卡方值作为商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。

[0066] 例如，对于商品短文本“Apple iPhone 6s Plus(A1699)64G玫瑰金色移动联通电信4G手机”，设经过步骤S108的聚类操作后，该短文本属于“数码产品”类别。分词“手机”的卡方值例如可以参考表1进行计算。表1为卡方算法的独立样本四格表。如表1所示，设A为“手机”在“数码产品”类中出现的短文本数量，B为“手机”在“数码产品”以外的类中出现的短文本数量，C为“数码产品”类中不包含“手机”的短文本数量，D为“数码产品”以外的类中不包含“手机”的短文本数量，N为参与聚类的所有商品短文本的数量。

[0067]

	属于“数码产品”	不属于“数码产品”	总计
包含“手机”	A	B	A+B
不包含“手机”	C	D	C+D
总计	A+C	B+D	N

[0068] 表1

[0069] 对于短文本“Apple iPhone 6s Plus(A1699)64G玫瑰金色移动联通电信4G手机”中分词“手机”的卡方值，可以采用公式(1)进行计算：

[0070] $$K^2 = \frac{N*(A*D-B*D)^2}{(A+B)*(A+C)*(B+D)*(C+D)} \quad (1)$$

[0071] 卡方值K^2即为该短文本中“手机”的聚类层级权重。

[0072] 然后，采用步骤S112确定核心词：

[0073] 步骤S112，根据各个分词的聚类层级权重确定商品短文本的核心词。

[0074] 例如，可以在确定商品短文本中各个分词的聚类层级权重后，将聚类层级权重的值最大的分词作为核心词。也可以根据需要，采用其他方法依据聚类层级权重的值确定核心词。

[0075] 通过参考商品短文本中分词的上下文信息来获得该商品短文本的文档向量，可以弥补短文本信息量少的缺陷，使基于文档向量的聚类结果更准确，进而依据分词在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

[0076] 在确定商品短文本的核心词时，除了参考分词的聚类层级权重以外，还可以结合其他权重对结果进行优化。例如，可以结合局部权重衡量分词对应的商品词性信息在单条商品短文本中的重要程度，或/和，还可以结合文档层级权重衡量分词在整个数据集内的重要程度。下面对聚类层级权重结合其他权重对结果进行优化的情形进行描述。

[0077] 下面参考图2描述本发明另一个实施例的商品短文本核心词提取方法。

[0078] 图2为本发明商品短文本核心词提取方法的另一个实施例的流程图。如图2所示，该实施例的方法包括：

[0079] 首先，对数据集内的商品短文本进行预处理，得到各个商品短文本的分词。预处理的实现可以例如参考前文中步骤S102～S105的描述。

[0080] 在预处理之后，除了确定分词的聚类层级权重(参考步骤S106～S110)，还可以确定分词的局部权重，以衡量分词在单条商品短文本中的重要程度。例如，可以采用步骤S206～S210确定局部权重：

[0081] 步骤S206，确定商品短文本中的各个分词的商品词性信息。

[0082] 其中，在本发明实施例中，商品词性信息是指引入具有商品特点的词性信息，例如品牌、系列名称、品类、属性词、款式、修饰词、书名、作者等等。同一个商品短文本中的同一个分词可以具有多种商品词性信息。

[0083] 在采用本实施例的方法之前，可以预先设置商品词性信息字典，字典包括若干词语以及其对应的商品词性信息。在确定商品短文本中的各个分词的商品词性信息时，可以逐一查找分词在商品词性信息字典中对应的商品词性信息并进行标注。

[0084] 步骤S208，根据商品词性及其词性权重的对应关系确定商品短文本中的分词对应的词性权重。

[0085] 词性权重指某一类别的商品词性信息在商品短文本中的重要程度。例如，在商品标题中，品牌和品类为相对重要的词语，而“500ml”、“2米”等属性词为重要程度较低的词语。因此，可以根据商品词性信息的重要性和具体的业务需求设置商品词性信息对应的词性权重。上述过程可以人工设置，也可以采用基于统计的方法确定。

[0086] 步骤S210，根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重。

[0087] 当一个分词具有多种商品词性信息时，可以将这些商品词性信息对应的若干词性权重进行累加。此外，确定局部权重时，可以直接将分词对应的词性权重作为局部权重，也可以将商品短文本中所有的分词的词性权重进行归一化处理，将归一化处理后的结果作为各个分词的局部权重。设x为归一化处理之前某分词的局部权重，y为归一化处理之后该分词的局部权重，min和max分别为该分词所属的商品短文本中局部权重的最小值和最大值，则y的值可以通过公式(2)进行归一化计算：

[0088] $y=(x-min)/(max-min)$ (2)

