(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 107133317 B
(45)授权公告日 2020.07.31

(21)申请号 201710304816.7

(22)申请日 2017.05.03

(65)同一申请的已公布的文献号
申请公布号 CN 107133317 A

(43)申请公布日 2017.09.05

(73)专利权人 成都云数未来信息科学有限公司
地址 610000 四川省成都市双流区西南航空港经济开发区电子科技大学成都研究院

(72)发明人 孙健 陆川 朱煜松

(74)专利代理机构 成都行之专利代理事务所（普通合伙）51220
代理人 温利平

(51)Int.Cl.
G06F 16/9535(2019.01)
G06F 40/284(2020.01)
G06F 16/33(2019.01)

(56)对比文件
CN 102750336 A,2012.10.24
CN 103955450 A,2014.07.30
CN 106339481 A,2017.01.18
CN 102043851 A,2011.05.04
林倩瑜.关联规则挖掘算法研究综述.《软件导刊》.2012,第11卷(第6期),第27-29页.
章博亨等.基于大数据和机器学习的微博用户行为分析系统.《电脑知识与技术》.2017,第13卷(第6期),第212-214页.

审查员 余佩玉

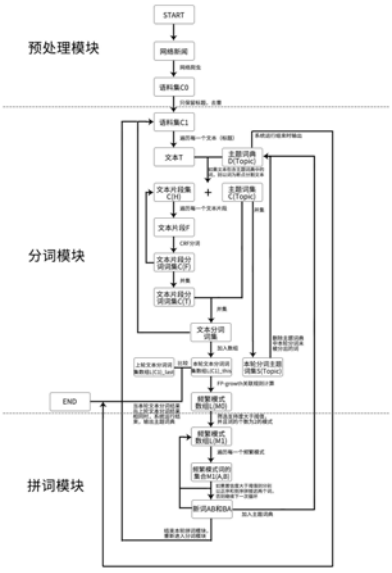
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种通过新词抽取网络舆情主题的方法

(57)摘要

本发明公开了一种基于新词的网络舆情主题抽取方法,通过两个或两个以上的关键词拼接得到的新词作为主题词;其中,对于关键词的选取,需要考虑代表文章的中心和主旨的文本标题,文本标题分得的词作为文本的关键词,同时还提高算法效率和减少因为文本差异对主题抽取的影响;同时,本发明以平均实词匹配度来量化这些新词对于网络舆情的主题的贡献程度,平均实词匹配度越高表示新词对于网络舆情的主题的贡献程度越大,因而,具有相当高的可信度,能够适应当今网络舆情日益增长的趋势。



CN 107133317 B

1.一种通过新词抽取网络舆情主题的方法,其特征在于,包括以下步骤:

(1)、构建一个空的主题词典;

(2)、利用网络爬虫从互联网中爬取网络新闻,对爬取的网络新闻进行去重处理后,将网络新闻的标题存入语料集中;

(3)、遍历语料集中的每一个网络新闻标题,利用CRF模型对网络新闻标题进行分词,把所有的分词结果存入到数组1中;

(4)、设置数据挖掘算法FP-growth的支持度阈值,利用数据挖掘算法FP-growth挖掘出数组1中的频繁项集,得到由频繁项集中的频繁项和对应支持度组成的频繁模式数组1;

(5)、筛选频繁项集中频繁项个数为2的频繁模式,得到频繁模式数组2;

(6)、遍历频繁模式数组2,并计算频繁模式数组2的频繁项集的置信度;设频繁模式数组2的频繁项集为M,其中的两个项分别为A和B,那么该频繁项集M对应的置信度C(M)为:

$$C(M)=S(M)\,(S(M\_A)+S(M\_B))/(2S(M\_A)S(M\_B))$$

其中,S(M_A)和S(M_B)分别为频繁模式数组1中频繁项个数为1且为A和B的频繁项集对应的支持度;

判断置信度是否大于预设的阈值,如果大于,则进入步骤(7);否则继续遍历频繁模式数组,直到遍历完成;

(7)、将频繁项集的两个项分别以正序和倒序组成两个新词,并加入到主题词典中;

