(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 106599181 A
(43)申请公布日 2017.04.26

(21)申请号 201611145855.9

(22)申请日 2016.12.13

(71)申请人 浙江网新恒天软件有限公司
地址 310012 浙江省杭州市西湖区教工路23号百脑汇科技大厦18楼

(72)发明人 庄郭冕 黄乔 彭志宇 付晗 王忆诗

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 刘静 邱启旺

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

权利要求书2页 说明书5页 附图3页

(54)发明名称

一种基于主题模型的新闻热点检测方法

(57)摘要

本发明公开了一种基于主题模型的新闻热点检测方法，通过网络爬虫定向爬取新闻流，首先对文章进行分词，去除停用词及无意义字符串等预处理，继而对预处理后的文章进行特征提取，构建文本模型，然后通过文本聚类算法将相似度程度高的文本加入到最相似的类别中，得到主题库，接着对新老主题进行相似度计算，对于相似度高的新老主题进行合并，最后进行主题热度计算，通过排序选出最热的主题。本发明创新性地将LDA算法应用在热点主题发现中，并提出了爆发性的概念，能够及时有效地发现最新热点新闻，同时提出了主题热度衰减概念，能够实时记录跟踪主题热度，真实地反映了新闻热点的发展变化，对于热点新闻的追踪展示具有重要意义。

CN 106599181 A

1.一种基于主题模型的新闻热点检测方法，其特征在于，包括以下步骤：

(1)采用网络爬虫的方式定向爬取新闻流，每到来N篇新文章进行一次批处理，对爬取数据进行数据清洗、文章分词得到预处理后的文章；

(2)构建向量空间模型：经过预处理操作，原始文档可以看做是由一堆词语构成的，如果把文档看做是一个向量的话，那么每个词语就是一维特征，通过将文档转化为向量，文本数据就变为可以被计算机处理的结构化数据，两个文档之间的相似性问题就转化为了两个向量之间的相似性问题。在计算文档向量每一维的权重时，采用改进的B-TFIDF算法，算法公式如下：

$$b_i(w)=\frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (1)$$

$$weight(d_i,w)=\frac{(tf_i(d_i,w)\log(^{(N+1)}/_{(df_i(w)+0.5)})\cdot b_i(w))}{\sqrt{\sum_{w'\in d}(tf(d,w')\log(^{(N+1)}/_{(df(w')+0.5)})\cdot b(w'))^2}} \quad (2)$$

公式(1)中w代表单词，A表示新文章中包含单词w的文章数，B表示新文章中不包含单词w的文章数，C表示历史文章中包含单词w的文章数，D表示历史文章中不包含单词w的文章数，公式(2)中$d_i$表示第i篇新文章，N表示新文章总数，tf(d,w)表示单词w在文章d中的词频，df(w)表示包含单词w的文章数。该算法将词语的爆发性考虑在内，爆发性即一个词语在短期内突然大量出现。通过以上算法计算构成文档的每个词语的权重，进而生成文章的向量空间模型$D_i$＝(weight($d_i$,$w_1$),weight($d_i$,$w_2$),weight($d_i$,$w_3$)……weight($d_i$,$w_n$))，其中n为总词数。

(3)文章聚类：经过步骤2，文本被表示为向量的形式，对文本向量进行聚类；采用LDA主题模型聚类算法，具体为：

LDA聚类过程：LDA是一个三层贝叶斯概率模型，包含词、主题和文档三层，将一篇文章的产生看做是这样一个过程：以一定概率选定某个主题，并在这个主题中以一定概率选定某个词，文档到主题服从多项式分布，主题到词服从多项式分布，通过LDA聚类得到“主题-词语”概率矩阵phi以及“文档-主题”概率矩阵theta，根据“文档-主题”概率矩阵theta得到m个主题和m个主题对应N篇文章的概率，theta的每一行i代表一篇文章，每一列j代表一个主题，对应矩阵值$theta_{ij}$是文章i属于主题j的概率。设置筛选阈值为thresholdT，若$theta_{ij}$>thresholdT则认为文章i属于主题j，由此选出每个主题对应的文章。