[0089] 通过归一化的处理，能够直观地反映各个分词的局部权重的统计分布特性。

[0090] 然后，采用步骤S212确定核心词：

[0091] 步骤S212，根据各个分词的聚类层级权重和局部权重确定商品短文本的核心词。

[0092] 在聚类层级权重的基础上，通过结合分词的局部权重，可以进一步针对商品短文本的特点，从其商品词性信息的角度对结果进行优化，提高了核心词确定的准确性。

[0093] 本发明还提供了基于统计方式建立商品词性信息与词性权重的对应关系的方法。

[0094] 一种示例性的方法，首先，获取训练语料，训练语料包括若干训练商品短文本；然后，标注各个训练商品短文本中的核心词和核心词对应的商品词性信息，例如，可以通过线下字典标注核心词及其商品词性，利用字典中记录的核心词及其对应的商品词性，对训练预料中的核心词进行商品词性的标注；最后，将具有相同商品词性信息的核心词的数量与

训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。

[0095] 另一种示例性的方法，还可以根据用户的搜索行为自动标注核心词。首先，建立用户搜索的搜索词和点击的商品短文本的对应关系；然后，将搜索词标注为对应的商品短文本的核心词，并标注核心词对应的商品词性信息；最后，将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。其中，建立用户搜索的搜索词和点击的商品短文本的对应关系并标注核心词时，可以通过搜索点击日志建立搜索词和用户点击的商品短文本的对应关系并进行标注。例如，根据搜索点击日志，用户使用搜索词“手机”搜索商品，搜索结果为“Apple iPhone 6s 64G深空灰色电信4G手机”以及“小米Note白色移动4G手机”，而用户点击了前者，即“Apple iPhone 6s 64G深空灰色电信4G手机”，则将“手机”标注为“Apple iPhone 6s 64G深空灰色电信4G手机”的核心词。

[0096] 通过上述基于统计的方法确定商品词性信息对应的词性权重，使词性权重的值更能适用于当前的使用环境，从而提高了核心词提取的准确性。

[0097] 下面参考图3A描述本发明又一个实施例的商品短文本核心词提取方法。

[0098] 图3A为本发明商品短文本核心词提取方法的又一个实施例的流程图。如图3A所示，该实施例的方法包括：

[0099] 首先，对数据集内的商品短文本进行预处理，得到各个商品短文本的分词。预处理的实现可以例如参考前文中步骤S102～S105的描述。

[0100] 在预处理之后，除了确定分词的聚类层级权重(参考步骤S106～S110)，还可以确定分词的文档层级权重，以衡量分词在数据集的所有商品短文本中的重要程度。例如，可以采用步骤S306确定文档层级权重：

[0101] 步骤S306，确定商品短文本中的各个分词在数据集内的逆向文件频率，将其确定为各个分词的文档层级权重。

[0102] 逆向文件频率(IDF或idf，inverse document frequency)表示具有某一分词的文档在语料库中的出现频率，该频率越高，则对应的idf值越小。即，如果一个分词在大量不同的文档中出现，说明该分词无法代表某一文档的特性。由于每条商品短文本相当于一个文档，因此商品短文本中的各个分词在数据集内的逆向文件频率idf值可以采用公式(3)进行计算：

[0103] $$\mathbf{idf} = \mathbf{log}\left(\frac{\mathbf{M}}{\mathbf{L}}\right) \quad (3)$$

[0104] 其中，M表示商品短文总数量，L表示包含某一分词的商品短文本数量，idf值表示商品短文本中该分词的分档层级权重。

[0105] 然后，采用步骤S312a确定核心词：

[0106] 步骤S312a，根据各个分词的聚类层级权重和文档层级权重确定商品短文本的核心词。

[0107] 在聚类层级权重的基础上，通过结合分词的文档层级权重，能够进一步从数据集层面的重要性对核心词提取过程进行优化，从而提升了核心词确定的准确性。

[0108] 显然，根据图1-3描述的核心词提取方法，还可以结合各个实施例分别得到聚类层级权重、局部权重和文档层级权重，然后根据聚类层级权重、局部权重和文档层级权重共同

作为核心词的确定依据，即，执行步骤S312b，根据各个分词的聚类层级权重、局部权重和文档层级权重确定商品短文本的核心词。上述商品短文本核心词提取方法例如可以参考图3B。

[0109] 当根据一种以上的权重确定核心词时，可以首先将各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重加权求和作为各个分词的核心权重，再将商品短文本中核心权重的值最大的分词确定为商品短文本的核心词。下面参考图4描述本发明确定各个权重的加权系数的方法。

