(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109062895 B
(45)授权公告日 2022.06.24

(21)申请号 201810813702.X

(22)申请日 2018.07.23

(65)同一申请的已公布的文献号
申请公布号 CN 109062895 A

(43)申请公布日 2018.12.21

(73)专利权人 挖财网络技术有限公司
地址 310000 浙江省杭州市西湖区华星路96号第18层

(72)发明人 康洪雨 尤志强 车曦 潘琪

(74)专利代理机构 杭州裕阳联合专利代理有限公司 33289
专利代理师 姚宇吉

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/216(2020.01)
G06F 40/284(2020.01)
G06F 16/35(2019.01)
G06F 16/33(2019.01)

(56)对比文件
CN 101067808 A,2007.11.07
CN 106202042 A,2016.12.07
CN 108170666 A,2018.06.15
CN 107526792 A,2017.12.29
CN 106682123 A,2017.05.17
CN 107885717 A,2018.04.06
CN 102779119 A,2012.11.14
CN 101719129 A,2010.06.02
CN 106557460 A,2017.04.05
刘勘等.基于关键词的科技文献聚类研究.《图书情报工作》.2012,第56卷(第4期),第6-11页.
刘啸剑等.结合主题分布与统计特征的关键词抽取方法.《计算机工程》.2017,第43卷(第7期),第217-222页.
钱爱兵等.基于改进TF-IDF的中文网页关键词抽取——以新闻网页为例.《情报理论与实践》.2008,第31卷(第06期),第945-950页.

审查员 刘莲花

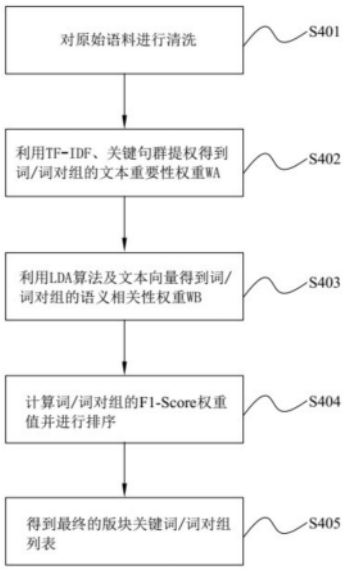
权利要求书2页 说明书8页 附图5页

(54)发明名称

一种智能语义处理方法

(57)摘要

本发明公开了一种智能语义处理方法,包括:获取语料中的关键句群信息,语料包括分类信息与正文信息;根据语料的分类信息,对语料中的词/词对组进行TF-IDF计算;将词/词对组与TF-IDF的计算结果对应;根据关键句群修正词/词对组的TF-IDF计算结果。本发明的有益之处在于:提供了一种智能语义处理方法。



CN 109062895 B

1.一种智能语义处理方法,包括:

获取各语料中的关键句群信息,所述语料包括分类信息与正文信息;

根据各语料的分类信息,对语料中的词/词对组进行TF-IDF计算;

将TF-IDF的计算结果匹配至对应的词/词对组;

根据所述关键句群对词/词对组的TF-IDF的计算结果进行结果修正,获得关键词候选组;

基于所述分类信息,将所有语料的关键词候选组中所包含的词/词对组进行合并,获得对应分类的关键词候选集;

计算每个分类对应的关键词候选集中每个词/词对组的位置权重分值之和,并按照这个分值进行降序排列,获得关键词集;

获得关键词候选组的具体步骤为:

按照TF-IDF值由大到小的顺序对词/词对组进行降序排列,并取前x个词作为待处理语料中的候选关键词;

根据词/词对组与关键句群之间的关系,对TF-IDF排序之后的候选关键词进行重排序调整,获得关键词候选组;

计算位置权重分值的具体步骤为:

计算每个关键词候选组中各词/词对组的位置权重分值,计算公式为

$$w_L = \sum_{i}^{k} \frac{1}{\log_2(i+1)}$$

其中k表示包含该词/词对组L的语料的数量,i为对应词/词对组L在某个语料所对应的关键词候选组中的排序位置。

2.根据权利要求1所述的方法,其特征在于:

所述语料为经过清洗处理的语料,所述清洗处理包括:

对原始语料进行分词,得到候选词集;

识别候选词集中有意义的词/词对组。

3.根据权利要求1所述的方法,其特征在于,所述结果修正包括:

若该词/词对组出现在所述关键句群中,对该词/词对组的TF-IDF计算结果增加预设数值,得到修正结果。

4.根据权利要求1所述的方法,其特征在于,所述结果修正包括:

根据所述TF-IDF计算结果对词/词对组的进行排序,标记各个词/词对组的序列值,所述序列值与该词/词对组根据所述TF-IDF计算结果排序的序列相关;

若该词/词对组出现在所述关键句群中,则该词/词对组的序列值减少预设数值。

5.根据权利要求1所述的方法,其特征在于,所述结果修正包括:

根据所述TF-IDF计算结果,按照从大到小的顺序,对词/词对组进行排序;

若该词/词对组出现在所述关键句群中,则对该词/词对组进行提权。

6.根据权利要求2所述的方法,其特征在于,对原始语料进行分词,包括:对分词的结果进行筛选,筛除结果中的停顿词。

7.根据权利要求2所述的方法,其特征在于,识别候选词集中有意义的词/词对组,包

括：

基于整个语料库，构建相邻的K个词对组；

统计非空类词/词对组的数量，将它们编入频次字典；

从频次字典中删除出现次数少于预设次数的词/词对组；

计算频次字典包含的剩余的词/词对组的成分值；对词/词对组进行阈值过滤。

8.根据权利要求1所述的方法，其特征在于，利用Text Rank、Textsum、Lex Rank中的一种或几种算法对语料进行识别得到所述关键句群信息。

9.根据权利要求1所述的方法，其特征在于，所述分类信息包括网上论坛或/和社区的版块信息。

10.一种智能语义处理方法，包括：

基于权利要求1至9任意一项所述的方法，获取某一分类的关键词候选集，并获取关键词候选集每个词/词对组的位置权重分值之和，将所述词/词对组作为关键词；

获取该分类关键词候选集中关键词的文本重要性权重WA，所述文本重要性权重WA为所述位置权重分值之和；

获取该分类关键词候选集中关键词的词语义重要性权重WB；

利用F1-Score对该分类下的关键词进行计算；

根据计算结果的顺序，取预设数量的关键词，作为该分类的关键词列表。

一种智能语义处理方法

技术领域

[0001] 本发明涉及一种智能语义处理方法。

背景技术

[0002] 现有的语义处理方法存在以下问题：

[0003] (1)中文文档进行分词目前工具较多，每个工具有不同的适应性和效果，效果不好的分词工具所得到的词语效果也不好，比如电脑被分成了“电”和“脑”，拿这样的词作为关键词的候选词集，所得到的关键词也是不准确的，反而引入了噪音。