(8)、重新遍历语料集中的每一个网络新闻标题,并用网络新闻标题检索主题词典,如果某一网络新闻标题中包含有主题词典中的新词,则以该新词为断点分割网络新闻标题,并进入步骤(9);如果某一网络新闻标题中不包含有主题词典中的新词,则进入步骤(10);

(9)、利用CRF模型分别对分割后网络新闻标题进行分词,再将其对应的分词结果和断点对应的新词作为网络新闻标题的最终分词结果;

(10)、利用CRF模型直接对网络新闻标题进行分词,得到最终分词结果;

(11)、重复步骤(8),直到所有的网络新闻标题遍历结束后,将所有的最终分词结果存入到数组2中,同时删除主题词典中通过网络新闻标题未被检索出的新词,再进入步骤(12);

(12)、将数组1和数组2中的每一项分词进行一一比对,如果每一项分词均相同,则网络舆情主题抽取结束,并进入步骤(13);如果有某一项不相同,令数组1等于数组2再返回步骤(4);

(13)、输出主题词典;

(13.1)、设置最小颗粒词集合;将数组1中的所有词并入到最小颗粒词集合中,再标记出最小颗粒词集合中每一个词的词性;

(13.2)、计算主题词典中所有新词的平均实词匹配度:设主题词典中某一新词为Topic,其中有n个网络新闻标题包含该新词Topic,记为T1,T2,...Tn;

计算新词Topic的平均实词匹配度ANMD(Topic):

$$ANMD(Topic)=(n(Topic)/n(T1)+(n(Topic)/n(T2)+\cdots+(n(Topic)/n(Tn))/n;$$

其中,n(Topic)为拼成新词Topic的最小颗粒词集合中使用实词的个数,n(T1),n(T2),…n(Tn)分别为对应网络新闻标题在最小颗粒词集合中使用实词的个数;

(13.3)、将主题词典中所有新词按平均实词匹配度大小进行降序排列,再输出主题词

典。

2.根据权利要求1所述的一种通过新词抽取网络舆情主题的方法，其特征在于，所述步骤(8)中，检索主题词典时，如果主题词典中的两个及以上的新词有重叠的部分，则取最后被检索出的新词作为断点。

一种通过新词抽取网络舆情主题的方法

技术领域

[0001] 本发明属于机器学习和信息挖掘技术领域，更为具体地讲，涉及一种通过新词抽取网络舆情主题的方法。

背景技术

[0002] 主题是指文本的中心思想，泛指主要内容。主题抽取技术是文本处理的基础技术之一，目前国内外主题抽取的普遍方法是应用各种加权算法，计算词对文本主题的贡献大小，并选定贡献大的词作为主题词，即由文本到关键词再到主题词的这样一个过程。但此类的加权算法大都是统计和经验的加权体系，并未考虑文本中词与词之间的关联和联系，尤其是在处理一个文本集的时候，每个文本长短不一，携带的信息量也参差不齐，使加权算法普遍性不强。另有一种主题抽取方法是基于语义对文本进行分析，但由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前还处于试验阶段。

[0003] 现有的主题抽取算法另一个缺陷是依赖对词的选择和处理，上面已经提到主题抽取的过程是由文本到关键词再到主题词，在当前中文文本处理领域，分词也一直是文本处理的前提和基础，分词的漏检和错误会影响抽取的关键词的性能，最终导致主题抽取的可读性不强，甚至错误。在网络信息爆炸的现在，大量新词的出现和流行，分词的漏检和错误主要就表现在新词的识别困难。

[0004] 事实上，通过对网络舆情的持续跟踪和研究中发现，大多数网络舆情的主题，或者热点，本身就是一个新词，这里的新词指的是两个含义，一个是字典中未登录的词，比如“十动然拒”，另一个是两个或多个字典中已登录的词组合而成的新词，比如“闺蜜干政”。因此基于新词发现的主题抽取能够具有相当高的可信度，尤其是在网络舆情和新闻报导中，正确率能达到半数以上。

发明内容

[0005] 本发明的目的在于克服现有技术的不足，提供一种通过新词抽取网络舆情主题的方法，将文本的标题作为处理对象，通过关联规则挖掘词与词之间的关联和联系，实现新词的发现和主题词的抽取。

