(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 107093021 A
(43)申请公布日 2017.08.25

(21)申请号 201710267800.3

(22)申请日 2017.04.21

(71)申请人 深圳市创艺工业技术有限公司
地址 518000 广东省深圳市南山区南山街道深南大道10128号南山数字文化产业基地南山软件园西塔楼1708

(72)发明人 不公告发明人

(51)Int.Cl.
G06Q 10/06(2012.01)
G06Q 50/06(2012.01)

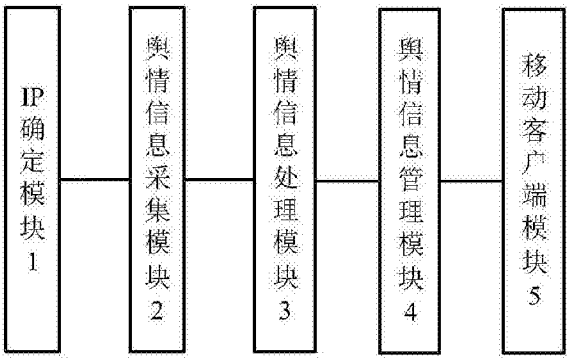
权利要求书3页 说明书6页 附图2页

(54)发明名称

电网工程物资合同履约诚信舆情监控系统

(57)摘要

电网工程物资合同卖方履约诚信舆情监控系统，包括IP确定模块、舆情信息采集模块、舆情信息处理模块、舆情信息管理模块和移动客户端，所述IP确定模块用于确定与电力行业相关的网页IP，述舆情信息采集模块用于搜集互联网上和电网工程物资合同履约情况相关的网页，所述舆情信息处理模块用于提取采集的网页信息的正文部分并进行网页特征项的提取和权重的计算，从而判断采集得到的网页是否为主题相关网页，所述用户管理模块用于向用户显示主题相关的网页，所述移动客户端为安装有相关应用的智能手机或平板电脑，用户可以通过移动客户端模块实时访问舆情信息管理模块的数据库单元，本发明的有益效果为：通过对网络舆情的监控，实时了解电网工程物资合同卖方履约诚信情况。



CN 107093021 A

1.电网工程物资合同卖方履约诚信舆情监控系统，其特征是，包括IP确定模块、舆情信息采集模块、舆情信息处理模块、舆情信息管理模块和移动客户端模块；

(1)IP确定模块：用于确定与电力行业相关的网页IP；

(2)舆情信息采集模块：设定IP确定模块确定的IP链接为种子链接，采用主题爬虫策略从此种子链接出发，搜集和电网工程物资合同履约主题相关的网页；

(3)舆情信息处理模块：用于提取采集的网页信息中的正文部分，从所述正文部分的中文分词结果中提取具有代表性的网页特征项并计算相应特征项的权重，从而进行网页分类；

(4)舆情信息管理模块：包括数据库单元、用户登录单元和信息检索单元，所述数据库单元用于存储舆情信息处理模块确定的主题相关网页，用户通过用户登录单元输入密码登录舆情信息管理模块，并可通过信息检索单元输入要检索的关键词进行电网工程物资合同履约情况的信息检索，信息检索单元即显示包含所述关键词的相关网页；

(5)移动客户端模块：为安装有相关应用的智能手机或平板电脑，用户可以通过移动客户端模块实时访问舆情信息管理模块的数据库单元，从而了解电网工程物资合同履约诚信情况。

2.根据权利要求1所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述舆情信息采集模块包括主题设置单元、电力猫接入单元和舆情信息采集单元，所述主题设置单元用于根据本系统的主题，设置主题关键词和主题爬虫的初始链接，所述电力猫接入单元用于判断电脑通过电力猫接入网络时，即令舆情信息采集单元采用主题爬虫策略搜集主题相关的网页。