LDA聚类个数m的确定：分别设置聚类个数为N/10-N/5重复执行LDA聚类算法，然后计算每一次执行结果的主题间相似度，选择主题间相似度最低的执行结果对应的主题个数。主题间相似度的计算根据LDA聚类得到的“主题-词语”概率矩阵phi，phi的每一行j代表一个主题$T_j$，每一列k代表一个单词$w_k$，$phi_{jk}$代表主题$T_j$包含单词$w_k$的概率。Phi的一行可以看做主题$T_j$的向量形式$T_j$＝($w_1$,$w_2$,$w_3$,…$w_k$…$w_n$)，n为总词数。计算主题两两之间的相似度，求相似度平均值，取最小值作为最终的主题间相似度。相似度的计算采用余弦相似度的计算方法，计算公式如下：

$$sim(T_i,T_j)=\frac{\sum_{k=1}^{n}\omega_k(T_i)\times\omega_k(T_j)}{\sqrt{((\sum_{k=1}^{n}\omega_k^2(T_i))(\sum_{k=1}^{n}\omega_k^2(T_j)))}} \quad (3)$$

公式(3)中的$T_i$和$T_j$代表两个主题，$\omega_k(T_i)$代表主题$T_i$在维度k上的值，n表示总词数。

(4)主题关键词提取：从主题下所有文章的题目中提取关键词，先对文章题目进行分词，过滤掉停止词，无意义词和标点符号，剩下的词作为主题关键词。

(5)话题合并：由步骤3得到m个主题和其对应的文章，接下来将m个新主题与旧主题进行合并，计算主题间相似度f1，若f1>0.5则认为两个主题相似，并合并两个主题。主题间相似度f1计算公式如下：

f1＝2*vectorSim*keywordSim/(vectorSim+keywordSim) (4)

公式(4)中的vectorSim代表以主题包含的所有单词作为维度计算主题余弦相似度，keywordSim代表以主题关键词为维度计算主题余弦相似度，余弦相似度的计算公式同公式(3)。

(6)热度计算：经过步骤5得到最终的所有主题，接下来计算主题热度h，筛选出热度高的主题，去掉热度低，即过时的主题。根据热点主题新闻聚集度s高的特点，热度计算公式如下：

$h_t=\sum sim(d_i,t)$ (5)

公式(5)中的$d_i$表示主题T包含的文章，主题T的热度$h_t$等于主题下文章与主题相似度的和，sim函数同公式(3)。

随着时间流逝，一个主题的热度会不断衰减，直至低于阈值该主题被舍弃。热度的衰减，在每次批处理过程中，如果主题T下面有新文章到来，那么主题T的热度$h_t$会相应的增加，$h_t=h_t$*Up，如果没有新的文章添加进主题T，那么热度$h_t$会衰减，$h_t=h_t$*Down，其中Up＞1，Down＜1。

# 一种基于主题模型的新闻热点检测方法

## 技术领域

[0001] 本发明提供了一种基于主题模型的新闻热点检测方法，涉及网络爬虫，聚类分析，文本相似度计算等核心技术与算法，及时有效的检测新闻热点，追踪热点新闻演变。

## 背景技术

[0002] 随着互联网技术的发展，海量信息时代已经来临，各类信息充斥于互联网，但只有少数新闻能引起轰动，即所谓的头条新闻，热点新闻，及时的新闻热点发现能够帮助人们实时关注社会状态。

[0003] 另一方面，一个新闻热点的爆发不是一瞬即逝的，往往伴随着一个跌宕起伏的发展过程，并引发其他潜在问题发生，所以追踪新闻热点的发展过程对于研究社会问题具有重要意义。

[0004] 互联网的发展，大数据的兴起，互联网充斥着大量信息，在这些低质量的信息中发现热点新闻变得极其重要。

## 发明内容

[0005] 本发明的目的在于针对如今互联网信息的繁杂，提供一种基于网络爬虫、聚类分析和主题模型的新闻热点检测方法。

[0006] 本发明的目的是通过以下技术方案来实现的：一种基于主题模型的新闻热点检测方法，通过网络爬虫定向爬取新闻流，首先对文章进行分词，去除停用词及无意义字符串等预处理，继而对预处理后的文章进行特征提取，构建文本模型，然后通过文本聚类算法将将相似度程度高的文本加入到最相似的类别中，得到话题库，接着对新老话题进行相似度计算，对于相似度高的新老话题进行合并，最后进行话题热度计算，通过排序选出最热的话题。具体包括以下步骤：

