(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 103544255 A
(43)申请公布日 2014.01.29

(21)申请号 201310482522.5

(22)申请日 2013.10.15

(71)申请人 常州大学
地址 213164 江苏省常州市武进区滆湖路 1 号

(72)发明人 陶宇炜 谢爱娟 熊长江 王娟琳

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

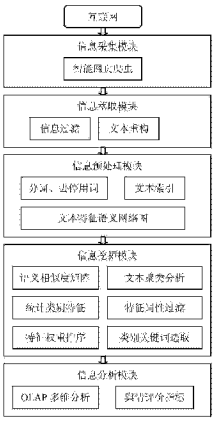
权利要求书3页 说明书9页 附图1页

(54)发明名称

基于文本语义相关的网络舆情信息分析方法

(57)摘要

本发明涉及一种基于文本语义相关的网络舆情信息分析系统，包括以下模块：网络舆情信息采集模块，从网页中采集蕴含丰富的各种舆情信息；舆情信息萃取模块和舆情信息预处理模块将采集的舆情信息进行初步过滤和切分，提取正文部分的元信息，建立文本的特征语义网络图，并进行加权计算和特征抽取，为舆情信息挖掘提供服务。舆情信息挖掘模块，采用基于语义相似度的改进文本聚类分析方法，将文本进行归类；舆情信息分析模块，把舆情信息经过挖掘的数据进行OLAP多维统计，分析舆情评测指标，为相关舆情信息决策提供支持。本发明解决文本中词语语义信息不完整的问题，高效实现大规模网络环境下对动态数据的聚类分析和热点话题发现。



CN 103544255 A

1. 基于文本语义相关的网络舆情信息分析方法，其特征在于：采用包括网络舆情信息采集模块、舆情信息萃取模块、舆情信息预处理模块、舆情信息挖掘模块、舆情信息分析模块和包含舆情信息数据库的网络舆情信息分析系统，并包括如下步骤：

a. 网络舆情信息采集模块从网页中采集各种舆情信息，并存储到舆情信息数据库中；

b. 舆情信息萃取模块和舆情信息预处理模块将步骤 a 采集的舆情信息进行初步过滤和切分，抽取文本所包含的内容信息，为舆情信息挖掘提供数据服务；

c. 在步骤 b 基础上，舆情信息挖掘模块采用基于语义相似度的改进文本聚类分析方法，生成类别描述信息，筛选出聚类分析结果中包含的文本信息；利用基于特征统计的 TFIDF 词频特征计算方法统计类别特征，获取类别特征词，选择名词作为候选类别特征词，按照候选特征词权重排序，以权重值较大的候选特征词作为类别关键词，利用类别关键词之间的语义关系，形成分类结果；识别和建立新的网络舆情主题，检测、跟踪已有舆情主题的相关内容；

d. 最后，舆情信息分析模块把舆情信息经过步骤 c 挖掘的数据进行 OLAP 多维统计分析，分析舆情主题内容关注度、舆情主题情感倾向等舆情评测指标。

2. 根据权利要求 1 所述的基于文本语义相关的网络舆情信息分析方法，其特征是，在步骤 a 中，所述舆情信息采集模块，是对网络舆情信息源进行采集，不仅要完成网页的爬取，而且要将网页内容进行格式化处理，提取舆情的主题和内容，所得数据存入 txt 格式或 html 格式文件，并存储到舆情信息数据库；网络舆情信息采集模块采用分时访问、定时更换 IP 地址和模拟浏览器进行单点登录三种技术结合进行防屏蔽。

3. 根据权利要求 2 所述的基于文本语义相关的网络舆情信息分析方法，其特征是，所述舆情信息采集模块执行的具体步骤为，从预先定义的主题相关网页的 URL 开始，获取网页中的文本信息，并从当前网页中抽取新的 URL 放入队列中，直到满足条件的舆情信息采集完毕，URL 队列为空为止；将采集到的网页文本信息按照字段分类存储到舆情信息数据库中，提供舆情信息萃取模块调用。

4. 根据权利要求 1 所述的基于文本语义相关的网络舆情信息分析方法，其特征是，在步骤 b 中，所述舆情信息萃取模块，是清除网页中的无关内容，提取对舆情分析有用的正文部分的元信息，对文本进行重构，将具有主题代表性的信息聚集在一起；所述舆情信息预处理模块，是对采集的舆情信息源经过所述舆情信息萃取模块萃取后，进行中文分词处理、过滤停用词、命名实体识别、词性标注、语法解析和特征词提取，建立正序索引和倒排索引；建立文本特征语义网络图，以文本中包含的实体 E 作为图的节点，两个实体之间的语义关系作为图的有向边，实体之间的语义关系结合词频信息作为节点的权重，有向边的权重表示实体关系在文本中的重要程度，所述实体 E 包括事物实体 NE、事件实体 VE、事件关系实体 RE；统计文本的词频和文本频率信息，然后进行特征词抽取，选取体现文本特征的词表示该文本。