3.根据权利要求2所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述主题设置单元用于根据本系统的主题，设置主题初始关键词和主题爬虫的初始链接，具体包括：

a.根据本系统的主题，设置主题初始关键词组G＝{“电网”、“电力物资”、“履约”、“物资合同”，“违约”}，设置关键词的权重分别为q_{g1}、q_{g2}、q_{g3}、q_{g4}和q_{g5}，则主题文档可以初步表示为$W_g=(t_{g1},t_{g2},t_{g3},t_{g4},t_{g5})$，其中，$t_{g1}$、$t_{g2}$、$t_{g3}$、$t_{g4}$、$t_{g5}$分别代表关键词电网、电力物资、履约、物资合同和违约；

b.设置IP确定模块所确定的IP为种子链接，主题爬虫从此种子链接出发，搜集主题相关网页；

c.从舆情处理模块确定的各个主题相关网页中，提取权重较高的前h个特征项加入关键词组G，形成新的关键词组G。

4.根据权利要求3所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述舆情信息处理模块包括正文提取单元、特征项提取单元和网页分类单元，所述正文提取单元用于根据电力行业网站的网页特点，采用文本分割的方式提取网页的正文部分，所述特征项提取单元用于从正文部分的中文分词结果中提取具有代表性的特征项并计算特征项在文本中的权重，所述网页分类单元用于判断采集得到的网页是否为主题相关网页。

5.根据权利要求4所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述正文提取单元用于根据电力行业网站的网页特点，采用文本分割的方式提取网页的正文部分，具体包括：

a.将采集得到的电力行业网站的网页进行滤波处理，去除网页中的噪声部分；

b.从网页源文件中按顺序提取文本块，得到文本块集合A＝{a_1,a_2,……a_n}，对文本块集合中的每个文本块的字符数进行统计，并将统计结果存入数组B_i对应的位置处，数组B_i＝{b_1,b_2,……b_n}，对数组B_i进行处理，其计算公式为：

$$b_i = \frac{\alpha_1 b_{i-1} + \alpha_2 b_i + \alpha_3 b_{i+1}}{\alpha_1 + \alpha_2 + \alpha_3} \qquad (i = 1,2,\ldots\ldots,n)$$

式中，b_{i-1}、b_i、b_{i+1}分别为文本块i-1、i、i+1的字符数总数，α_1、α_2、α_3分别为b_{i-1}、b_i、b_{i+1}的权重，且α_1、α_2、α_3>0；

c.定义文本块分界阈值f_1和f_2，则f_1和f_2分别为：

$$f_1 = \frac{\min\limits_{1\le i\le \frac{n}{v}} b_i + \min\limits_{\frac{n}{v}\le i\le 2*(\frac{n}{v})} b_i + \cdots + \min\limits_{(v-1)*(\frac{n}{v})\le i\le n} b_i}{v}$$

$$f_2 = \rho_1 \frac{\sum_{i=0}^{n} b_i}{n} + \rho_2 f_1$$

式中，n为文本块的总数，b_i为文本块i中的字符总数，ρ_1和ρ_2分别为文本块中的平均字符数和f_1的权重，ρ_1、ρ_2>0，且ρ_1+ρ_2＝1，v为对数组B_i的分组数；

d.根据文本块字符数与分界阈值之间的关系进行正文部分提取，定义文本块子集C＝{b_i,b_{i+1},……,b_{i+m}}，其中i+m≤n，且C∈B，则当文本块子集C满足{b_i,b_{i+1},……,b_{i+m}}中的值全部大于f_1且{b_i,b_{i+1},……,b_{i+m}}中的值大于f_2的个数$k \ge \frac{m}{2}$时，则文本块子集C为网页正文部分。