[0110] 图4为本发明确定各个权重的加权系数的方法的流程图。如图4所示，该实施例的方法包括：

[0111] 步骤S402，对训练数据集内的各个商品短文本进行分词处理。

[0112] 步骤S404，标注训练数据集内各个商品短文本的核心词和非核心词。

[0113] 例如，核心词与非核心词的标注可以采用人工手动标注的方式；也可以根据搜索点击日志，将用户搜索时采用的搜索词作为用户点击的搜索结果对应的核心词。

[0114] 步骤S406，计算训练数据集内各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重。

[0115] 其中，具体的计算方法采用前述各个实施例中计算各个分词的局部权重、文档层级权重中和聚类层级权重的方法。

[0116] 步骤S408，将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重，采用机器学习的方法，例如线性回归、决策树、神经网络等算法，计算局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数。

[0117] 通过采用上述方法，当根据多个权重确定商品短文本中的核心词时，能够调节各个权重之间的比重，提升核心词确定的准确性。

[0118] 显然，根据需要，也可以直接将不同的权重进行直接求和作为核心权重，即各个权重的加权系数均为1，从而方便计算。

[0119] 下面参考图5描述本发明一个实施例的商品短文本核心词提取装置。

[0120] 图5为本发明商品短文本核心词提取装置的一个实施例的结构图。如图5所示，该实施例的装置包括：预处理模块52、权重确定模块54和核心词确定模块56。预处理模块52包括：短文本获取单元522，用于获取数据集内的商品短文本；分词单元524，用于对商品短文本进行分词处理。权重确定模块54包括聚类层级权重确定子模块542，用于确定商品短文本中各个分词的权重。聚类层级权重确定子模块542包括：文档向量确定单元5422，用于根据商品短文本的上下文信息获得商品短文本的文档向量；聚类单元5424，用于根据文档向量对数据集内的商品短文本进行聚类；聚类层级权重确定单元5426，用于确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。核心词确定模块56用于根据各个分词的聚类层级权重确定商品短文本的核心词。

[0121] 通过采用文档向量确定单元参考商品短文本中分词的上下文信息来获得该商品短文本的文档向量，可以弥补短文本信息量少的缺陷，使聚类单元基于文档向量的聚类结果更准确，进而聚类层级权重确定单元在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

[0122] 其中，商品短文本可以包括商品标题或商品评论。

[0123] 其中，文档向量确定单元5422可以用于根据数据集内商品短文本的平均长度确定窗口参数，采用向量运算工具word2vec，将数据集作为输入语料、确定的窗口参数作为窗口大小进行计算，得到数据集内的商品短文本的文档向量。

[0124] 其中，聚类单元5424聚类的数量可以根据商品品类的数量确定，使基于文档向量的聚类结果更准确，进而依据分词在其商品短文本所属聚类类别的权重可以更加准确地从商品短文本中提取出核心词。

[0125] 其中，聚类层级权重确定单元5426可以用于采用卡方公式计算商品短文本中的各个分词在商品短文本所属的类别的卡方值，将卡方值作为商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重。

[0126] 下面参考图6描述本发明另一个实施例的商品短文本核心词提取装置。

[0127] 图6为本发明商品短文本核心词提取装置的另一个实施例的结构图。如图6所示，该实施例的权重确定模块54还包括局部权重确定子模块644和/或文档层级权重确定子模块646。

[0128] 其中，局部权重确定子模块644包括：商品词性信息确定单元6442，用于确定商品短文本中的各个分词的商品词性信息；词性权重确定单元6444，用于根据商品词性及其词性权重的对应关系确定商品短文本中的分词对应的词性权重；局部权重确定单元6446，用于根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重。在设置聚类层级权重确定子模块的基础上，通过结合局部权重确定子模块，可以进一步针对商品短文本的特点，从其商品词性信息的角度对结果进行优化，提高了核心词确定的准确性。

[0129] 其中，文档层级权重确定子模块646用于确定商品短文本中的各个分词在数据集内的逆向文件频率，将其确定为各个分词的文档层级权重；核心词确定模块56用于根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重确定商品短文本的核心词。在设置聚类层级权重确定子模块的基础上，通过结合文档层级权重确定子模块，能够进一步从数据集层面的重要性对核心词提取过程进行优化，从而提升了核心词确定的准确性。

[0130] 其中，商品词性信息可以包括品牌、系列名称、品类、名词、属性词、款式、修饰词中的一个或多个。

[0131] 其中，局部权重确定单元6446可以用于对商品短文本中的各个分词对应的词性权重进行归一化处理，获得各个分词的局部权重。通过归一化的处理，能够直观地反映各个分词的局部权重的统计分布特性。

[0132] 其中，核心词确定模块可以56包括：核心权重计算单元662，用于将各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重加权求和作为各个分词的核心权重；核心词确定单元664，用于将商品短文本中核心权重的值最大的分词确定为商品短文本的核心词。

[0133] 下面参考图7描述本发明又一个实施例的商品短文本核心词提取装置。

[0134] 图7为本发明商品短文本核心词提取装置的又一个实施例的结构图。如图7所示，局部权重确定子模块644还可以包括第一词性权重对应单元7442或者第二词性权重对应单元7444。第一词性权重对应单元7442用于获取包括若干训练商品短文本的训练语料，并标