[0006] 为实现上述发明目的，本发明一种通过新词抽取网络舆情主题的方法，其特征在于，包括以下步骤：

[0007] (1)、构建一个空的主题词典；

[0008] (2)、利用网络爬虫从互联网中爬取网络新闻，对爬取的网络新闻进行去重处理后，将网络新闻的标题存入语料集中；

[0009] (3)、遍历语料集中的每一个网络新闻标题，利用CRF模型对网络新闻标题进行分词，把所有的分词结果存入到数组1中；

[0010] (4)、设置数据挖掘算法FP-growth的支持度阈值，利用数据挖掘算法FP-growth挖

掘出数组1中的频繁项集，得到由频繁项集中的频繁项和对应支持度组成的频繁模式数组1；

[0011] (5)、筛选频繁项集中频繁项个数为2的频繁模式，得到频繁模式数组2；

[0012] (6)、遍历频繁模式数组2，并计算频繁模式数组2的频繁项集的置信度；

[0013] 设频繁模式数组2的频繁项集为M，其中的两个项分别为A和B，那么该频繁项集M对应的置信度C(M)为：

[0014] C(M)＝S(M)(S(M_A)+S(M_B))/(2S(M_A)S(M_B))

[0015] 其中，S(M_A)和S(M_B)分别为频繁模式数组1中项的个数为1且为A和B的频繁项集对应的支持度；

[0016] 判断置信度是否大于预设的阈值，如果大于，则进入步骤(7)；否则继续遍历频繁模式数组，直到遍历完成；

[0017] (7)、将频繁项集的两个项分别以正序和倒序组成两个新词，并加入到主题词典中；

[0018] (8)、重新遍历语料集中的每一个网络新闻标题，并用网络新闻标题检索主题词典，如果某一网络新闻标题中包含有主题词典中的新词，则以该新词为断点分割网络新闻标题，并进入步骤(9)；如果某一网络新闻标题中不包含有主题词典中的新词，则进入步骤(10)；

[0019] (9)、利用CRF模型分别对分割后网络新闻标题进行分词，再将其对应的分词结果和断点对应的新词作为网络新闻标题的最终分词结果；

[0020] (10)、利用CRF模型直接对网络新闻标题进行分词，得到最终分词结果；

[0021] (11)、重复步骤(8)，直到所有的网络新闻标题遍历结束后，将所有的最终分词结果存入到数组2中，同时删除主题词典中通过网络新闻标题未被检索出的新词，再进入步骤(12)；

[0022] (12)、将数组1和数组2中的每一项分词进行一一比对，如果每一项分词均相同，则网络舆情主题抽取结束，并进入步骤(13)；如果有某一项不相同，令数组1等于数组2再返回步骤(4)；

[0023] (13)、输出主题词典；

[0024] (13.1)、设置最小颗粒词集合；将数组1中的所有词并入到最小颗粒词集合中，再标记出最小颗粒词集合中每一个词的词性；

[0025] (13.2)、计算主题词典中所有新词的平均实词匹配度：设主题词典中某一新词为Topic，其中有n个网络新闻标题包含该新词Topic，记为T1,T2,…Tn；

[0026] 计算新词Topic的平均实词匹配度ANMD(Topic)：

[0027] ANMD(Topic)＝(n(Topic)/n(T1)+(n(Topic)/n(T2)+...+(n(Topic)/n(Tn))/n；

[0028] 其中，n(Topic)为拼成新词Topic的最小颗粒词集合中使用实词的个数，n(T1)，n(T2)，...n(Tn)分别为对应网络新闻标题在最小颗粒词集合中使用实词的个数；

[0029] (13.3)、将主题词典中所有新词按平均实词匹配度大小进行降序排列，再输出主题词典。

[0030] 本发明的发明目的是这样实现的：

[0031] 本发明一种通过新词抽取网络舆情主题的方法，通过两个或两个以上的关键词拼

接得到的新词作为主题词;其中,对于关键词的选取,需要考虑代表文章的中心和主旨的文本标题,文本标题分得的词作为文本的关键词,同时还提高算法效率和减少因为文本差异对主题抽取的影响;同时,本发明以平均实词匹配度来量化这些新词对于网络舆情的主题的贡献程度,平均实词匹配度越高表示新词对于网络舆情的主题的贡献程度越大,因而,具有相当高的可信度,能够适应当今网络舆情日益增长的趋势。