[0004] (2)文档经过切词之后直接进行下一步词性判断这一过程考虑不够充分，因为有些词经过拼接之后会变成词组或者短语，比如“机器”和“学习”拼成“机器学习”，拼接之后和拼接之前意义完全不同。不进行短语识别对关键词提取效果影响很大，也会影响关键词与主题的相关性。

[0005] (3)没有进行词性标注与词性筛选，比如介词、形容词、代词等都会被保留，这样会有很多根本不会是关键词的词语参与到词性标注及后续的权值计算当中，增加了噪音，浪费计算资源。

[0006] (4)单个字的词义表达效果一般弱于两字及以上，作为候选关键词也是不合适，而且单个字也可能是因为分词效果欠佳导致产生，作为关键词会产生误解。

[0007] (5)单一使用TFIDF算法，会存在问题。IDF的引入，其初衷是抑制某一文档内无意义高频词的负面影响，但是在总文档于关键词出现文档比值较大时，低频词将因此而被凸现出来，常见词并不等于无意义词，比如一些公众人物，热点事件等等，同样的，低频词的偶然出现将被当作高权值关键词，这过渡放大了生僻词的重要性。因此必须对输出的结果进行多信息融合调整。

[0008] (6)TFIDF算法没有体现出位置信息的区分性，对于出现在文章不同位置的词语都是一视同仁的，而我们知道，在文章重要句群中出现的词语势必重要性要相对高点。将处于文章不同位置的词语赋予不同的权重才是合理的。

[0009] (7)仅仅考虑了统计信息，并没有将文本语义信息融合进来，没有从语义层面，考虑关键词与文章主题的相关性，一方面会影响关键词与文章的契合度，另一方面容易丢失一些语义上含义相似的词语。

发明内容

[0010] 一种智能语义处理方法，包括：获取语料中的关键句群信息，语料包括分类信息与正文信息；根据语料的分类信息，对语料中的词/词对组进行TF-IDF计算；将词/词对组与TF-IDF的计算结果对应；根据关键句群修正词/词对组的TF-IDF计算结果。

[0011] 进一步的，语料为清洗过的语料，清洗过程包括：对原始语料进行分词，得到候选词集；识别候选词集中有意义的词/词对组。

[0012] 进一步的，根据关键句群修正词/词对组的TF-IDF计算结果，包括：

[0013] 若该词/词对组出现在关键句群中，对该词/词对组的TF-IDF计算结果+1，得到修正结果；重复上述步骤至遍历每个词/词对组与关键句群。

[0014] 进一步的，根据关键句群修正词/词对组的TF-IDF计算结果，包括：根据TF-IDF计算结果，按照从大到小的顺序，对词/词对组的进行排序，并标记各个词/词对组的序列值；序列值为该词/词对组的排序；若该词/词对组出现在关键句群中，则该词/词对组的序列值-1，重复该步骤至遍历每个词/词对组与关键句群；按照最终序列值对词/词对组进行排序。

[0015] 进一步的，根据关键句群修正词/词对组的TF-IDF计算结果，包括：根据TF-IDF计算结果，按照从大到小的顺序，对词/词对组的进行排序；若该词/词对组出现在关键句群中，则对该词/词对组进行提权，重复该步骤至遍历每个词/词对组与关键句群。

[0016] 进一步的，语料进行分词，得到候选词集，包括：对候选词集进行筛选，筛除候选词集中的停顿词。

[0017] 进一步的，识别候选词集中有意义的词/词对组，包括：基于整个语料库，构建相邻K个词对组；统计不包含空词的词对组(pair)以及词出现的次数，形成词/词对组的频次字典(vocab)；从vocab中删除出现次数少于一定次数的词/词对组；计算vocab包含的词/词对组的成分值；对词/词对组进行阈值过滤。

[0018] 进一步的，对词/词对组进行阈值过滤，包括：对这些词/词对组进行词性识别与筛选，保留包含具体含义的实体词，删除语气词、副词、形容词。

[0019] 进一步的，关键句群信息，包括利用Text Rank、Textsum、LexRank等算法对语料进行识别得到的关键句群信息。

[0020] 进一步的，分类信息为网上论坛、社区的版块信息。

[0021] 一种智能语义处理方法，包括：获取某一分类的关键词候选集；获取该分类关键词候选集的关键词文本重要性权重WA；获取该分类关键词候选集的关键词的词语义重要性权重WB；利用F1-Score对该分类下的关键词进行计算；根据计算结果的顺序，取一定数量的关键词，作为该分类的关键词列表

[0022] 本发明的有益之处在于：提供了一种智能语义处理方法。

附图说明

[0023] 图1是本发明对语料进行清洗的流程示意图；

[0024] 图2是本发明识别有意义的短语、词组及单个存在的词的流程示意图；

[0025] 图3是本发明根据关键句群对词/词对组进行关键度排序的流程示意图；

[0026] 图4是本发明获取某分类最相关的词/词对组的流程示意图；

[0027] 图5是本发明整体系统运行流程示意图。

具体实施方式

[0028] 如图1所示为本发明的智能语义提取方法中对于待处理语料的语义清洗方法的流程图。

[0029] 在步骤S101，输入待处理语料。待处理语料采用的文字通常为中文，可以包括数字、符号、少量英文等其他方式表示的语料。输入的语料内容上包括分类名称与正文内容。

在本实施例中，输入的语料来源于网上论坛或社区的发帖。其中，分类名称即为该帖子所属的论坛版块名称，正文内容即为该帖子的正文内容。作为可选的实施方式，正文内容还可以包括该帖子的标题。