5. 根据权利要求 4 所述的基于文本语义相关的网络舆情信息分析方法，其特征是，在步骤 c 中，所述舆情信息挖掘模块，是在对文本集进行预处理，包括中文分词处理、停用词过滤和结构化标签信息分析后，将信息萃取模块生成的文本数据集，根据文本特征语义网络图构建的文本语义特征描述结构，利用相似度评价方法计算文本之间的语义相似度，构建相似度矩阵，采用基于语义 相似度的改进文本聚类分析方法生成聚类结果；聚类分

析结果生成类别描述信息，筛选出聚类分析结果中包含的文本信息；利用基于特征统计的TFIDF词频特征计算方法统计类别特征，获取候选类别特征词，选择名词作为候选类别特征词，按照候选特征词权重排序，以权重值确定候选特征词作为类别关键词，利用类别关键词之间的语义关系，形成分类结果；将挖掘结果构建知识库。

6. 根据权利要求4或5所述的基于文本语义相关的网络舆情信息分析方法，其特征是，文本特征语义网络图是利用实体及其语义关系来表达舆情信息的有向图，通过网络节点表示的词语合并，节点权值相加；再合并有向边，有向边权值相加，构建文本特征语义网络图，描述文本中的语义信息和主题特征。

7. 根据权利要求5所述的基于文本语义相关的网络舆情信息分析方法，其特征是，文本之间的语义相似度评价方法为：

设经过步骤b的舆情信息萃取和预处理后的文本为$D_1(t_{11}, t_{12}, t_{13}, \cdots, t_{1m})$，$D_2(t_{21}, t_{22}, t_{23}, \cdots, t_{2m})$，计算文本$D_1$中所有关键词$t_{1i}$与文本$D_2$中所有关键词$t_{2i}$的相似度，形成相似度矩阵如下：

$$M(D_1,D_2)=\begin{pmatrix} Sim_{11} & Sim_{1m} \\ Sim_{m1} & Sim_{mm} \end{pmatrix}$$

$Sim_{ij}(1=i, j=m)$表示文本D_1关键词t_{1i}与文本D_2关键词t_{2j}的相似度；$M(D_1, D_2)$表示文本D_1与文本D_2之间的相似度矩阵；i为文本D_1的关键词数；m为文本D_2的关键词数；

词语相似度计算公式$S(T_1, T_2)=Max_{(i=1,2,\cdots,n;j=1,2,\cdots,m)}S(y_{1i}, y_{2j})$，即词语相似度为两词语所有义项相似度中的最大值，所述义项是指一个词语所包含的多个词义；

依次遍历相似度矩阵M，找到相似度Sim值最大的关键词对应组合，并删除对应的行和列；然后继续遍历相似度矩阵M找到Sim值最大的关键词组合，反复 循环直至矩阵M为零值矩阵；最后利用得到的相似度最大关键词组合序列，求得文本D_1和D_2的语义相似度，计算公式如下：

$$Sim(D_1,D_2)=\frac{1}{m}\sum_{k=1}^{m} Sim_{k-\max}(t_{1i}, t_{2j})$$

其中，max为相似度Sim的最大值；i为文本D_1的关键词数；j为文本D_2的关键词数。

8. 根据权利要求7所述的基于文本语义相关的网络舆情信息分析方法，其特征是，基于语义相似度的改进文本聚类分析方法为：

1）首先对所有采集的文本经过预处理后，采用TFIDF加权法对所有类别关键词进行特征加权，提取m个最优特征关键词形成原始的基于关键词特征向量Di*；

2）依据所述知识库对原始的基于关键词特征向量Di*中关键词进行预处理：在知识库中找到与关键词匹配的词汇并将其替换，形成新的特征向量D_i，$D_i=(T_1, T_2, \cdots, T_i)$，i=1,2,3,…,m；

3）形成n个文本的m个特征向量D_i，利用文本语义相似度计算公式计算采集的文本之间的语义相似度，形成文本集的相似度矩阵M，并求出所有特征向量的平均相似度MA；计算公式如下：

$M=\begin{pmatrix} S_{11} & S_{1n} \\ S_{n1} & S_{nn} \end{pmatrix}$，$MA=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}Sij-n\sum_{i=1}^{n}Sii}{n*(n-1)}$；其中，n 为文本数；

4）设定三个相似度阈值，一个重复度阈值为 0.9，一个主题中心阈值为 0.5，以及一个新主题阈值为 0.3；

5）将文本与中心主题比较，如果文本与中心主题的初始中心相似度大于重复度阈值 0.9，认为该文本属于同一主题同一内容文本；如果相似度小于新主题 阈值 0.3，则该文本需要新建一个类；如果相似度在 0～0.5 范围内，则该文本属于同一主题的不同侧面讨论的核心内容文本，标记为第二个中心，以此类推，形成多个中心的层次化的聚类结果；

6）针对多个中心的主题表示方法，选择文本与类内每个中心的相似度的最大值作为该类文本的相似度。

9. 根据权利要求 1 所述的基于文本语义相关的网络舆情信息分析方法，其特征是，在步骤 d 中，所述舆情信息分析模块，是对已存入舆情信息数据库中的经过步骤 c 挖掘的数据进行 OLAP 多维统计分析。