[0032] 同时,本发明一种通过新词抽取网络舆情主题的方法还具有以下有益效果:

[0033] (1)、基于CRF模型的中文分词方法,在现在的中文分词领域,CRF代表了新一代的机器学习技术,其基本思路是对汉字进行标注即由字构词(组词),不仅考虑了文字词语出现的频率信息,同时考虑上下文语境,具备较好的学习能力,从而避免了词典存在的不足,并且增加了对歧义词和未登录词的识别,提高了分词的可读性和主题抽取的质量;

[0034] (2)、基于FP-Tree关联规则的合成主题词方法,在数据挖掘领域,数据项之间的关联规则称为关联模式,FP-growth算法使其中的主要算法之一。利用FP-growth算法可以挖掘出关键词之间的关联和联系,从而提高了主题抽取的准确率。

附图说明

[0035] 图1是本发明一种通过新词抽取网络舆情主题的方法流程图;

[0036] 图2是CRF分词模型的训练流程图。

具体实施方式

[0037] 下面结合附图对本发明的具体实施方式进行描述,以便本领域的技术人员更好地理解本发明。需要特别提醒注意的是,在以下的描述中,当已知功能和设计的详细描述也许会淡化本发明的主要内容时,这些描述在这里将被忽略。

[0038] 实施例

[0039] 图1是本发明一种通过新词抽取网络舆情主题的方法流程图。

[0040] 在本实施例中,如图1所示,本发明一种基于新词的网络舆情主题抽取方法,包括以下步骤:

[0041] S1、构建一个空的主题词典;

[0042] S2、利用网络爬虫从互联网中爬取网络新闻,对爬取的网络新闻进行去重处理后,将网络新闻的标题存入语料集中;例如:通过爬虫爬取新浪、百度、腾讯……,爬取当日的网络新闻,再对相同的网络新闻进行去重;

[0043] S3、遍历语料集中的每一个网络新闻标题,利用CRF分词模型对网络新闻标题进行分词,把所有的分词结果存入到数组1中;

[0044] 在本实施例中,CRF分词模型的训练步骤为:

[0045] 1)、提取语料集中的每一个网络新闻标题;

[0046] 2)、对每一个网络新闻标题进行半自动的分块和标注,即模型给出候选结果,人工进行判别、修改和再标注,得到标注集;

[0047] 3)、随机选择一部分标注集在条件随机场中进行训练,其余的标注集在所述条件随机场中进行测试,最终得到训练好的CRF分词模型;

[0048] S4、、设置数据挖掘算法FP-growth的支持度阈值,利用数据挖掘算法FP-growth挖

掘出数组1中的频繁项集，得到由频繁项集中的频繁项和对应支持度组成的频繁模式数组1；

[0049] S5、筛选频繁项集中频繁项个数为2的频繁模式，得到频繁模式数组2；

[0050] S6、遍历频繁模式数组2，并计算频繁模式数组2的频繁项集的置信度；

[0051] 设频繁模式数组2的频繁项集为M，其中的两个项分别为A和B，那么该频繁项集M对应的置信度C(M)为：

[0052] C(M)＝S(M)(S(M_A)+S(M_B))/(2S(M_A)S(M_B))

[0053] 其中，S(M_A)和S(M_B)分别为频繁模式数组1中项的个数为1且为A和B的频繁项集对应的支持度；

[0054] 判断置信度是否大于预设的阈值，如果大于，则进入步骤S7；否则继续遍历频繁模式数组，直到遍历完成；

[0055] S7、将频繁项集的两个项分别以正序和倒序组成两个新词，并加入到主题词典中；

[0056] S8、重新遍历语料集中的每一个网络新闻标题，并用网络新闻标题检索主题词典，如果某一网络新闻标题中包含有主题词典中的新词，则以该新词为断点分割网络新闻标题，并进入步骤S9；如果某一网络新闻标题中不包含有主题词典中的新词，则进入步骤S10；