[0007] (1)采用网络爬虫的方式定向爬取新闻流，每到来N篇新文章进行一次批处理，对爬取数据进行数据清洗、文章分词得到预处理后的文章；

[0008] (2)构建向量空间模型：经过预处理操作，原始文档可以看做是由一堆词语构成的，如果把文档看做是一个向量的话，那么每个词语就是一维特征，通过将文档转化为向量，文本数据就变为可以被计算机处理的结构化数据，两个文档之间的相似性问题就转化为了两个向量之间的相似性问题。在计算文档向量每一维的权重时，采用改进的B-TFIDF算法，算法公式如下：

[0009] $$b_i(w)=\frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (1)$$

[0010] $$weight(d_i,w)=\frac{(tf_i(d_i,w)\log(\frac{(N+1)}{(df_i(w)+0.5)})\cdot b_i(w))}{\sqrt{\sum_{w'\in d}(tf(d,w')\log(\frac{(N+1)}{(df(w')+0.5)})\cdot b(w'))^2}} \quad (2)$$

[0011] 公式(1)中w代表单词，A表示新文章中包含单词w的文章数，B表示新文章中不包含单词w的文章数，C表示历史文章中包含单词w的文章数，D表示历史文章中不包含单词w的文章数，公式(2)中$d_i$表示第i篇新文章，N表示新文章总数，tf(d,w)表示单词w在文章d中的词频，df(w)表示包含单词w的文章数。该算法将词语的爆发性考虑在内，爆发性即一个词语在短期内突然大量出现。通过以上算法计算构成文档的每个词语的权重，进而生成文章的向量空间模型$D_i$＝(weight($d_i$,$w_1$),weight($d_i$,$w_2$),weight($d_i$,$w_3$)…..weight($d_i$,$w_n$))，其中n为总词数。

[0012] (3)文章聚类：经过步骤2，文本被表示为向量的形式，对文本向量进行聚类；采用LDA主题模型聚类算法，具体为：

[0013] LDA聚类过程：LDA是一个三层贝叶斯概率模型，包含词、主题和文档三层，将一篇文章的产生看做是这样一个过程：以一定概率选定某个主题，并在这个主题中以一定概率选定某个词，文档到主题服从多项式分布，主题到词服从多项式分布，通过LDA聚类得到“主题-词语”概率矩阵phi以及“文档-主题”概率矩阵theta，根据“文档-主题”概率矩阵theta得到m个主题和m个主题对应N篇文章的概率，theta的每一行i代表一篇文章，每一列j代表一个主题，对应矩阵值$theta_{ij}$是文章i属于主题j的概率。设置筛选阈值为thresholdT，若$theta_{ij}$>thresholdT则认为文章i属于主题j，由此选出每个主题对应的文章。

[0014] LDA聚类个数m的确定：分别设置聚类个数为N/10-N/5重复执行LDA聚类算法，然后计算每一次执行结果的主题间相似度，选择主题间相似度最低的执行结果对应的主题个数。主题间相似度的计算根据LDA聚类得到的“主题-词语”概率矩阵phi，phi的每一行j代表一个主题$T_j$，每一列k代表一个单词$w_k$，$phi_{jk}$代表主题$T_j$包含单词$w_k$的概率。Phi的一行可以看做主题$T_j$的向量形式$T_j$＝($w_1$,$w_2$,$w_3$,…$w_k$…$w_n$)，n为总词数。计算主题两两之间的相似度，求相似度平均值，取最小值作为最终的主题间相似度。相似度的计算采用余弦相似度的计算方法，计算公式如下：

[0015] $$sim(T_i,T_j)=\frac{\sum_{k=1}^{n}\omega_k(T_i)\times\omega_k(T_j)}{\sqrt{((\sum_{k=1}^{n}\omega_k^2(T_i))(\sum_{k=1}^{n}\omega_k^2(T_j)))}} \quad (3)$$