基于文本语义相关的网络舆情信息分析方法

技术领域

[0001] 本发明涉及网络信息技术领域，具体是一种基于文本语义相关的网络舆情信息分析方法。

背景技术

[0002] 当今社会，互联网已经渗透到人们的日常生活中，微博、论坛、博客等即时通信工具已经成为人们获取信息，进而发表看法、传播信息的重要渠道。借助网络平台，舆情信息迅速传播，引起广泛关注，其传播的速度之快、范围之广、影响力之大，远非传统媒体可比，网络空间的匿名交互性、非时空限制性等特点，使网络舆情这股强大的社会舆论力量，对社会发展和稳定产生一定的冲击和影响。正面的网络舆情似“正能量”，推动和促进社会发展；负面的网络舆情对社会稳定形成负面效应，引发舆情危机。由此，加强网络舆情信息监测、分析、管理，对稳定社会秩序、构建和谐社会具有重要的现实意义。对网络舆情信息及时监测、正确判断决策、迅速及时回应，积极采取有效措施化解舆情危机，成为网络舆情管理工作的重点和难点问题。

发明内容

[0003] 针对上述背景技术中网络舆情信息的特点和网络舆情信息管理中需要解决的问题，本发明提供一种基于文本语义相关的网络舆情信息分析方法。

[0004] 本发明解决其技术问题所采用的技术方案是，一种基于文本语义相关的网络舆情信息分析方法。采用包括网络舆情信息采集模块、舆情信息萃取模块、舆情信息预处理模块、舆情信息挖掘模块、舆情信息分析模块和包含舆情信息数据库的网络舆情信息分析系统，并包括如下步骤：

[0005] a. 网络舆情信息采集模块从网页中采集各种舆情信息，并存储到舆情信息数据库中；

[0006] b. 舆情信息萃取模块和舆情信息预处理模块将步骤 a 采集的舆情信息进行初步过滤和切分，抽取文本所包含的内容信息，为舆情信息挖掘提供数据服务；

[0007] c. 在步骤 b 基础上，舆情信息挖掘模块采用基于语义相似度的改进文本聚类分析方法，生成类别描述信息，筛选出聚类分析结果中包含的文本信息；利用基于特征统计的 TFIDF 词频特征计算方法统计类别特征，获取类别特征词，选择名词作为候选类别特征词，按照候选特征词权重排序，以权重值较大的候选特征词作为类别关键词，利用类别关键词之间的语义关系，形成分类结果；识别和建立新的网络舆情主题，检测、跟踪已有舆情主题的相关内容；

[0008] d. 最后，舆情信息分析模块把舆情信息经过步骤 c 挖掘的数据进行 OLAP 多维统计分析，分析舆情主题内容关注度、舆情主题情感倾向等舆情评测指标。

[0009] 在步骤 a 中，所述舆情信息采集模块，是对网络舆情信息源进行采集，与一般的网络爬虫不同的是，它不仅要完成网页的爬取，而且要将网页内容进行格式化处理，提取舆情

的主题和内容，所得数据存入 txt 格式或 html 格式文件，并存储到舆情信息数据库；网络舆情信息采集模块采用分时访问、定时更换 IP 地址和模拟浏览器进行单点登录三种技术结合进行防屏蔽。网络舆情信息采集模块采用分时访问、定时更换 IP 地址和模拟浏览器进行单点登录三种技术结合进行防屏蔽。网络舆情信息采集模块执行的具体步骤为：所述舆情信息采集模块执行的具体步骤为，从预先定义的主题相关网页的 URL 开始，获取网页中的文本信息，并从当前网页中抽取新的 URL 放入队列中，直到满足条件的舆情信息采集完毕，URL 队列为空为止；将采集到的网页文本信息按照字段分类存储到舆情信息数据库中，提供舆情信息萃取模块调用。

[0010] 所述舆情信息萃取模块，是清除网页中的无关内容，如网页中的广告、导航信息、图片、版权说明等噪声数据，提取对舆情分析有用的正文部分的元信息，对文本进行重构，将具有主题代表性的信息聚集在一起；所述舆情信息预处理模块，是对采集的舆情信息源经过所述舆情信息萃取模块萃取后，进行中文分词处理、过滤停用词、命名实体识别、词性标注、语法解析和特征词提取，建立正序索引和倒排索引；建立文本特征语义网络图，以文本中包含的实体 E 作为图的节点，两个实体之间的语义关系作为图的有向边，实体之间的语义关系结合词频信息作为节点的权重，有向边的权重表示实体关系在文本中的重要程度，所述实体 E 包括事物实体 NE、事件实体 VE、事件关系实体 RE；统计文本的词频和文本频率信息，然后进行特征词抽取，选取体现文本特征的词表示该文本。