[0030] 在步骤S102，将在S101中输入的待处理语料进行分词，即将汉字序列切分成多个词。作为可选的实施方式，分词的方法可以是基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法中的一种或多种。作为可选的实施方式，分词采用的工具可以是pyltp、jieba、hanlp、snowNLP中的一种或多种。在本实施例中，采用pyltp对待处理语料进行分词处理。

[0031] 在得到分词后的词组合后，对该词组合中的词进行筛选，筛掉常见的停顿词。

[0032] 待处理语料经过分词步骤，输出每组待处理语料所对应的候选词集。

[0033] 在步骤S103，在通过S102得到待处理语料所对应的候选词集之后，对候选词集中的组成元素进行短语和词组的识别。

[0034] 有些短语和词组在文档中具有实际意义，但是分词的时候由于分词工具的限制，无法准确切分，会把短词和词组进行切分成词语，甚至是把少数有效的词语切分成单独的两个字。为了还原文档的准确意义，作为一种实施方式，如图2所示，采用如下方法对候选词集进行短语和词组的识别：

[0035] 为了清楚地描述本短语、词组识别方法的原理，假设在步骤S2中输出的候选词集包括两条文本，分别为[w1,w2,w3,w4,w5]、[w2,w3,w6]

[0036] 在步骤S1031，基于整个语料库，构建相邻K个词对组。作为一种可选的实施方式，K取2。在K取2的情况下，构建的词对组(pair)列表分别为[(w1,w2),(w2,w3),(w3,w4),(w4,w5),(w5,None)]、[(w2,w3),(w3,w6),(w6,None)]。

[0037] 在步骤S1032，统计不包含None的词对组(pair)以及词出现的次数形成词对组/词的频次字典(vocab)。在K取2时，vocab为{(w1,w2):1,(w2,w3):2,(w3,w4):1,(w4,w5):1,(w3,w6):1,(w1):1,(w2):2,(w3):2,(w4):1,(w5):1,(w6):1}。

[0038] 同时，统计总的词对数(包含None在内，但不包含单个词)，将总的词对数标记为train_words。在本实施例中，train_words为8。

[0039] 在步骤S1033，从vocab中删除出现次数少于一定次数的词或词对组。将出现次数阈值定义为min_count。在本实施例中，min_count设定为1，此时，最终得到的vocab为[(w2,w3):2,(w2):2,(w3):2]。

[0040] 在步骤S1034，按照公式score＝(pab-min_count)*train_words/(pa*pb)计算S1033中得到的vocab包含的词对组(pair)的成分值。其中，pab为该pair通过查询vocab得到的出现次数，在本实施例中，词对组(w2,w3)在待处理语料中出现的次数为2，亦即该词对组的pab＝0。其中，pa为词对组中第一个词的出现次数，pb为词对组中第二个词出现的次数，即w2的出现次数对应pa，w3的次数对应pb。通过查询vocab，可知pa为2，pb为2。那么词对组(w2,w3)能合成w2w3的score(以该例展示)：(2-1)*8/(2*2)＝2。如果pab，pa，pb其中任一值缺失，则score为0。

[0041] 在步骤S1035，对步骤S1034中得到的score进行阈值过滤：如果score>threshold，那么我们认为该词对组属于有意义的词对组。在本实施例中，设threshold为1(实际中会更大，比如threshold＝100)。此时，由于score(w2,w3)>threshold，则判断w2与w3能够组成词

组。如果score(w2,w3)≤threshold,则判断w2与w3不能组成词组,则在对该语句进行理解时,w2、w3应当被当做单个存在的词。

[0042] 在步骤S1036,通过以上步骤判断得到有意义的词对组、单个存在的词后,将原始的语料根据判断结果进行重新分类,得到每组待处理语料对应的词语和短语的集合。

[0043] 至此,步骤S103中对候选词集中的组成元素进行短语和词组的一种识别方法执行完毕。

[0044] 在步骤S104,对S103中得到的短语和词组集中的元素进行词性识别与筛选,保留名词、动词、动名词、机构缩写等包含具体含义的实体词。删除语气词、副词、形容词等词性的词语。作为一种可选的实施方式,词性识别、筛选的过程可以通过pyltp软件的词性标注功能实现。

[0045] 在步骤S105,对步骤S104中得到待处理语料进行词长过滤:经过步骤S104的处理,能够去除掉部分词长过短的或非实体的词汇,比如“是”、“很”等,但依然可能存在其他的单字。在实际的语言表达中,长度为1的词往往不能准确表达文章主旨。而这样的词往往会在机器理解语言含义时引入噪音,很难成为有效的关键词。该步骤能够把词集中长度为1的词(即单字)删除,进一步减少噪声,提高机器理解语义的精度。

[0046] 如图4所示为本发明的智能语义提取方法中对于待处理语料得到语料中关键句群识别,并根据关键句识别该语料对应的最相关关键词/词对组的流程图。作为一种一般的处理方式,该对关键句群的识别方法可以应用在未经清洗处理的语料中。

[0047] 在步骤S201,对待处理语料中的每个元素(这里的元素包括词/词对组)进行版块下的TF-IDF计算。

[0048] 其中,TF-IDF=TF*IDF,由词频与逆文档频率两部分组成:

[0049] 词频(term frequency,TF)指的是某一个给定的词/词对组在对应版块中出现的频率,计算公式如下:

[0050] $$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0051] 公式中的分子是每个词在对应分类下所有文档中出现的次数,分母是对应分类中所有文档的词/词对组总数。由于本实施例的待处理语料来源于线上论坛或社区,在本实施例中,分子即是每个词/词对组在该线上论坛或社区的对应版块下所有文档中出现的次数,分母是对应版块中所有文档的词/词对组的总数。

[0052] 在本实施例中,公式中的参数$tf_{i,j}$表示词/词对组i在主题j中的次数,$n_{i,j}$表示词/词对组i在版块j中出现的次数,$n_{k,j}$表示词/词对组k在版块j中出现的次数。也就是说说,分母表示所有词/词对组在版块j中出现的次数,即版块j中词总数。

[0053] 逆文档频率(inverse document frequency,IDF)是一个词语普遍重要性的度量。在本实施例中,某一个词/词对组的IDF可以由总版块数目除以包含该词/词对组的版块数目,再将结果取对数得到:

[0054] $$\mathrm{idf}_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

[0055] 上式中分子|D|是总的版块数,分母是包含词i的版块数目(若版块总数如果为1,

则不进行计算idf值)。TF-IDF的权重值一定程度上体现出词/词对组与版块的关联度。

[0056] 根据该公式,可以得到待处理语料里所有词/词对组所对应的TF-IDF值,按照TF-IDF值由大到小的顺序对词/词对组进行降序排列,并取前x个词作为待处理语料中的候选关键词。在本实施例中,作为一种可选的实施方式,x取5。

[0057] 在步骤S202,定位关键句群。在步骤S201中,通过TF-IDF法筛选,得到每篇文档的5个词语作为关键词候选集,并按照TF-IDF值由大到小排序。

[0058] 通过TF-IDF值进行排序的词/词对组,没有考虑该词/词对组在待处理语料中所处的位置的重要性:若该词/词对组来自标题、摘要、体现主旨意思的句子,该词/词对组则应该获得更高的权重。相比之下,出现在其他次要的语句中的词/词对组的权重则应当相对较低。

[0059] 为了提高通用性,即使没有明确的标题、摘要等段落的情况下,依然能够奏效,我们引入了Text Rank算法来进行文档中关键句群识别模块。

[0060] 关键句群即是文章中体现主旨主题的关键句子,能够反映一篇文章的核心意思。如果关键词来自于这些关键句群中,其权重相对于来自其他位置将会更高。以凸显出关键位置的重要性。

[0061] Text Rank是由Page Rank改进而来。将每个句子看成图中的一个节点,若两个句子之间有相似性,认为对应的两个节点之间有一个无向有权边,权值是相似度。通过Text Rank算法计算得到的重要性最高的若干句子可以当作关键句群。

[0062] 作为可选的实施方式,识别关键句群的方法还可以是Textsum、LexRank等算法。

[0063] 在步骤S203,根据词/词对组与通过Text Rank得到的关键句群之间的关系,对TF-IDF排序之后的关键词进行重排序调整。具体方法包括:

[0064] 待处理语料的关键词候选集按照TF-IDF权重降序排列,取该降序排列序列中排名前5的词/词对组,作为候选词/词对组;

[0065] 利用关键句群集合数据,查询每一个候选词/词对组是否在关键句群中出现。

[0066] 如果一个词/词对组出现在该待处理语料的关键句群中,此时,触发词/词对组提权机制。作为一种可选的实施方式,提权方法包括:对出现在关键句群中的词/词对组的排序位置序列号-1,再按照位置序号数从小到大的顺序排列。这种计算结果体现在具体排序中,也就是该词/词对组向前移动一位。

[0067] 采用数码标记排序位置的方法,在该词/词对组的排名已经在第一位的情况下也能够记录该词/词对组的提权数量,防止由于排序的“溢出”(排在第一位的词/词对组,如果只进行提权操作,该词/词对组仍然排在第一位,相当于该提权操作的结果被“溢出”了)。

[0068] 如果5个词/词对组都在关键句群中,则每一个词/词对组的排序号都需要-1(或保持不变),亦即该5个词/词对组保持原有的TF-IDF排序不变。

[0069] 举例来说,假设一篇文章TF-IDF权重降序排列后候选关键词为[w1,w2,w3,w4,w5]。在这种情况下,w1与w3出现在了关键句群中,那么经过排序序号数码处理后再进行排序,候选关键词顺序变为[w1,w3,w2,w4,w5],其中,w1、w3被“提权”。

[0070] 在步骤S204,对多个分类下的多组待处理语料重复S201-S203的步骤,得到多个经过提权排序的关键词候选组,每个关键词候选组中包含5个按顺序排列的词/词对组。

[0071] 在步骤S205,把同属于一个分类下面所有待处理语料的关键词候选组中所包含的

词/词对组合并到一起作为这个分类的关键词候选集，将每一个关键词候选集中的词/词对组排序。在本实施例中，分类及对应着在线论坛、社区的版块。排序的具体方法包括：

[0072] 计算每个待处理语料中得到5个词/词对组的位置权重，对同一分类下所有词语进行合并，每个词语的位置权重的计算公式如下：

[0073] $$\omega_L = \sum_{i}^{k} \frac{1}{log_2(i+1)}$$

[0074] 其中k表示包含该词/词对组L的待处理语料的数量，i为对应词/词对组L在某个待处理语料中的排序位置。这个排序位置就是经过关键句群重排序后的顺序，在本实施例中的取值范围为[1,2,3,4,5]。

[0075] 在S206，计算每个分类对应的关键词候选集中每个词/词对组的位置权重分值之和，并按照这个分值进行降序排列，得到每个版块下面的关键词集，至此，即得到了基于文本重要性刻画的关键词列表，以及列表中各个词/词对组对应的权重分WA。

[0076] 作为另一种根据TF-IDF计算结果与获取的关键句群信息识别该语料对应的最相关关键词/词对组的方法，步骤如下：

[0077] 对待处理语料中的每个元素(这里的元素包括词/词对组)进行版块下的TF-IDF计算，得到每个词/词对组的TF-IDF计算结果；获取该语料的关键句群信息；根据关键句群信息修正每个词/词对组的TF-IDF计算结果，得到每个词/词对组的修正结果；根据每个词/词对组的修正结果对该词/词对组进行排序。

[0078] 该方法在最终排序之前仍然考虑所有的词/词对组，因此，采用该方法得到的计算结果将更加精确。

[0079] 如图4所示，为本发明的智能语义提取方法中提取与某个分类最相关的词/词对组的方法流程图。判断的对象可以是通过清洗的语料，也可以是未经过清洗的原始语料。

[0080] 在S301，设置隐含主题，隐含主题的数量小于等于版块数，也就是：topic1，topic2，…，topicN，假设版块数为y，那么N≤y。

[0081] 在S302，通过LDA算法计算出词语(为了方便起见，词/词对组的概念在这里直接用“词语”表示)的主题分布。

[0082] 一篇文档是有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文章的第一个词。不断重复这个过程，就生成了整片文章。在LDA算法中，主题是隐含主题。