[0016] 公式(3)中的$T_i$和$T_j$代表两个主题，$\omega_k(T_i)$代表主题$T_i$在维度k上的值，n表示总词数。

[0017] (4)主题关键词提取：从主题下所有文章的题目中提取关键词，先对文章题目进行分词，过滤掉停止词，无意义词和标点符号，剩下的词作为主题关键词。

[0018] (5)话题合并：由步骤3得到m个主题和其对应的文章，接下来将m个新主题与旧主题进行合并，计算主题间相似度f1，若f1>0.5则认为两个主题相似，并合并两个主题。主题间相似度f1计算公式如下：

[0019] f1＝2*vectorSim*keywordSim/(vectorSim+keywordSim) (4)

[0020] 公式(4)中的vectorSim代表以主题包含的所有单词作为维度计算主题余弦相似度，keywordSim代表以主题关键词为维度计算主题余弦相似度，余弦相似度的计算公式同公式(3)。

[0021] (6)热度计算：经过步骤5得到最终的所有主题，接下来计算主题热度h，筛选出热

度高的主题，去掉热度低，即过时的主题。根据热点主题新闻聚集度s高的特点，热度计算公式如下：

[0022] $h_t = \sum sim(d_i, t)$ (5)

[0023] 公式(5)中的$d_i$表示主题T包含的文章，主题T的热度$h_t$等于主题下文章与主题相似度的和，sim函数同公式(3)。

[0024] 随着时间流逝，一个主题的热度会不断衰减，直至低于阈值该主题被舍弃。热度的衰减，在每次批处理过程中，如果主题T下面有新文章到来，那么主题T的热度$h_t$会相应的增加，$h_t = h_t * Up$，如果没有新的文章添加进主题T，那么热度$h_t$会衰减，$h_t = h_t * Down$，其中$Up > 1$，$Down < 1$。

[0025] 本发明的有益效果是：本发明创新性地将LDA算法应用在热点主题发现中，并提出了爆发性的概念，能够及时有效地发现最新热点新闻，同时提出了提出了主题热度衰减概念，能够实时记录跟踪主题热度，真实地反映了新闻热点的发展变化，对于热点新闻的追踪展示具有很重要的意义。

## 附图说明

[0026] 图1是基于主题模型的新闻热点检测流程示意图；

[0027] 图2是文章建模过程示意图；

[0028] 图3是LDA聚类过程示意图；

[0029] 图4是新旧主题合并示意图；

[0030] 图5是主题热度计算示意图。

## 具体实施方式

[0031] 下面结合附图和具体实施例对本发明作进一步详细说明。

[0032] 如图1所示，本发明提出的一种基于主题模型的新闻热点检测方法，包括以下步骤：

[0033] (1)采用网络爬虫的方式定向爬取新闻流，每到来N篇文章进行一次批处理，对爬取数据进行数据清洗、文章分词得到预处理后的文章；

[0034] (2)构建向量空间模型：如图2所示，经过预处理操作，原始文档可以看做是由一堆词语构成的，如果把文档看做是一个向量的话，那么每个词语就是一维特征，通过将文档转化为向量，文本数据就变为可以被计算机处理的结构化数据，两个文档之间的相似性问题就转化为了两个向量之间的相似性问题。在计算文档向量每一维的权重时，采用了改进的B-TFIDF算法，算法公式如下：

[0035] $$b_i(w) = \frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (1)$$

[0036] $$weight(d_i, w) = \frac{(tf_i(d_i, w)\log(\frac{(N+1)}{(df_i(w)+0.5)}) \cdot b_i(w))}{\sqrt{\sum_{w' \in d}(tf(d, w')\log(\frac{(N+1)}{(df(w')+0.5)}) \cdot b(w'))^2}} \quad (2)$$

[0037] 公式1中w代表单词，A表示新文章中包含单词w的文章数，B表示新文章中不包含单

词w的文章数，C表示历史文章中包含单词w的文章数，D表示历史文章中不包含单词w的文章数，公式2中$d_i$表示第i篇新文章，N表示新文章总数，tf(d,w)表示单词w在文章d中的词频，df(w)表示包含单词w的文章数。该算法将词语的爆发性考虑在内，爆发性即一个词语在短期内突然大量出现。通过以上算法计算构成文档的每个词语的权重，进而生成文章的向量空间模型$D_i$＝(weight($d_i$,$w_1$),weight($d_i$,$w_2$),weight($d_i$,$w_3$)…..weight($d_i$,$w_n$))，其中n为总词数。