[0011] 在步骤 b 中，所述舆情信息萃取模块，是清除网页中的无关内容，提取对舆情分析有用的正文部分的元信息，对文本进行重构，将具有主题代表性的信息聚集在一起；所述舆情信息预处理模块，是对采集的舆情信息源经过所述舆情信息萃取模块萃取后，进行中文分词处理、过滤停用词、命名实体识别、词性标注、语法解析和特征词提取，建立正序索引和倒排索引；建立文本特征语义网络图，以文本中包含的实体 E 作为图的节点，两个实体之间的语义关系作为图的有向边，实体之间的语义关系结合词频信息作为节点的权重，有向边的权重表示实体关系在文本中的重要程度，所述实体 E 包括事物实体 NE、事件实体 VE、事件关系实体 RE；统计文本的词频和文本频率信息，然后进行特征词抽取，选取体现文本特征的词表示该文本。

[0012] 要实现网络舆情信息文本挖掘、自然语言处理等文本分析，首先要进行分词处理，借鉴国内中文分词领域的研究成果，使用中国科学院计算技术研究所 研制的 ICTCLAS 汉语词法分析系统所具有的词语切分、词性标注、命名实体识别等功能，通过对舆情信息文本内容进行分词，提取长度大于二的词语。在文本分词之后，过滤对计算机理解文本无用的停用词，保留名词、动词、名形词、动形词等词性的词，得到备选特征词集，有效减少索引的大小，增加检索效率，提高准确率。经过分词处理的文本文档，建立正序索引和倒排索引，实现用户的查询交互。文本经过分词、词性标注、去停用词后，建立文本的特征语义网络图，统计文本的词频和文本频率等信息，然后进行加权计算和特征抽取等。

[0013] 在步骤 c 中，所述舆情信息挖掘模块，是在对文本集进行预处理，包括中文分词处理、停用词过滤和结构化标签信息分析后，将信息萃取模块生成的文本数据集，根据文本特征语义网络图构建的文本语义特征描述结构，利用相似度评价方法计算文本之间的语义相似度，构建相似度矩阵，采用基于语义相似度的改进文本聚类分析方法生成聚类结果；聚类分析结果生成类别描述信息，筛选出聚类分析结果中包含的文本信息；利用基于特征统计

的 TFIDF 词频特征计算方法统计类别特征，获取候选类别特征词，选择名词作为候选类别特征词，按照候选特征词权重排序，以权重值确定候选特征词作为类别关键词，利用类别关键词之间的语义关系，形成分类结果；将挖掘结果构建知识库，知识库还可以设置成具有同时支持舆情主题发现、舆情倾向性分析等文本挖掘功能。

[0014] 在步骤 d 中，所述舆情信息分析模块，是对已存入舆情信息数据库中的经过步骤 c 挖掘的数据进行 OLAP 多维统计分析，分析舆情主题关注度、舆情内容敏感度、舆情传播扩散度、舆情发布影响度等舆情评测指标，为相关部门及时掌握舆情动态、适时发布舆情信息、做出正确决策提供支持。

[0015] 与现有技术相比，本发明具有以下有益效果：

[0016] 1. 当前网络舆情信息反映出了海量性、动态性、不完整性、表现形式多样性 等特点，而现有的舆情信息分析方法往往忽视了舆情信息文本内容的相关关系，导致舆情信息分析结果不准确；本发明采用构建舆情信息文本的文本特征语义网络图模型，在文本描述结构中引入词语语义关联及上下文语境之间的联系；结合基于语义相似度的改进文本聚类算法，挖掘分析出舆情信息文本中上下文语义相关的内容。

[0017] 2. 通过建立舆情信息文本的文本特征语义网络图，将舆情信息文本中词语间的上下文关系形成特征项和权重组成的有向图结构，在保留文本词语上下文信息结构的同时，强化了文本中词语上下文语义的内涵，较好地描述文本中隐含的语义信息和主题特征，解决文本中词语语义信息缺失的问题。

[0018] 3. 基于语义相似度的改进文本聚类算法适合于大规模网络环境下对动态数据的聚类分析和舆情主题热点发现，通过对文本语义相似度计算，构建文本语义相似度矩阵，深度挖掘出舆情信息文本中上下文语义相关的内容，及时检测、跟踪新的主题事件；采用类内多个中心的主题表示方法，选择文本与类内每个中心的相似度最大值作为该类文本的相似度，有效地提高了系统运行效率，随着文本数量的增加，聚类分析效果会更加明显。

附图说明

[0019] 图 1 是本发明实施例基于文本语义相关的网络舆情信息分析方法的工作流程图。

具体实施方式

[0020] 下面将结合附图和具体实施例对本发明做进一步说明。但本发明的实施方式不限于此。

[0021] 如图 1 所示，本发明的方法中，包括网络舆情信息采集模块、舆情信息萃取模块、舆情信息预处理模块、舆情信息挖掘模块、舆情信息分析模块和包含舆情信息数据库的网络舆情信息分析系统。其处理流程是：