[0083] 选择其中权重最大的主题，就可以得到“词语—topic”配对数据，也就是词语对应所属的最可能的topic。

[0084] 在步骤S303，由于一篇文档中会包含多个topic，将一篇文档就转换成形式S，即(docid,word1:topic1,word2:topic2…,wordN:topicN)。之所以要用到主题topic，是为了将其作为文档、词之间的桥梁关系，更好得获取两者的内在联系。将三者信息放在一块进行学习向量表达，可以得到更好的效果。

[0085] 在步骤S304，将形式S的文档语料(即docid,word1:topic1,word2:topic2…,

wordN:topicN)通过Doc2vec算法进行计算,得到了每个docid、word和topic的向量,通过Doc2vec的计算过程,我们对docid,topic,word三者的共现关系加以利用,实现了在同一个空间中,对这三种实体进行向量化表征,可以得到docid:[0.1,0.2,0.12,0.3,0.13,0.5],topic1:[0.3,0.1,0.1,0.25,0.6,0.8],word1:[0.25,0.01,0.3,0.2,0.16,0.78]这样的向量形式,其中所有向量的维度都是一致的。那么每一个版块下的所有文档都可以得到docid:vector这样的结果。

[0086] 在步骤S305,对该版块下所有文档的docid的向量进行向量元素依次求和取平均,得到的平均后的向量作为版块主旨的向量。作为一种可选的实施方式,假设版块1中有doc1和doc2两篇文档,且doc1对应的docid向量为[0.1,0.3],doc2对应的docid向量为[0.4,0.32],那么版块1的主旨向量为[(0.1+0.4)/2,(0.3+0.32)/2],即[0.25,0.31]。

[0087] 在步骤S306,计算每个版块下候选关键词向量与版块主旨向量的内积,该内积的大小反应了词与版块的词义相关性,内积越大,说明词与版块主旨越相关。将该内积值作为该词的语义权重分值,至此,我们得到了基于语义模型的关键词集及其与对应主题的相关性权重WB。

[0088] 在步骤S307,通过F1-Score进行融合计算得到一个分值作为词语最终的分值,并将各个词语按照分值的降序进行排序。其中,F1-Score的计算公式如下:

[0089] $$F_1 = 2 \cdot \frac{WA \cdot WB}{WA + WB}$$

[0090] WA是词语在文本重要性模块中的权重分值;WB是词语在语义相关性模块下的权重分。

[0091] 在步骤S308,对对应版块下的每一个候选关键词经过F1-Score权重值降序排列后取前x个关键词,得到最终的版块关键词列表。

[0092] 将版块主旨与词语都进行向量化表达,可以很方便、快速地计算版块主旨与词语之间的相似程度。

[0093] 图5具体描述了针对某个网上社区自动化识别出每个版块对应的关键词/词对组的流程图。

[0094] 在步骤S401,对原始语料进行清洗。

[0095] 对语料的清洗步骤可以包括:对待处理语料进行分词处理;对分词后的语料进行筛选,筛除常用的停顿词;对剩下的语料内容进行词/词对组的识别,识别出有意义的短语和词组以及单个存在的词;对词/词对组进行词性识别与筛选,筛除语气词、副词、形容词等词性的词语;对剩下的词/词对组进行词长过滤,去除词长过短的或非实体的词汇。

[0096] 在步骤S402,得到词/词对组的文本重要性权重WA。

[0097] 获取WA的步骤可以包括:对词/词对组进行板块下的TF-IDF计算,并取x个词/词对组作为候选关键词;利用TextRank、Textsum、LexRank等算法对关键句群进行识别,并利用关键句群的识别结果,对TF-IDF得到的排序进行提权,得到新的排序;将同属于一个分类下面所有候选关键词合并,作为该分类的关键词候选集;计算每个分类对应的关键词候选集中每个词/词对组的位置权重分值之和,得到WA。

[0098] 在步骤S403,得到词/词对组的语义相关性权重WB。

[0099] 获取WB的步骤可以包括:设置N个数量的隐含主题;通过LDA算法计算出词/词对组

的主题分布;将文档转换为形式S;将形式S的文档通过Doc2vec算法,得到docid:vector这样的结果;该版块下所有文档的docid的向量进行向量元素依次求和取平均,得到的平均后的向量作为版块主旨的向量;计算每个版块下候选关键词向量与版块主旨向量的内积,得到语义相关性权重WB。

[0100] 在步骤S404,利用F1-Score进行融合计算,得到一个分值作为词/词对组最终的分值,并将各个词/词对组按照分值的降序进行排序。

[0101] 在步骤S405,针对某个版块按照顺序取前x个关键词,得到最终的版块关键词列表。

[0102] 由于版块的设置是根据业务需求设定,同时一个文档一般包含多个主题,往往会造成多个版块中文档的内容会在某些方面存在相似性,造成这些词与所属版块的相关性会较弱,不具有较好的区分性。比如“信贷”版块与“征信”版块,会有内容的交叉,但词语往往会在某个主题上表现出最高的倾向,将词语归类到最合适的版块,能够提升词语对版块的主旨刻画能力,同时也能带来更好的用户体验。

[0103] 以上显示和描述了本发明的基本原理、主要特征和优点。本行业的技术人员应该了解,上述实施例不以任何形式限制本发明,凡采用等同替换或等效变换的方式所获得的技术方案,均落在本发明的保护范围内。