[0038] (3)文章聚类：经过步骤2，文本被表示为向量的形式，对文本向量进行聚类；如图3所示，这里采用了LDA主题模型聚类算法，LDA是一个三层贝叶斯概率模型，包含词、主题和文档三层，将一篇文章的产生看做是这样一个过程：以一定概率选定某个主题，并在这个主题中以一定概率选定某个词，文档到主题服从多项式分布，主题到词服从多项式分布，通过LDA聚类分析得到“主题-词语”概率矩阵以及“文档-主题”概率矩阵，详细过程见下面描述。

[0039] LDA聚类过程：LDA是一个三层贝叶斯概率模型，通过LDA聚类可以得到“主题-词语”概率矩阵phi以及“文档-主题”概率矩阵theta，根据“文档-主题”概率矩阵theta得到m个主题和m个主题对应N篇文章的概率，theta的每一行i代表一篇文章，每一列j代表一个主题，对应矩阵值$theta_{ij}$是文章i属于主题j的概率。设置筛选阈值为thresholdT(优选值0.32)，若$theta_{ij}$>thresholdT则认为文章i属于主题j，由此选出每个主题对应的文章。

[0040] LDA聚类个数m的确定：由于N篇文章聚类个数在N/10到N/5之间较符合现实情况(例如，当新文章总数N为150时，聚类个数在15到30之间较符合现实情况)，所以分别设置聚类个数为N/10-N/5重复执行LDA聚类算法，然后计算每一次执行结果的主题间相似度，选择主题间相似度最低的执行结果对应的主题个数。主题间相似度的计算需要根据LDA聚类得到的“主题-词语”概率矩阵phi，phi的每一行j代表一个主题$T_j$，每一列k代表一个单词$w_k$，$phi_{jk}$代表主题$T_j$包含单词$w_k$的概率。Phi的一行可以看做主题$T_j$的向量形式$T_j$＝($w_1$,$w_2$,$w_3$,…$w_k$…$w_n$)，n为总词数。计算主题两两之间的相似度，求相似度平均值，取最小值作为最终的主题间相似度。相似度的计算采用了余弦相似度的计算方法，计算公式如下：

[0041] $$sim(T_i,T_j)=\frac{\sum_{k=1}^{n}\omega_k(T_i)\times\omega_k(T_j)}{\sqrt{((\sum_{k=1}^{n}\omega_k^2(T_i))(\sum_{k=1}^{n}\omega_k^2(T_j)))}} \quad (3)$$

[0042] 公式3中的$T_i$和$T_j$代表两个主题，$\omega_k(T_i)$代表主题$T_i$在维度k上的值，n表示总词数。

[0043] (4)主题关键词提取：从主题下所有文章的题目中提取关键词，先对文章题目进行分词，过滤掉停止词，无意义词和标点符号，剩下的词作为主题关键词。

[0044] (5)话题合并：由步骤3得到m个主题和其对应的文章，接下来将m个新主题与旧主题进行合并，如图4所示，计算主题间相似度f1，若f1>0.5则认为两个主题相似，并合并两个主题。主题间相似度f1计算公式如下：

[0045] f1＝2*vectorSim*keywordSim/(vectorSim+keywordSim) (4)

[0046] 公式4中的vectorSim代表以主题包含的所有单词作为维度计算主题余弦相似度，keywordSim代表以主题关键词为维度计算主题余弦相似度，余弦相似度的计算公式同公式3。

[0047] (6)热度计算：如图5所示，经过步骤5得到最终的所有主题，接下来计算主题热度

h，筛选出热度高的主题，去掉热度低，即过时的主题。根据热点主题新闻聚集度s高的特点，热度计算公式如下：

[0048] $h_t = \sum sim(d_i, t)$ (5)

[0049] 公式5中的di表示主题T包含的文章，主题T的热度$h_t$等于主题下文章与主题相似度的和，sim函数同公式3。

[0050] 随着时间流逝，一个主题的热度会不断衰减，直至低于阈值该主题被舍弃。热度的衰减，在每次批处理过程中，如果主题T下面有新文章到来，那么主题T的热度$h_t$会相应的增加，$h_t = h_t * Up$(优选值1.05)，如果没有新的文章添加进主题T，那么热度$h_t$会衰减，$h_t = h_t * Down$(优选值0.9)。