6.根据权利要求4所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述特征项提取单元用于从正文部分的中文分词结果中提取具有代表性的特征项并计算特征项在文本中的权重，具体包括：

a.采用一种改进的信息增益计算方法进行特征项的选择，定义采集得到的网页中类别为C_i(1≤i≤m)的文本有{w_{i1},w_{i2},……w_{ie}}，则改进的信息增益IG(C_i,t_j)的计算方法为：

$$IG(C_i,t_j) = \frac{u(t_j,C_i)}{\sum_{i=1}^{m} u(t_j,C_i)} \left(\sum_{i=1}^{m} \left[P(t_j|C_i) ln \frac{p(t_j|C_i)}{p(t_j)p(C_i)} + p(\overline{t_j}|C_i) ln \frac{p(\overline{t_j}|C_i)}{p(\overline{t_j})p(C_i)} \right) ln \left(1+\sum_{k=1}^{e} \frac{q_{ik}(t_j)}{\max\limits_{1\le k\le e} q_{ik}(t_j)} \right)$$

式中，u(t_j,C_i)为C_i类文本出现的特征词t_j的次数，p(t_j)为特征词t_j出现的概率，则$p(\overline{t_j})$为t_j不出现的概率，P(t_j|C_i)为特征词t_j存在的文本属于C_i类的概率，$p(\overline{t_j}|C_i)$为t_j不存在的文本属于C_i类的概率，m为类别数，p(C_i)为C_i类文本出现的概率，$q_{ik}(t_j)$为特征项t_j在文本w_{ik}(1≤k≤e)中出现的频率，IG(C_i,t_j)为特征词t_j的信息增益值；

将计算所得的特征词的信息增益值按从大到小顺序排列，选取前n个特征词作为文本的特征项；

b.定义文本W_i的特征项为{t_1,t_2,……t_n}，则对应特征项的权重{w_{i1},w_{i2},……w_{in}}的计算公式为：

$$w_{ij}=\frac{\sqrt{\frac{p_{ij}}{e^{\sum_{r=1}^{b}p_{rj}}}}*ln(\frac{H}{s_j}+1)}{\sqrt{\sum_{j=1}^{n}(p_{ij}*ln(\frac{H}{s_j}+1))^2}}\quad(j=1,2\ldots\ldots n)$$

式中，w_{ij}表示特征项t_j在文本中W_i中的权重，而p_{ij}表示特征项t_j在文本W_i中出现的频率，H表示全部文本集中的文本数量，s_j表示文本集中包含t_j的文本数，b表示文本集中文本的数量，p_{rj}表示特征项t_j在文本W_r中出现频率。

7.根据权利要求4所述的电网工程物资合同卖方履约诚信舆情监控系统，其特征是，所述网页分类单元用于判断采集得到的网页是否为主题相关网页，定义采集得到的文档为$W_i=(t_{i1},t_{i2},t_{i3},\cdots\cdots t_{in})$，描述主题文档为$W_g=(t_{g1},t_{g2},t_{g3},\cdots\cdots t_{gv})$，则文档$W_i$和文档$W_g$之间的文档相似性系数$\omega_{ig}$的计算公式为：

$$f_{(t_{i1},t_{gl})}=\max_{1\le l\le v}sim(t_{i1},t_{gl})$$

$$f_{(t_{i2},t_{gl})}=\max_{1\le l\le v}sim(t_{i2},t_{gl})$$

……

$$f_{(t_{in},t_{gl})}=\max_{1\le l\le v}sim(t_{in},t_{gl})$$

$$\rho_{ig}=\frac{\sum_{k=1}^{n}f_{(t_{ik},t_{gl})}e^{(w_{ik}+q_{gl})}}{n*v}$$

式中，ρ_{ig}为文档W_i和文档W_g之间的文档相似性系数，$sim(t_{i1},t_{gl})$为特征项t_{i1}和特征项t_{gl}的概念词语相似度，$sim(t_{i2},t_{gl})$为特征项t_{i2}和特征项t_{gl}的概念词语相似度，$sim(t_{in},t_{gl})$为特征项t_{in}和特征项t_{gl}的概念词语相似度，$f_{(t_{i1},t_{gl})}$、$f_{(t_{i2},t_{gl})}$和$f_{(t_{in},t_{gl})}$表示文档W_i中的特征项t_{i1}、t_{i2}、t_{in}分别和文档W_g中所有特征项之间的概念词语相似度的最大值，w_{ik}和q_{gl}分别为特征项t_{ik}和t_{gl}的权重；