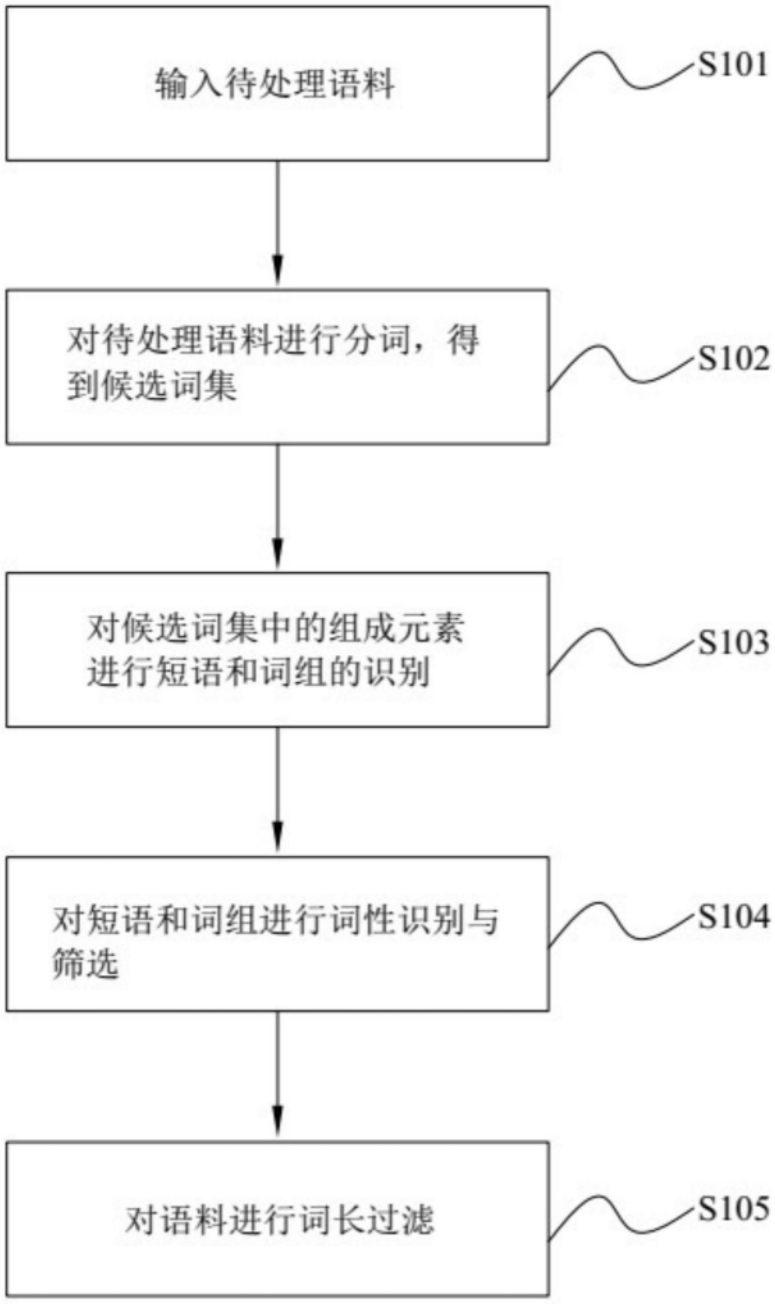

输入待处理语料
S101
对待处理语料进行分词，得到候选词集
S102
对候选词集中的组成元素进行短语和词组的识别
S103
对短语和词组进行词性识别与筛选
S104
对语料进行词长过滤
S105

图1

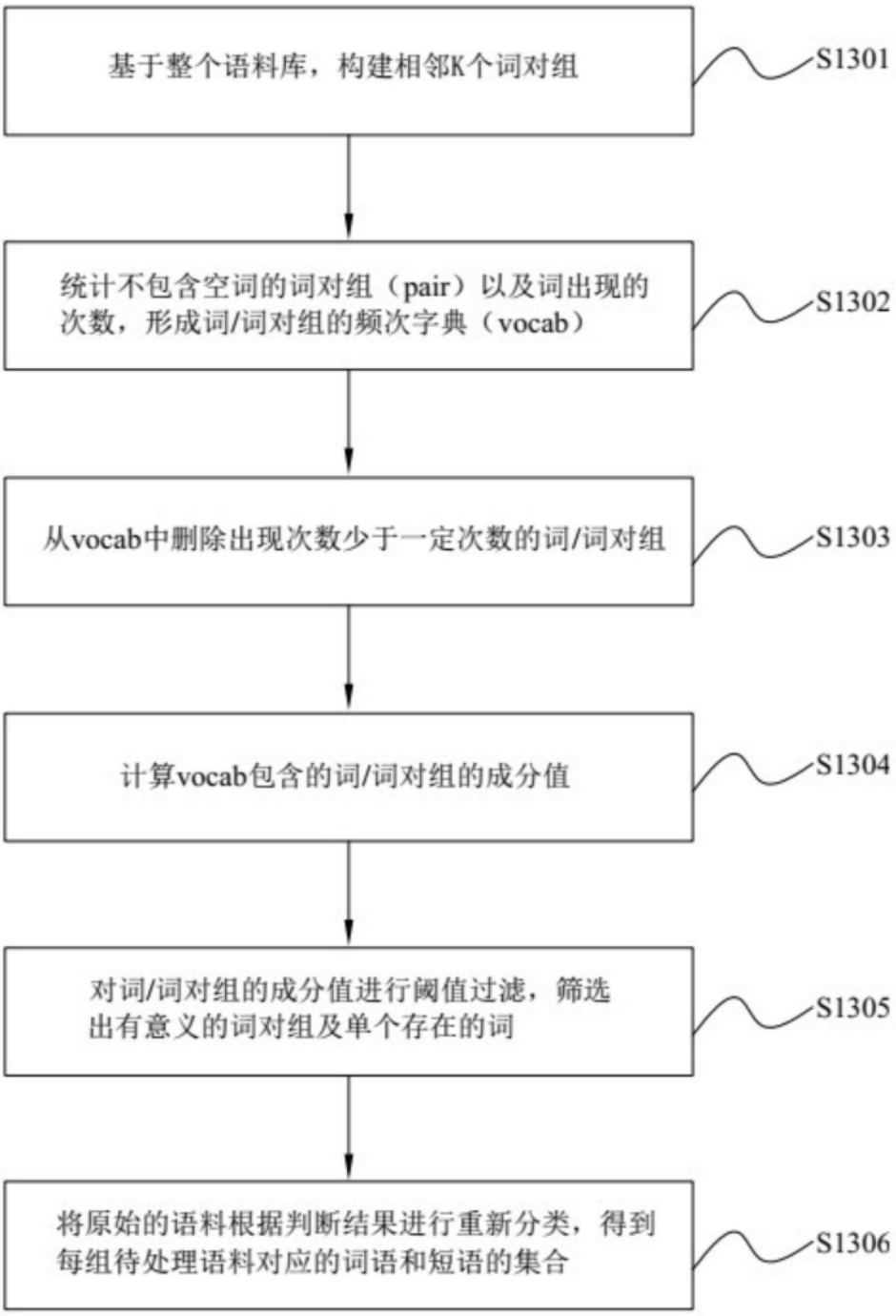

基于整个语料库，构建相邻K个词对组
S1301
统计不包含空词的词对组（pair）以及词出现的次数，形成词/词对组的频次字典（vocab）
S1302
从vocab中删除出现次数少于一定次数的词/词对组
S1303
计算vocab包含的词/词对组的成分值
S1304
对词/词对组的成分值进行阈值过滤，筛选出有意义的词对组及单个存在的词
S1305
将原始的语料根据判断结果进行重新分类，得到每组待处理语料对应的词语和短语的集合
S1306