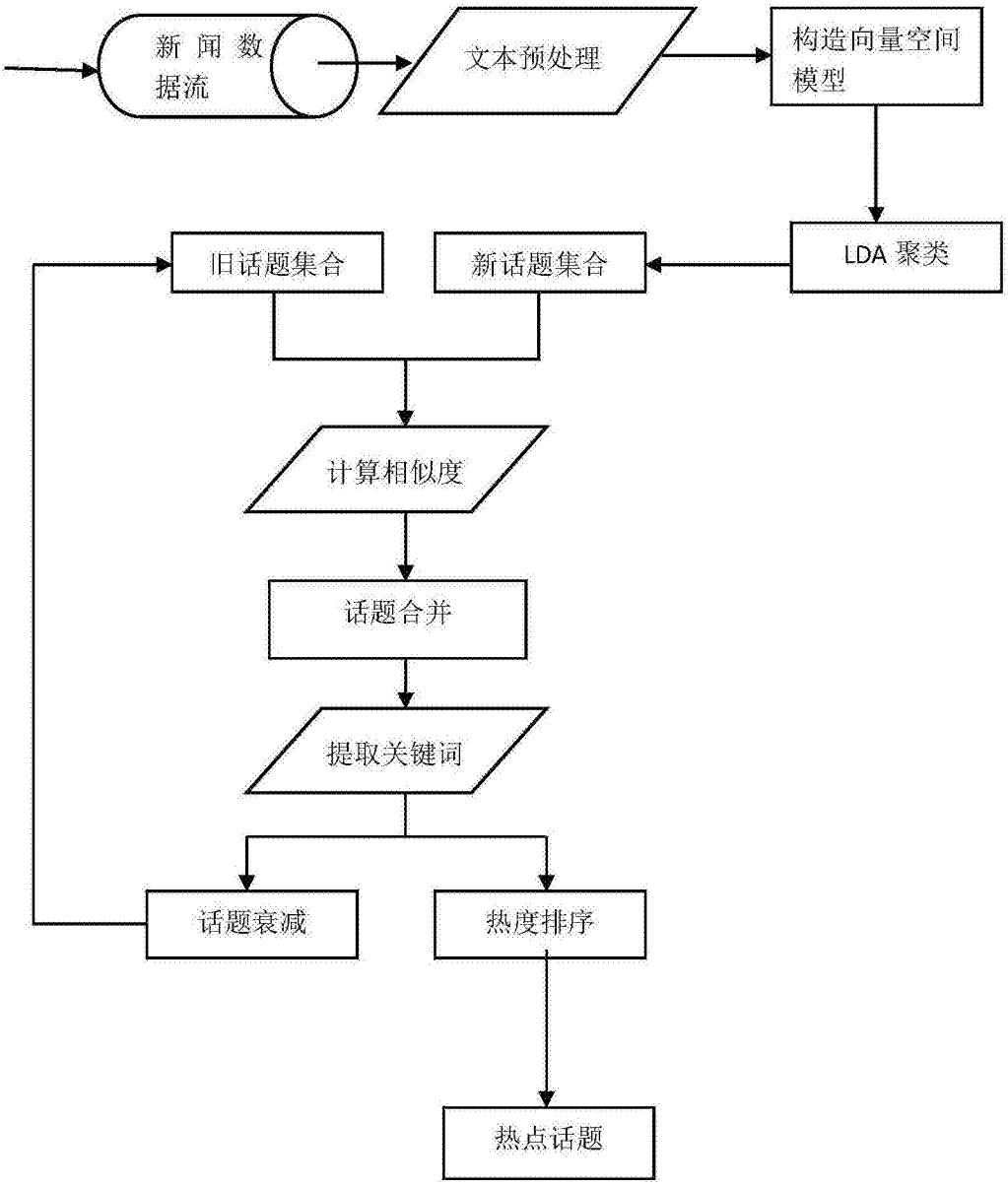
新 闻 数 据流
文本预处理
构造向量空间模型
LDA 聚类
旧话题集合
新话题集合
计算相似度
话题合并
提取关键词
话题衰减
热度排序
热点话题