注各个训练商品短文本中的核心词和核心词对应的商品词性信息，将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。第二词性权重对应单元7444用于建立用户搜索的搜索词和点击的商品短文本的对应关系，将搜索词标注为对应的商品短文本的核心词，并标注核心词对应的商品词性信息，再将具有相同商品词性信息的核心词的数量与训练语料中所有核心词的数量的比值作为该商品词性信息对应的词性权重。其中，第二词性权重对应单元7444例如可以通过搜索点击日志建立搜索词和用户点击的商品短文本的对应关系。

[0135] 通过采用上述基于统计的第一词性权重对应单元或者第二词性权重对应单元确定商品词性信息对应的词性权重，能够使词性权重的值更能适用于当前的使用环境，从而提高了核心词提取的准确性。

[0136] 此外，核心词确定模块56还可以包括加权系数确定单元762，用于确定各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数并发送给核心权重计算单元662。加权系数确定单元762包括：训练数据分词子单元7622，用于对训练数据集内的各个商品短文本进行分词处理；训练数据标注子单元7624，用于标注训练数据集内各个商品短文本的核心词和非核心词；分词权重计算子单元7626，用于计算训练数据集内各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重；机器学习子单元7628，用于将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重，采用机器学习的方法计算局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数。

[0137] 通过采用加权系数确定单元，当根据多个权重确定商品短文本中的核心词时，能够调节各个权重之间的比重，提升核心词确定的准确性。

[0138] 此外，预处理模块52还可以包括数据清理单元722，用于去掉商品短文本中的停用词和标点符号，并将处理后的商品短文本发送给分词单元524。从而，可以减少对非必要词汇的处理，提升后续处理的效率。

[0139] 此外，预处理模块52还可以包括分词过滤单元724，用于在对商品短文本进行分词处理之后，统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词。从而，可以提升后续核心词的提取效率。

[0140] 此外，根据本发明的方法还可以实现为一种计算机程序产品，该计算机程序产品包括计算机可读介质，在该计算机可读介质上存储有用于执行本发明的方法中限定的上述功能的计算机程序。本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

[0141] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

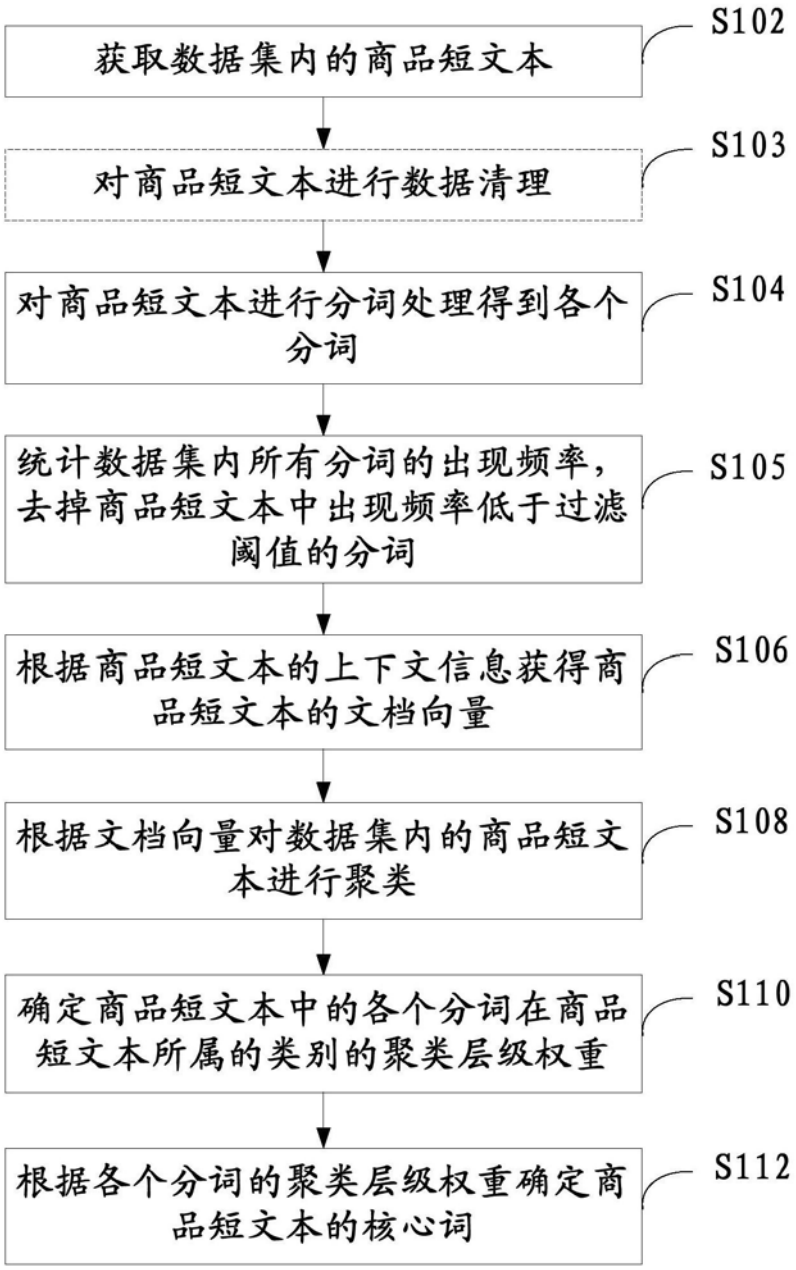
S102
获取数据集内的商品短文本
S103
对商品短文本进行数据清理
S104
对商品短文本进行分词处理得到各个分词
S105
统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词
S106
根据商品短文本的上下文信息获得商品短文本的文档向量
S108
根据文档向量对数据集内的商品短文本进行聚类
S110
确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重
S112
根据各个分词的聚类层级权重确定商品短文本的核心词