图2

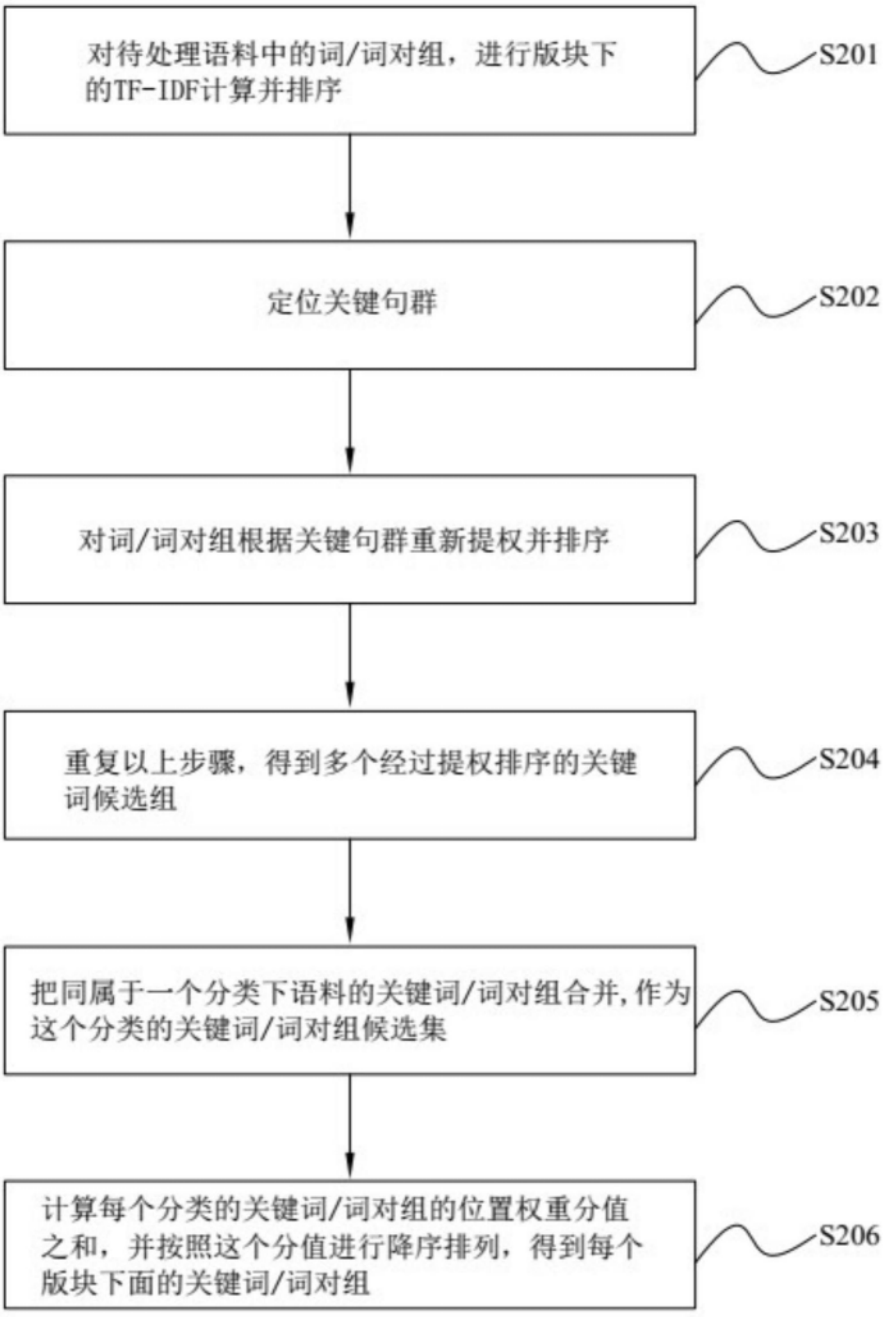

对待处理语料中的词/词对组，进行版块下的TF-IDF计算并排序
S201
定位关键句群
S202
对词/词对组根据关键句群重新提权并排序
S203
重复以上步骤，得到多个经过提权排序的关键词候选组
S204
把同属于一个分类下语料的关键词/词对组合并，作为这个分类的关键词/词对组候选集
S205
计算每个分类的关键词/词对组的位置权重分值之和，并按照这个分值进行降序排列，得到每个版块下面的关键词/词对组
S206