[0057] 其中，检索主题词典时，如果主题词典中的两个及以上的新词有重叠的部分，则取最后被检索出的新词作为断点；

[0058] 在本实施例中，如果两个及以上的新词在文本中有重叠的部分，则选取检索到的最后一个新词作为断点，忽略其他的新词，如文本为：“华为超三星成最赚钱安卓手机”中，主题词典同时包含了新词“超三星”和“华为超三星”，由于“华为超三星”是后加入词典的新词，最后被检索出来，因此选取“华为超三星”作为文本的断点；

[0059] S9、利用CRF分词模型分别对分割后网络新闻标题进行分词，再将其对应的分词结果和断点对应的新词作为网络新闻标题的最终分词结果；

[0060] S10、利用CRF模型直接对网络新闻标题进行分词，得到最终分词结果；

[0061] S11、重复步骤S8，直到所有的网络新闻标题遍历结束后，将所有的最终分词结果存入到数组2中，同时删除主题词典中通过网络新闻标题未被检索出的新词，再进入步骤S12；

[0062] S12、将数组1和数组2中的每一项分词进行一一比对，如果每一项分词均相同，则网络舆情主题抽取结束，并进入步骤S13；如果有某一项不相同，令数组1等于数组2再返回步骤S4；

[0063] S13、输出主题词典；

[0064] S13.1、设置最小颗粒词集合；将数组1中的所有词并入到最小颗粒词集合中，再标记出最小颗粒词集合中每一个词的词性；

[0065] S13.2、计算主题词典中所有新词的平均实词匹配度：设主题词典中某一新词为Topic，其中有n个网络新闻标题包含该新词Topic，记为T1,T2,...Tn；

[0066] 计算新词Topic的平均实词匹配度ANMD(Topic)：

[0067] ANMD(Topic)＝(n(Topic)/n(T1)+(n(Topic)/n(T2)+...+(n(Topic)/n(Tn))/n；

[0068] 其中，n(Topic)为拼成新词Topic的最小颗粒词集合中使用实词的个数，n(T1)，n(T2)，...n(Tn)分别为对应网络新闻标题在最小颗粒词集合中使用实词的个数；

[0069] S13.3、将主题词典中所有新词按平均实词匹配度大小进行降序排列，再输出主题词典。

[0070] 尽管上面对本发明说明性的具体实施方式进行了描述，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。