[0022] (1) 舆情信息采集

[0023] 对网络舆情信息源进行采集，与一般的网络爬虫不同的是，它不仅要完成网页的爬取，而且要将网页内容进行格式化处理，提取有用的舆情信息，如舆情的主题和内容，所得数据存入 txt 格式或 html 格式文件，写入原始舆情信息数据库。具体步骤为：按照预设的网络舆情信息采集策略，从多个种子网页的 URL 开始，通过各类端口发送遵循 http 协议的指令(采用 GET 方法)；远程服务器根据申请指令的内容返回 HTML 类型的文档。舆情信

息采集模块收集返回文档中所有的信息后先保存至缓存，然后传送到数据库中保存，获取网页中的文本信息；在获取网页文本信息过程中，不断从当前网页中抽取新出现的超链接URL访问，并剔除已经访问过的超链接URL，如此反复循环，直到满足搜索策略的网页文本信息采集完毕，未访问的URL队列为空为止。将采集的网页文本信息按照字段分类存储到数据库中，提供舆情信息萃取模块调用。

[0024] 网络舆情信息采集模块通常采用分时访问、定时更换IP地址、模拟浏览器进行单点登录等多种技术结合的防屏蔽策略。针对许多网站如论坛、博客、微博等通过用户登录方式才能访问，这里采用模拟浏览器的策略较易实现，利用微软.NET开发工具Visual Studio2008提供的Web Browser控件为微软IE浏览器的API调用，利用SSO单点登录模拟提交用户名及密码登录，等待用户登录信息加载完成后，页面跳转至相应URL地址，通过提交关键词进行检索，获得所需网页的源文件。

[0025] 采集的网页文本信息包括Web内容信息、Web结构和使用记录信息两部分。 Web内容信息包含新闻标题、正文内容、评论信息等文本内容信息，Web结构和Web使用记录信息包含点击量、浏览量、评论量等统计信息。

[0026] (2)舆情信息萃取

[0027] 采集的网页信息含有广告、导航信息、图片、版权说明等噪声数据，对舆情信息分析来说真正需要的是正文部分的元信息，清除掉这些无关内容，提取对舆情信息分析有用的正文部分的元信息，为文本后续的挖掘、分析提供服务。具体流程如下：

[0028] (2-1)首先使用Tidy工具对正文网页进行HTML标记规范化，然后利用html parser工具构建HTML树，将HTML标记作为树的节点，这样表示便于对HTML代码的管理和操作，可以更好地对代码进行结构化挖掘。

[0029] (2-2)从采集的舆情信息源中提取标题、关键词、正文、长度、更新时间和URL等相关信息，标题可截取标签<TITLE>与</TITLE>之间的信息；关键词包含在HTML文件头部的META标签，可从META标签信息中提取；时间信息可通过模式匹配分析和网页分析提取。

[0030] (2-3)正文提取的具体步骤为：选择适当的关键词，获取相关网页的URL地址，通过访问URL地址所在的服务器，得到网页的HTML源代码；删除网页源代码中的无用标记行，保留网页主体内容；将HTML代码中的段落符号(如</p>、
等)替换为特殊符号(如*[/p]*、*[/br]*等)，回车符和换行符替换为行分隔符，采用行结构存储方式，保留网页内容格式；提取每一行HTML标记“<”与“>”之间的文本；用回车符替换特殊符号(如*[/p]*、*[/br]*等)，保持正文原有的段落；对结果字符串进行去除HTML特殊转义字符(如&quot、&lt等)处理，结合正则表达式，匹配并提取最终的正文结果。

[0031] 从采集的舆情信息源中提取标题、关键词、正文、长度、更新时间和URL 等相关信息后，舆情信息萃取模块还要实现文本信息的重构。

[0032] 文本重构通过分析网络新闻、论坛帖子、微博博文等舆情信息存在形式和文本的结构特征，将具有代表性话题的信息组成“主旨块”，其余部分的信息组成“内容块”，以提高聚类分析效果。

[0033] 对于网页新闻的文本重构，是把网页新闻的标题和首段信息组成“主旨块”，其余的新闻描述信息和评论内容组成“内容块”。

[0034] 对于论坛帖子的文本重构，是将帖子的标题和主帖组成“主旨块”，将回帖和跟帖

信息净化处理，去除没有汉字内容的帖子和使用常用评价词的帖子，选择若干条帖子构成“内容块”。

[0035] (3) 舆情信息预处理

[0036] 舆情信息萃取后，接下来进行中文分词处理、命名实体识别、词性标注、语法解析、特征词提取等预处理，将结果保存到数据库中。要实现网络舆情信息文本挖掘、自然语言处理等文本分析，首先要进行分词处理，借鉴国内中文分词领域的研究成果，采用中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS 进行文本的分词及词性标注，通过中文分词处理，提取长度大于二的词语。ICTCLAS 的功能有中文文本的分词、词性标注、新词识别等；使用角色模型（role model）的方法进行命名实体识别；同时支持用户根据需要定义个性化词典，不仅具有较高的分词精度，分词效果也较好。其实现代码如下：

[0037]

```
//生成ICTCLAS类实例
ICTCLAS clas=ICTCLAS.Getinstance();
List<ResultTerm>terms=clas.Segment(this.txtinput.Text);
StringBuilder sb_seg=new StringBuilder();
foreach(ResultTerm term in terms)
```