图1

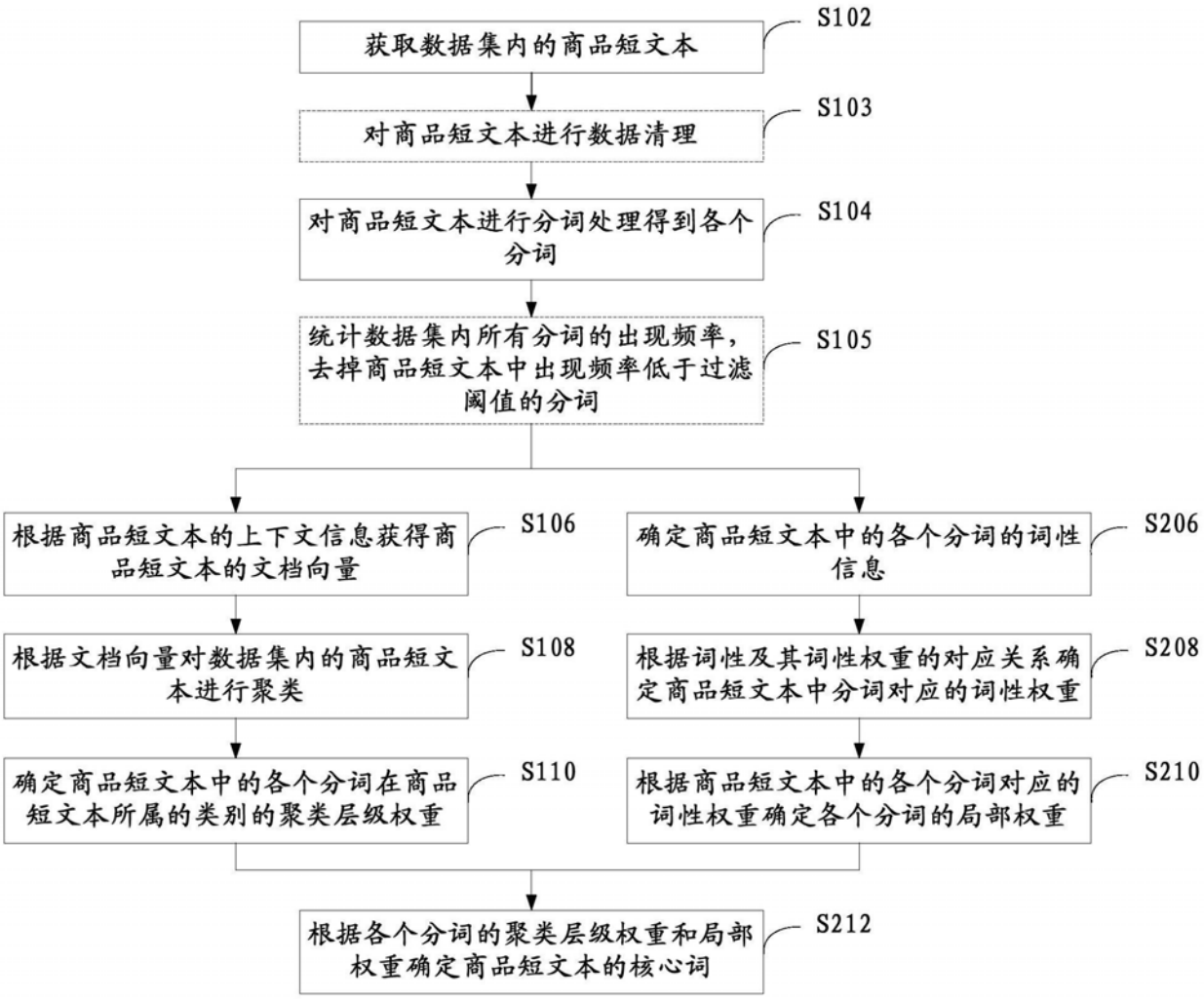
获取数据集内的商品短文本
S102
对商品短文本进行数据清理
S103
对商品短文本进行分词处理得到各个分词
S104
统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词
S105
根据商品短文本的上下文信息获得商品短文本的文档向量
S106
确定商品短文本中的各个分词的词性信息
S206
根据文档向量对数据集内的商品短文本进行聚类
S108
根据词性及其词性权重的对应关系确定商品短文本中分词对应的词性权重
S208
确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重
S110
根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重
S210
根据各个分词的聚类层级权重和局部权重确定商品短文本的核心词
S212