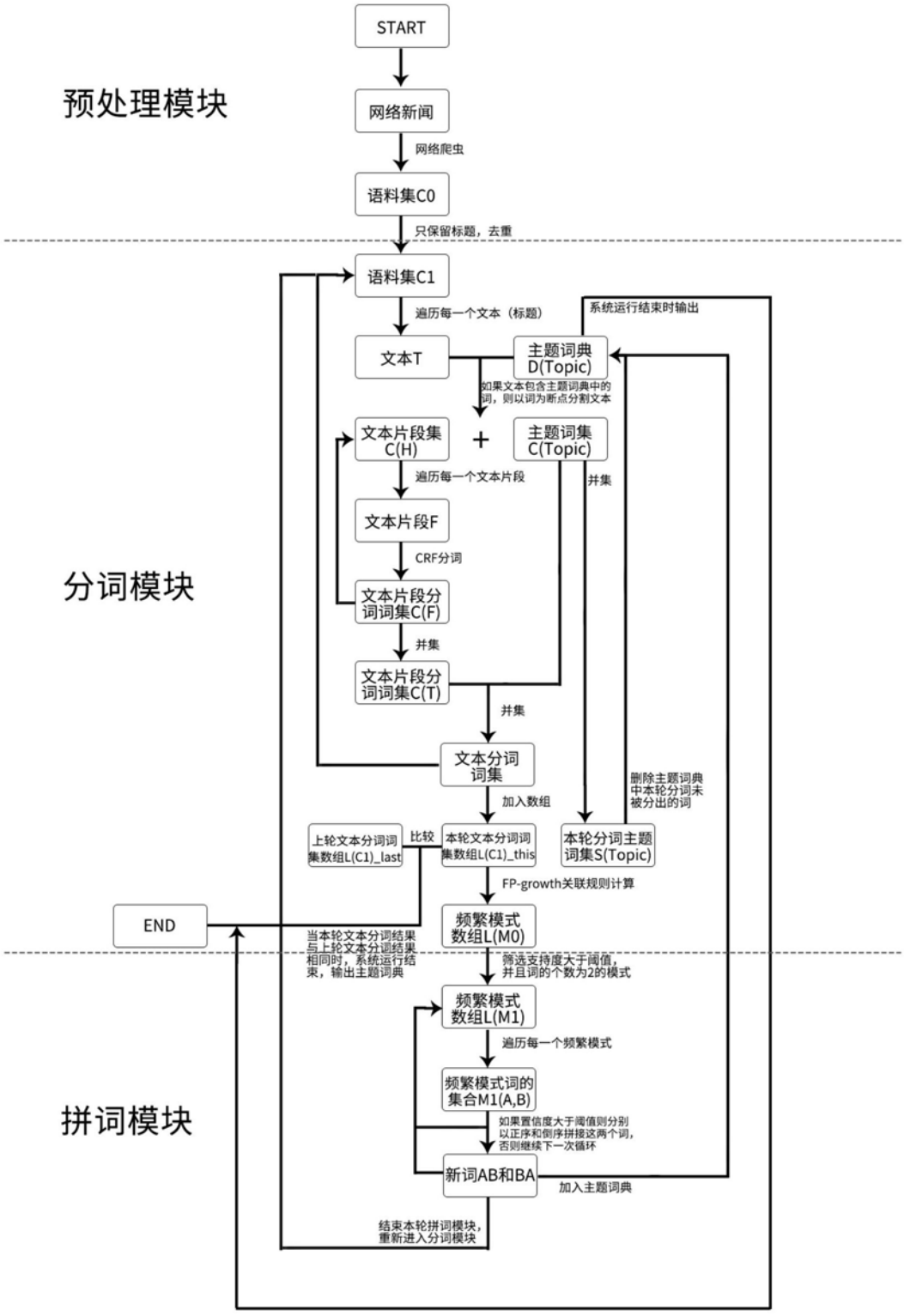

预处理模块
分词模块
拼词模块
START
网络新闻
网络爬虫
语料集C0
只保留标题，去重
语料集C1
遍历每一个文本（标题）
文本T
主题词典D(Topic)
系统运行结束时输出
如果文本包含主题词典中的词，则以词为断点分割文本
文本片段集C(H)
+
主题词集C(Topic)
并集
遍历每一个文本片段
文本片段F
CRF分词
文本片段分词词集C(F)
并集
文本片段分词词集C(T)
并集
文本分词词集
加入数组
删除主题词典中本轮分词未被分出的词
上轮文本分词词集数组L(C1)_last
比较
本轮文本分词词集数组L(C1)_this
本轮分词主题词集S(Topic)
FP-growth关联规则计算
END
当本轮文本分词结果与上轮文本分词结果相同时，系统运行结束，输出主题词典
频繁模式数组L(M0)
筛选支持度大于阈值，并且词的个数为2的模式
频繁模式数组L(M1)
遍历每一个频繁模式
频繁模式词的集合M1(A,B)
如果置信度大于阈值则分别以正序和倒序拼接这两个词，否则继续下一次循环
新词AB和BA
加入主题词典
结束本轮拼词模块，重新进入分词模块

图1

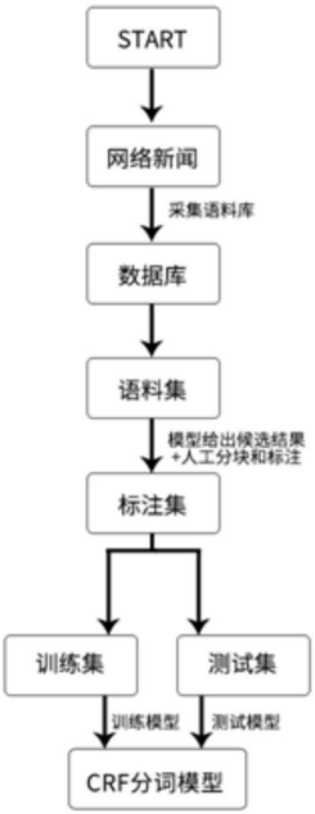

START
网络新闻
采集语料库
数据库
语料集
模型给出候选结果
+人工分块和标注
标注集
训练集
测试集
训练模型
测试模型
CRF分词模型

图2