图1

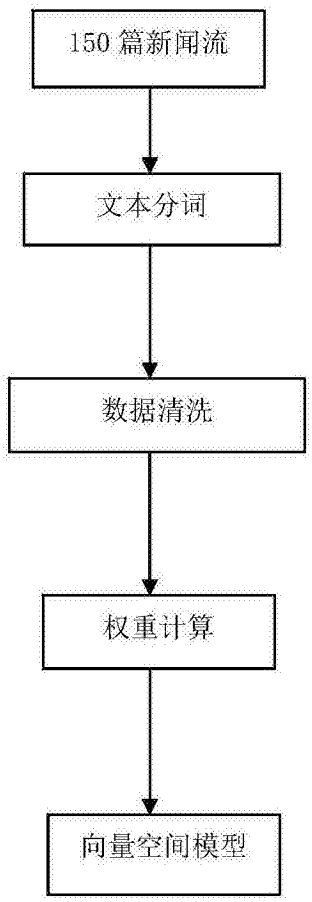
150 篇新闻流
文本分词
数据清洗
权重计算
向量空间模型

图2

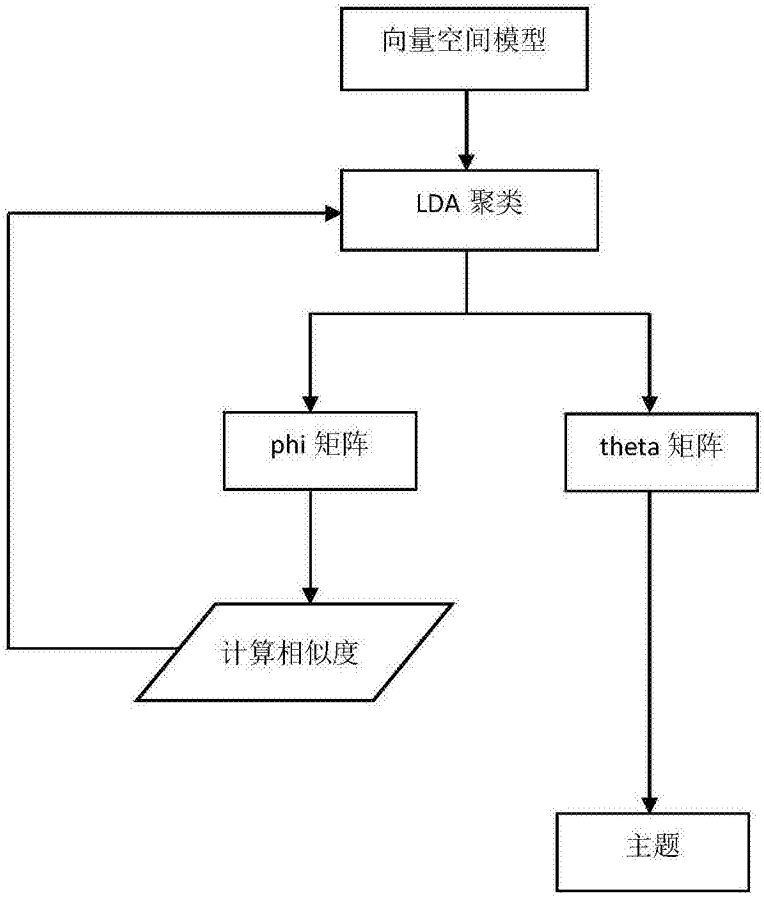
向量空间模型
LDA 聚类