图2

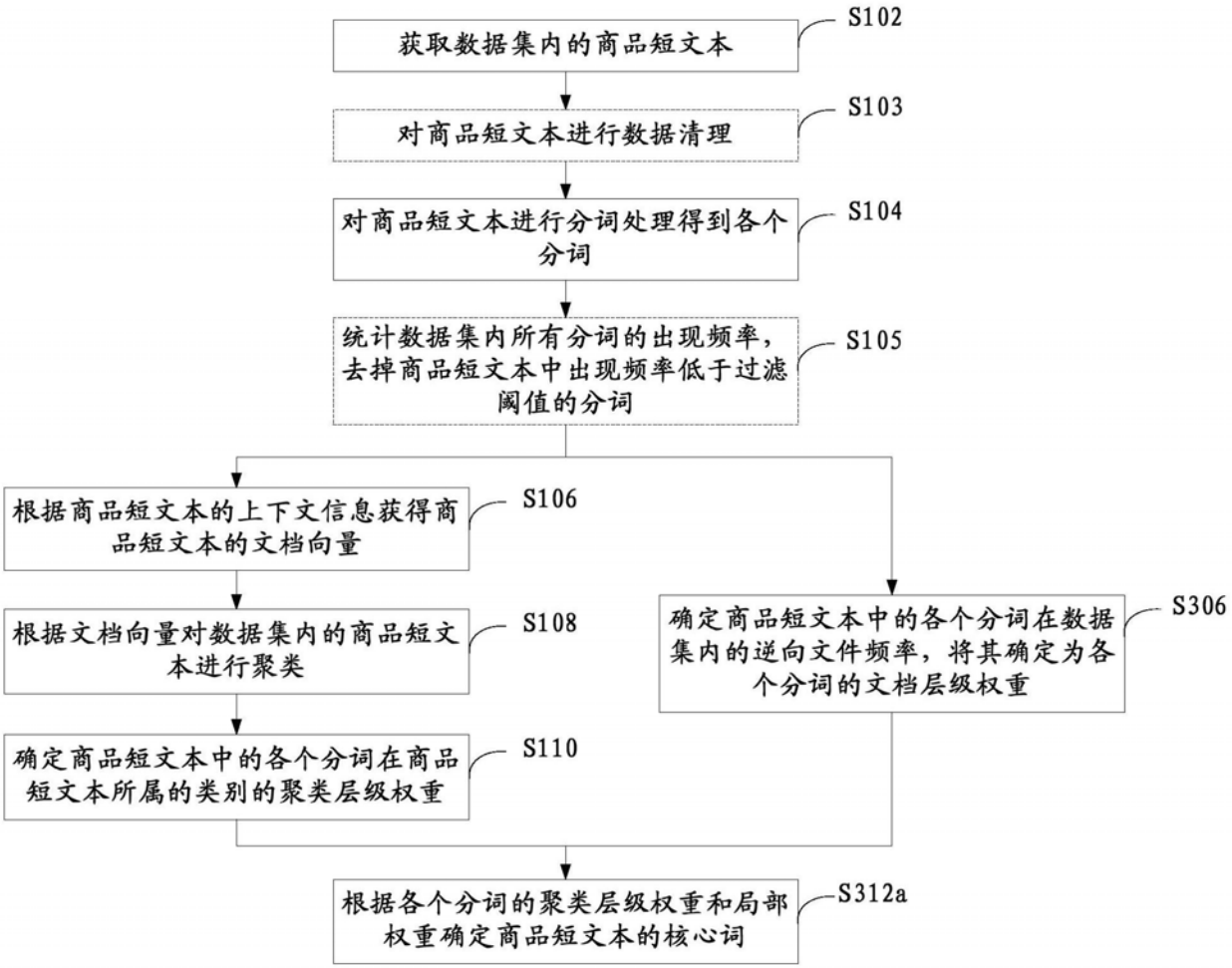
获取数据集内的商品短文本
S102
对商品短文本进行数据清理
S103
对商品短文本进行分词处理得到各个分词
S104
统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词
S105
根据商品短文本的上下文信息获得商品短文本的文档向量
S106
根据文档向量对数据集内的商品短文本进行聚类
S108
确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重
S110
确定商品短文本中的各个分词在数据集内的逆向文件频率，将其确定为各个分词的文档层级权重
S306
根据各个分词的聚类层级权重和局部权重确定商品短文本的核心词
S312a