图3

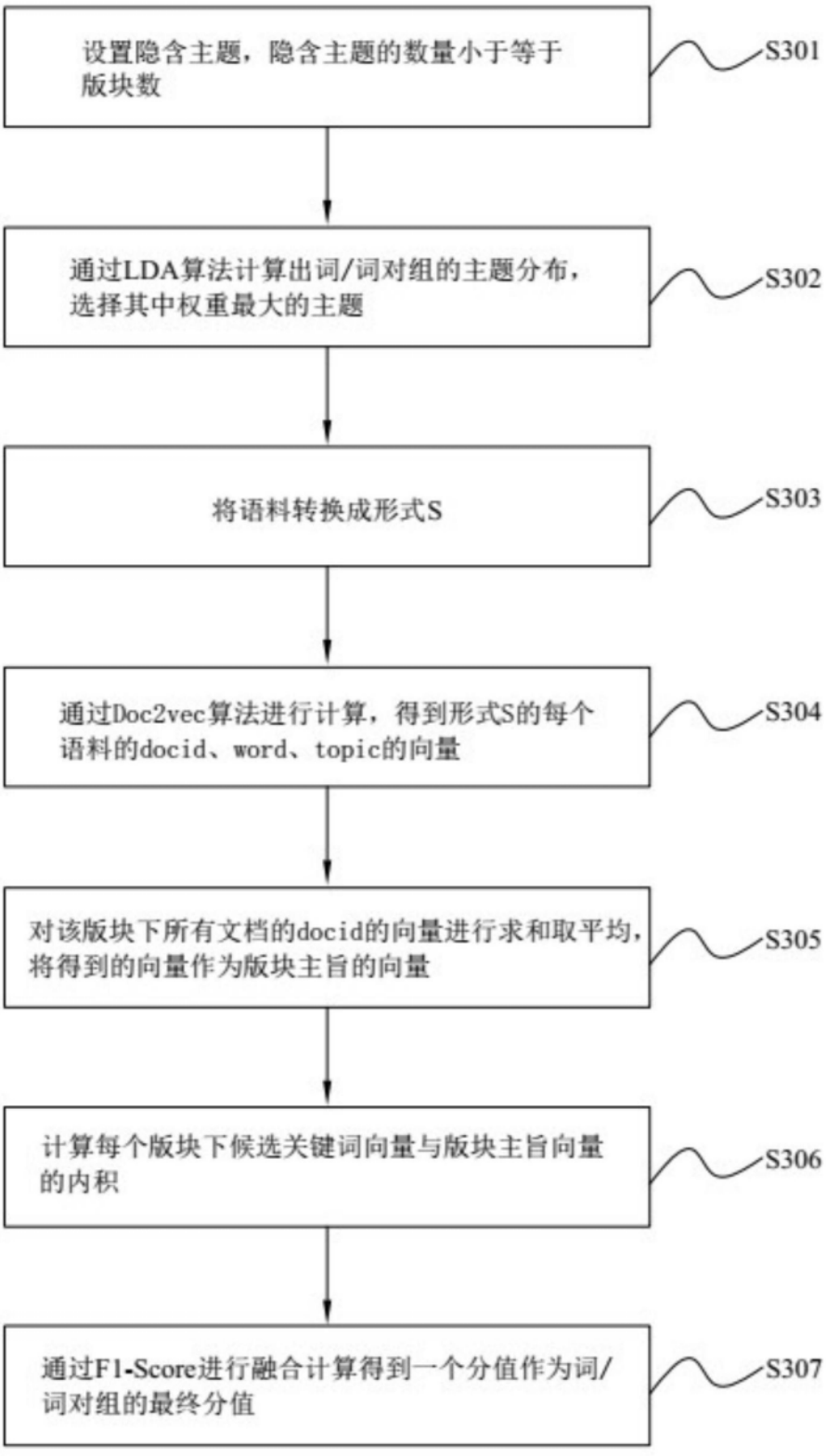

设置隐含主题，隐含主题的数量小于等于版块数
S301
通过LDA算法计算出词/词对组的主题分布，选择其中权重最大的主题
S302
将语料转换成形式S
S303
通过Doc2vec算法进行计算，得到形式S的每个语料的docid、word、topic的向量
S304
对该版块下所有文档的docid的向量进行求和取平均，将得到的向量作为版块主旨的向量
S305
计算每个版块下候选关键词向量与版块主旨向量的内积
S306
通过F1-Score进行融合计算得到一个分值作为词/词对组的最终分值
S307

图4

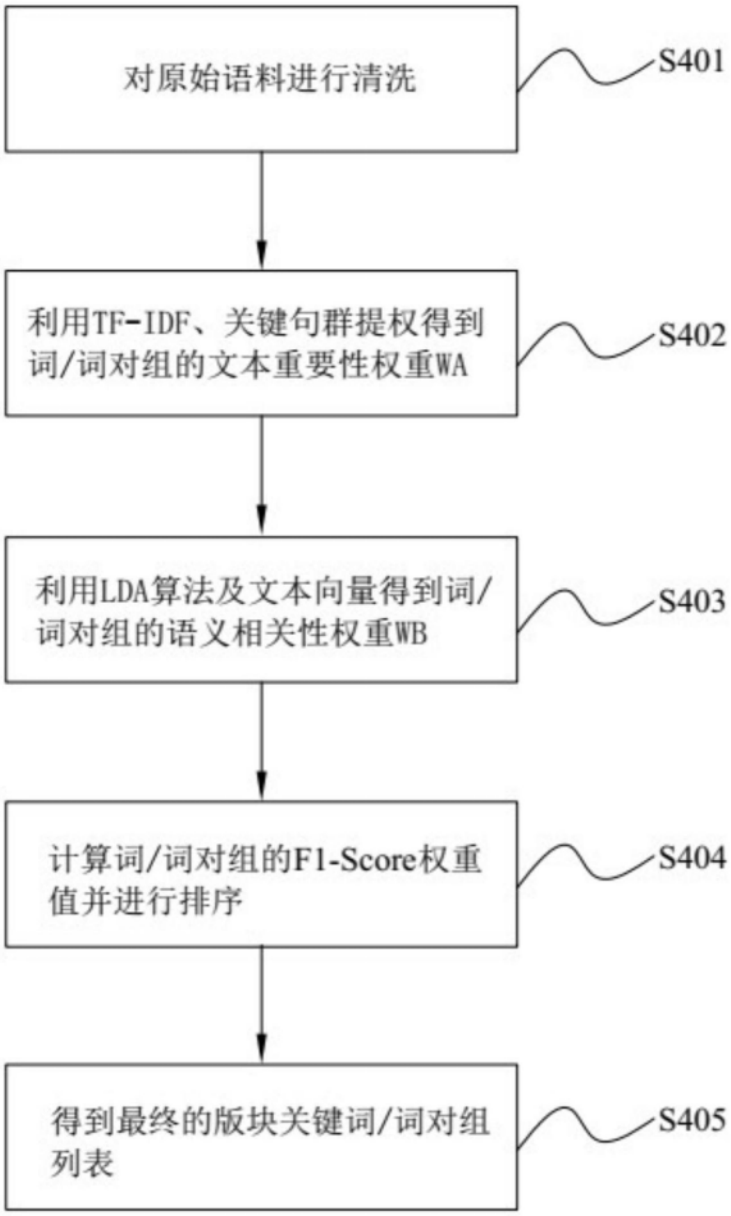

对原始语料进行清洗
S401
利用TF-IDF、关键句群提权得到词/词对组的文本重要性权重WA
S402
利用LDA算法及文本向量得到词/词对组的语义相关性权重WB
S403
计算词/词对组的F1-Score权重值并进行排序
S404
得到最终的版块关键词/词对组列表
S405

图5