phi 矩阵
theta 矩阵
计算相似度
主题

图3

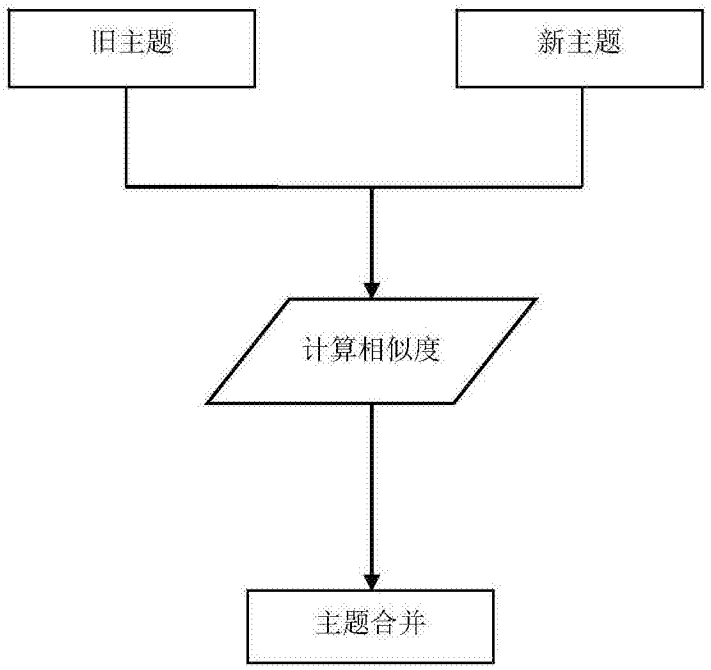
旧主题
新主题
计算相似度
主题合并

图4

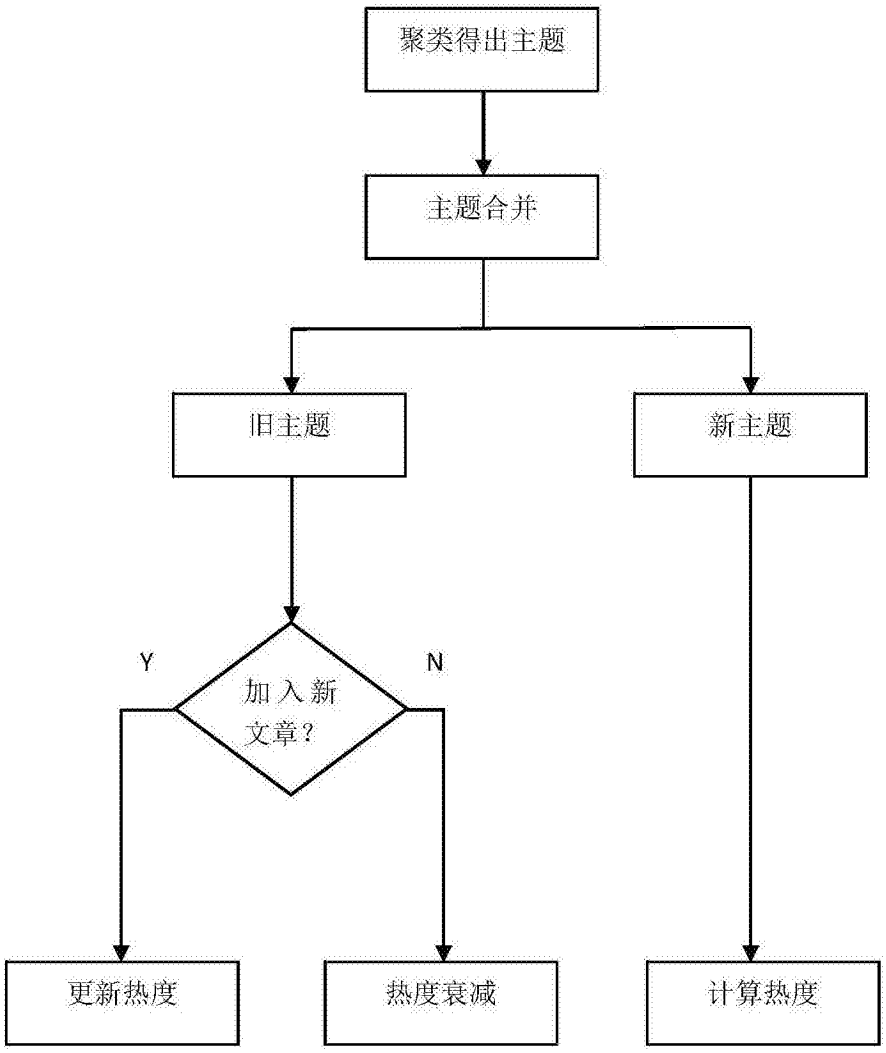


图5