图3A

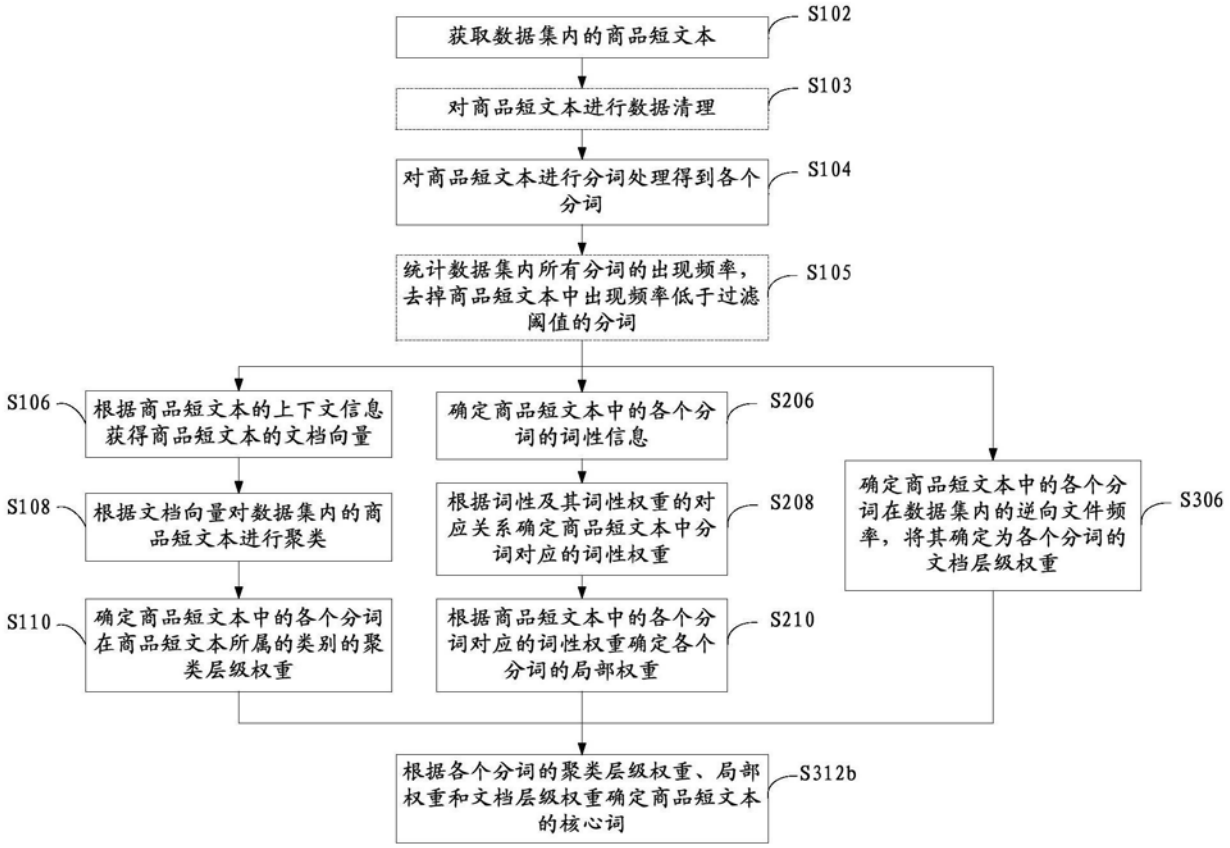
获取数据集内的商品短文本
S102
对商品短文本进行数据清理
S103
对商品短文本进行分词处理得到各个分词
S104
统计数据集内所有分词的出现频率，去掉商品短文本中出现频率低于过滤阈值的分词
S105
S106
根据商品短文本的上下文信息获得商品短文本的文档向量
S108
根据文档向量对数据集内的商品短文本进行聚类
S110
确定商品短文本中的各个分词在商品短文本所属的类别的聚类层级权重
S206
确定商品短文本中的各个分词的词性信息
S208
根据词性及其词性权重的对应关系确定商品短文本中分词对应的词性权重
S210
根据商品短文本中的各个分词对应的词性权重确定各个分词的局部权重
S306
确定商品短文本中的各个分词在数据集内的逆向文件频率，将其确定为各个分词的文档层级权重
S312b
根据各个分词的聚类层级权重、局部权重和文档层级权重确定商品短文本的核心词