[0038]

```
{
//得到分词后的词语
sb_seg.Append(term.Word)
//得到分词后的词性
sb_seg.AppendFormat("/{0}",term.POSStr);
sb_seg.Append("");
}
```

[0039] 在文本分词之后，过滤对计算机理解文本无用的停用词，保留名词、动词、名形词、动形词等词性的词，得到备选特征词集，以避免文本的冗杂，有效减少索引的大小，增加检索效率，提高检索准确率。

[0040] 经过分词处理的文本，建立正序索引和倒排索引，实现用户的查询交互。对于正序索引，根据词频的排序，选择前 N 个词语表示文本，用哈希表表示为：< 文件名，关键词词组 >；建立正序索引后，搜索文本中的关键词，找出包含此关键词的所有文件名，建立文件名词组，可得倒排索引，用哈希表表示为：< 关键词，文件名词组 >。

[0041] 索引的建立和索引的检索服务基于 Apache 开源项目 Lucene 实现，Lucene 提供完整的查询引擎和索引引擎，文本分析引擎；采用 Hadoop 存储和管理海量的索引文件。

[0042] 索引的建立过程如下：

[0043] 1. 创建索引写对象 IndexWriter。该对象创建时需提供词汇解析器，不同的词汇解析器采用不同的词库。选用 ThesaurusAnalyzer，能够提取内容摘要；

[0044] 2. 为取自数据库中的每个结果集创建一个 Document 对象；

[0045] 3. 将结果集中的数据元分别创建一个 Field 对象，并添加到 Document 对象；

[0046] 4. 写入该 Document 对象。

[0047] 索引检索的过程为：首先创建查询解析器，该查询解析器需要 Field 对象名以及对应的词汇解析器等参数；再由查询解析器和关键字获得查询对象；通过查询对象获取检索的结果集，结果集由 Document 对象构成。

[0048] 文本经过分词、词性标注、去停用词后，建立文本的特征语义网络图，统计文本的词频和文本频率等信息，然后进行加权计算和特征抽取等。

[0049] 文本特征语义网络图是一种用实体及其语义关系来表达舆情信息的有向图，以文本中包含的实体 E(包括事物实体 NE、事件实体 VE、事件关系实体 RE)作为图的节点，两个实体之间的语义关系作为图的有向边，实体之间的语义关系结合词频信息作为节点的权重，有向边的权重表示实体关系在文本中的重要程度。通过网络节点权值的引入和基于概念的合并与简化，构建文本特征语义网络图，提取文本的核心语义。即通过网络节点表示的词语合并，节点权值相加；再合并有向边，有向边权值相加，构建文本特征语义网络图，描述文本中的语义信息和主题特征。具体概念描述如下：

[0050] C1：事物实体 NE 定义为 NE（id,concept,property,power)。id 代表实体标识，concept 代表实体概念，property 代表实体属性，power 代表权重。

[0051] C2：事件实体 VE 定义为 VE（id,concept,property,power,isN,subT,objT1,objT2)。除了包含 NE 的几个数据项外，isN 代表是否为否定，subT 代表主体实体表头，objT1 和 objT2 代表客体实体 1 与 2 的表头。

[0052] C3：事件关系实体 RE 定义为 RE（id,concept,property,power,isN,subT,objT)。RE 用一对主客体实体就可完全描述。

[0053] 文本特征语义网络图模型分析步骤如下：

[0054] S1：在分析文本时，首先以语句为单位，构建各条语句对应的特征语义网 络图。逐句分析每句产生了哪些 NE，将 NE 及其属性信息记入实体信息表。

[0055] S2：NE 分析完毕后，分析 VE，登记 VE 的概念，属性，主体和客体。主客体相同的 VE 实体表示为同一 VE，否则设置不同的 id。

[0056] S3：接下来分析 RE。分析 RE 要注意与 NE、VE 区分开来，把 RE 的概念、属性、主体、客体登记到实体信息表。

[0057] S4：分析结束后，得到该语句的实体信息表。实体信息表描述了实体之间的关系，用来构造实体关系图，NE 与 VE 之间，RE 与 NE、VE 之间，实体 E 与属性 T 之间通过不同的连线把实体关系可视化。

[0058] S5：在分析构建第一条语句的特征语义网络图基础上，将后续语句的特征语义网络图合并，先合并节点，再合并有向边。

[0059] S6：合并节点时，把节点之间词语相同或者语义相似度满足阈值条件的节点合并，节点权值相加；否则保留该节点。

[0060] S7：有向边合并，是把合并后的节点间存在的有向边进行合并，有向边权值相加。

[0061] S8：更新新合并节点邻接边的权值为该节点的权值，强化节点之间的语义关系。

[0062] S9：输出所有合并语句的特征语义网络图后，完成整个文本的特征语义网络图的构造。

[0063] 下一步对词性特征权重赋值，以准确标示文本。按照汉语词性特点及完整事件描述要素(时间、地点、人物以及事件内容)，结合中国科学院汉语词性标记集，文本特征权重赋值分为：标题权重值为 3，子标题和关键词权重值为 2，摘要权重值为 1.5，段首句和段尾句权重值为 1.3。