图3B

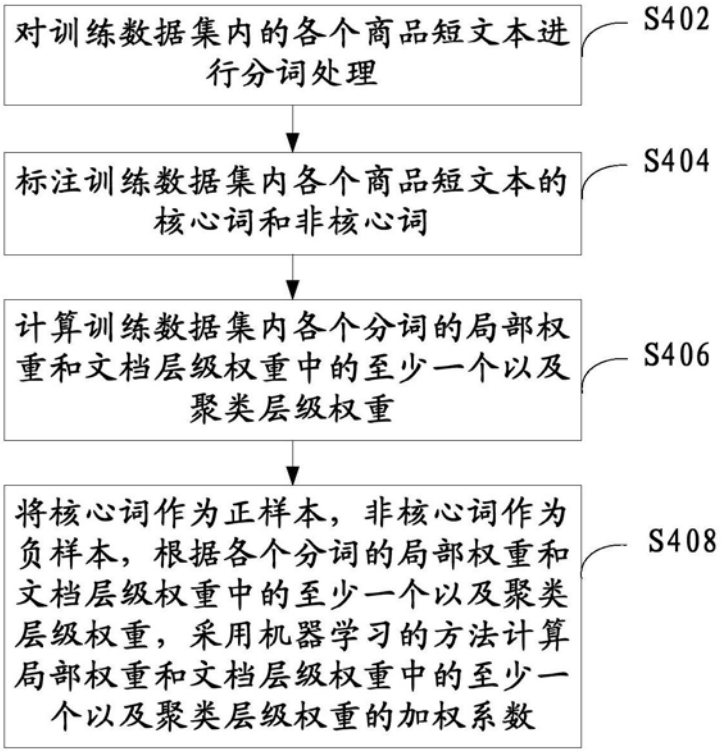
对训练数据集内的各个商品短文本进行分词处理
S402
标注训练数据集内各个商品短文本的核心词和非核心词
S404
计算训练数据集内各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重
S406
将核心词作为正样本，非核心词作为负样本，根据各个分词的局部权重和文档层级权重中的至少一个以及聚类层级权重，采用机器学习的方法计算局部权重和文档层级权重中的至少一个以及聚类层级权重的加权系数
S408

图4

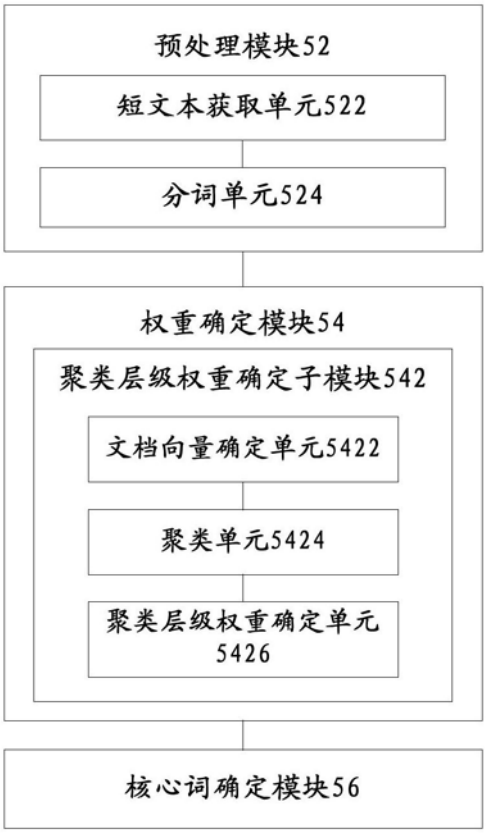
预处理模块52
短文本获取单元522
分词单元524
权重确定模块54
聚类层级权重确定子模块542
文档向量确定单元5422
聚类单元5424
聚类层级权重确定单元5426
核心词确定模块56

图5


预处理模块52
短文本获取单元522
分词单元524
权重确定模块54
聚类层级权重确定子模块542
文档向量确定单元5422
聚类单元5424
聚类层级权重确定单元5426
局部权重确定子模块644
词性信息确定单元6442
词性权重确定单元6444
局部权重确定单元6446
文档层级权重确定子模块646
核心词确定模块56
核心权重计算单元662
核心词确定单元664

图6


预处理模块52
短文本获取单元522
数据清理单元722
分词单元524
分词过滤单元724
权重确定模块54
聚类层级权重确定子模块542
文档向量确定单元5422
聚类单元5424
聚类层级权重确定单元5426
文档层级权重确定子模块646
局部权重确定子模块644
词性信息确定单元6442
词性权重确定单元6444
局部权重确定单元6446
第一词性权重对应单元7442
第二词性权重对应单元7444
核心词确定模块56
加权系数确定单元762
训练数据分词子单元7622
训练数据标注子单元7624
分词权重计算子单元7626
机器学习子单元7628
核心权重计算单元662
核心词确定单元664

图7