[0064] 舆情信息经过预处理后，为文本的标题、正文和回复设置不同的标签，在 计算权重时，读取关键词的标签信息，完成词语的位置权重的赋值。

[0065] (4) 舆情信息挖掘

[0066] 舆情信息挖掘模块，是在对文本集进行预处理，包括中文分词处理、停用词过滤和结构化标签信息分析后，将信息萃取模块生成的文本数据集，根据文本特征语义网络图构建的文本语义特征描述结构，利用相似度评价方法计算文本之间的语义相似度，构建相似度矩阵，采用基于语义相似度的改进文本聚类分析方法生成聚类结果；聚类分析结果生成类别描述信息，筛选出聚类分析结果中包含的文本信息；利用基于特征统计的 TFIDF 词频特征计算方法统计类别特征，获取候选类别特征词，选择名词作为候选类别特征词，按照候选特征词权重排序，以权重值确定候选特征词作为类别关键词，利用类别关键词之间的语义关系，形成分类结果；将挖掘结果构建知识库，知识库还可以设置成具有同时支持舆情主题发现、舆情倾向性分析等文本挖掘功能。

[0067] 首先定义和计算文本之间的相似度，即文本之间所讨论主题的相关程度，用 $Sim(D_1,D_2)$ 表示文本 D_1 和文本 D_2 之间的相似度。相似度取值范围在 0 和 1 之间，与文本 D_1 和 D_2 的相似程度成正比。文本之间的相似度越大，表明文本之间的主题相关程度越大。文本之间的语义相似度评价方法如下：

[0068] 设经过步骤 b 的舆情信息萃取和预处理后的文本为 $D_1(t_{11}, t_{12}, t_{13}, \cdots, t_{1m})$，$D_2(t_{21}, t_{22}, t_{23}, \cdots, t_{2m})$，计算文本 D_1 中所有关键词 t_{1i} 与文本 D_2 中所有关键词 t_{2i} 的相似度，形成相似度矩阵如下：

[0069] $$M(D_1,D_2)=\begin{pmatrix} Sim_{11} & Sim_{1m} \\ Sim_{m1} & Sim_{mm} \end{pmatrix}$$

[0070] $Sim_{ij}(1=i, j=m)$ 表示文本 D_1 关键词 t_{1i} 与文本 D_2 关键词 t_{2j} 的相似度；$M(D_1,D_2)$ 表示文本 D_1 与文本 D_2 之间的相似度矩阵；i 为文本 D_1 的关键词数；m 为文本 D_2 的关键词数；

[0071] 词语相似度计算公式为：$S(T_1,T_2)=Max(i=1,2,\cdots,n;j=1,2,\cdots,m)S(y_{1i},y_{2j})$，即词语相似度为两词语所有义项(一个词语所包含的多个词义)相似度中的最大值。

[0072] 依次遍历相似度矩阵 M，找到相似度 Sim 值最大的关键词对应组合，并删除对应的行和列。然后继续遍历相似度矩阵 M 找到相似度值最大的关键词组合，反复循环直至矩阵 M 为零值矩阵。最后利用得到的相似度最大关键词组合序列，求得文本 D_1 和 D_2 的语义相似

度，计算公式如下：

[0073] $$\mathrm{Sim}(D_1,D_2)=\frac{1}{m}\sum_{k=1}^{m}Sim_{k-\max}(t_{1i},t_{2j})$$

[0074] 其中，max 为相似度 Sim 的最大值；i 为文本 D_1 的关键词数；j 为文本 D_2 的关键词数。

[0075] 基于语义相似度的改进文本聚类分析方法，描述如下：

[0076] 1. 首先对所有采集的文本经过预处理后，采用 TFIDF 加权法对所有类别关键词进行特征加权，提取 m 个最优特征关键词形成原始的基于关键词特征向量 Di*。

[0077] 2. 依据所述知识库对原始的基于关键词特征向量 Di* 中关键词进行预处理：在知识库中找到与关键词匹配的词汇并将其替换，形成新的特征向量 D_i，$D_i=(T_1,T_2,\cdots,T_i)$，i=1,2,3,…,m。

[0078] 3. 形成 n 个文本的 m 个特征向量 D_i，利用文本语义相似度计算公式计算采集的文本之间的语义相似度，形成文本集的相似度矩阵 M，并求出所有特征向量的平均相似度 MA。计算公式如下：

[0079] $$M=\begin{pmatrix} S_{11} & S_{1n} \\ S_{n1} & S_{nn} \end{pmatrix},\quad MA=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}Sij-n\sum_{i=1}^{n}Sii}{n*(n-1)}$$；其中，n 为文本数；

[0080] 4. 设定三个相似度阈值，一个重复度阈值为 0.9，一个主题中心阈值为 0.5，以及一个新主题阈值为 0.3；

[0081] 5. 将文本与中心主题比较，如果文本与中心主题的初始中心相似度大于重复度阈值 0.9，认为该文本属于同一主题同一内容文本；如果相似度小于新主题阈值 0.3，则该文本需要新建一个类；如果相似度在 0～0.5 范围内，则该文本属于同一主题的不同侧面讨论的核心内容文本，标记为第二个中心，以此类推，形成多个中心的层次化的聚类结果。

[0082] 6. 针对多个中心的主题表示方法，选择文本与类内每个中心的相似度的最大值作为该类文本的相似度。

[0083] 基于语义相似度的改进文本聚类算法适合于大规模网络环境下对动态数据的聚类分析和舆情主题热点发现，能及时检测到新事件，检测、跟踪新的舆情主题；采用类内多个中心的舆情主题表示方法，有效地提高了系统运行效率，随着文本数量的增加，效果会更加明显。

[0084] 5) 舆情信息分析

[0085] 所述舆情信息分析模块对已存入舆情信息数据库中的经过步骤 c 挖掘的数据进行 OLAP 多维统计分析，分析舆情主题内容关注度、舆情主题情感倾向等舆情评测指标，为相关部门及时掌握舆情动态、适时发布舆情信息、做出正确决策提供支持。

[0086] 通过采集、处理和挖掘分析产生的舆情主题，表示为：$T=(T_1,T_2,\cdots,T_n)$，其中 T_i 表示舆情主题的文本。舆情主题文本的关注度表示为：$T_i=\alpha N_p+\beta N_r$，舆情 主题的关注度度量公式为：$F=\sum_{i=1}^{i=n}(T_i)=\sum_{i=1}^{i=n}(\alpha Np\_i+\beta Nr\_i)$，其中 α，β 表示权重，$N_p$ 表示舆情主题文本的点击数，N_r 表示评论数；Np_i 表示第 i 个舆情主题文本的点击数，Nr_i 表示第 i 个舆情主题

文本的评论数。由于 $N_p > N_r$，经过统计，α 取值为 0.02，β 取值为 0.98。

[0087] 舆情主题的情感倾向基于舆情主题文本的聚类分析数据描述。首先设定一个阈值，只有当文本的倾向度量值大于阈值，文本才表现出极性（正面性、负面性）。文本的倾向度量值为正，则该文本为正面的评论，反之则为负面的评论。

[0088] 舆情信息经过采集、预处理、信息萃取、挖掘和分析，可以得到舆情主题的详细数据，按照建立的舆情指标评价体系进行处理，处理的结果提供决策帮助。

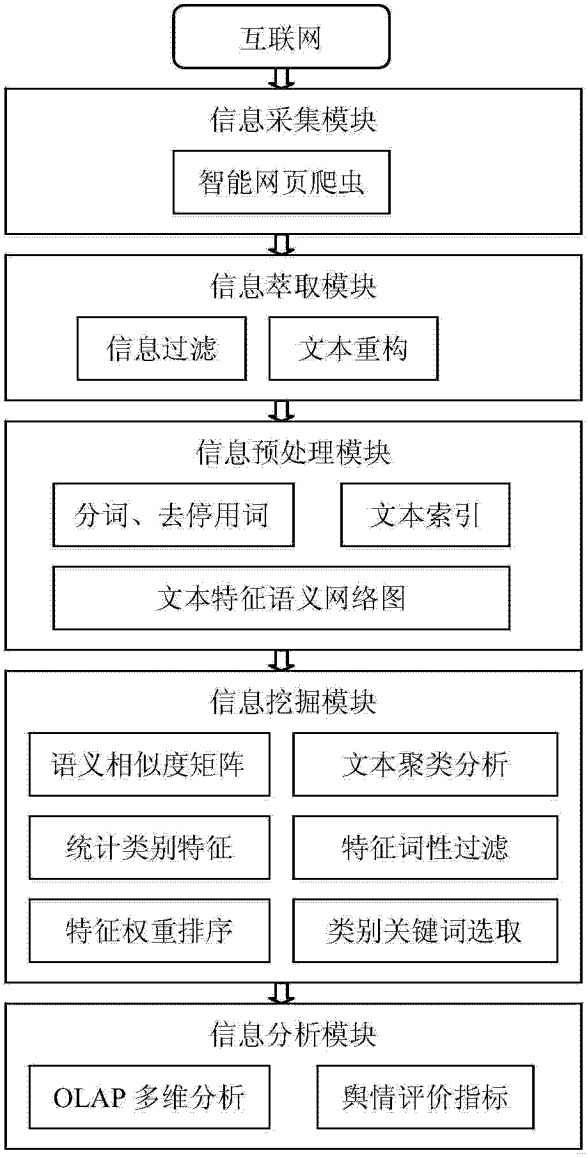

互联网
信息采集模块
智能网页爬虫
信息萃取模块
信息过滤
文本重构
信息预处理模块
分词、去停用词
文本索引
文本特征语义网络图
信息挖掘模块
语义相似度矩阵
文本聚类分析
统计类别特征
特征词性过滤
特征权重排序
类别关键词选取
信息分析模块
OLAP 多维分析
舆情评价指标

图 1