定义主题阈值为μ，判断系数为r，根据判断系数r的大小判断采集得到的文档W_i是否为与电网工程物资合同履约相关的网页，具体为：

$$r=ln^{\frac{\rho_{ig}}{\mu}}$$

$$\begin{cases}r>0, & \text{则为主题相关网页}\\ r<0, & \text{则为非主题相关网页}\end{cases}$$

式中，ρ_{ig}为文档W_i和文档W_g之间的文档相似性系数，μ为主题阈值；

当判断为主题相关网页时，即将网页送入数据库单元进行存储，并将主题相关网页中包含的链接加入主题爬虫的等待队列，当判断为非主题相关网页时即舍弃。

电网工程物资合同履约诚信舆情监控系统

技术领域

[0001] 本发明创造涉及舆情监控领域，具体涉及一种电网工程物资合同履约诚信舆情监控系统。

背景技术

[0002] 电网工程物资是整个电网的基础，而电网工程物资合同卖方的履约情况，而电网工程物资合同卖方的履约情况关系到整个电力系统的物资供应。近年来，随着国民经济的快速发展激增了用电量的需求，加大了电网工程项目的建设，因此也产生了大量的电网工程物资合同，传统的物资合同管理方法通常在签订合同后建立专业的合同管理机构和人员进行实时了解和监管物资合同的履约情况，这种方法不仅增加了工作人员的工作量，而且不能有效预防物资合同的违约情况。

[0003] 当今社会网络舆情的影响越来越大，受人民关注度越来越高的特点，通过对互联网舆情信息的获取和监控，便于通过民众的影响进行电网工程物资合同卖方履约情况的舆情监控，进一步实现良好的电网工程物资交易。

发明内容

[0004] 针对上述问题，本发明旨在提供一种电网工程物资合同履约诚信舆情监控系统。

[0005] 本发明创造的目的通过以下技术方案实现：

[0006] 电网工程物资合同卖方履约诚信舆情监控系统，包括IP确定模块、舆情信息采集模块、舆情信息处理模块、舆情信息管理模块和移动客户端模块；

[0007] (1) IP确定模块：用于确定与电力行业相关的网页IP；

[0008] (2) 舆情信息采集模块：设定IP确定模块确定的IP链接为种子链接，采用主题爬虫策略从此种子链接出发，搜集和电网工程物资合同履约主题相关的网页；

[0009] (3) 舆情信息处理模块：用于提取采集的网页信息中的正文部分，从所述正文部分的中文分词结果中提取具有代表性的网页特征项并计算相应特征项的权重，从而进行网页分类；

[0010] (4) 舆情信息管理模块：包括数据库单元、用户登录单元和信息检索单元，所述数据库单元用于存储舆情信息处理模块确定的主题相关网页，用户通过用户登录单元输入密码登录舆情信息管理模块，并可通过信息检索单元输入要检索的关键词进行电网工程物资合同履约情况的信息检索，信息检索单元即显示包含所述关键词的相关网页；

[0011] (5) 移动客户端模块：为安装有相关应用的智能手机或平板电脑，用户可以通过移动客户端模块实时访问舆情信息管理模块的数据库单元，从而了解电网工程物资合同履约诚信情况。

[0012] 本发明创造的有益效果：提出一种电网工程物资合同卖方履约诚信舆情监控系统，通过对具有较高可信度的新闻网页以及电力行业专用的门户网站的信息的抓取和科学有效的分析，得到了反应“电网工程物资合同履约”的舆情热点话题，实现了电网工程物资

合同卖方履约诚信的有效监控。

附图说明

[0013] 利用附图对发明创造作进一步说明，但附图中的实施例不构成对本发明创造的任何限制，对于本领域的普通技术人员，在不付出创造性劳动的前提下，还可以根据以下附图获得其它的附图。

[0014] 图1是本发明结构示意图；

[0015] 图2是本发明舆情信息采集模块结构示意图

[0016] 图3是本发明舆情信息处理模块结构示意图。

[0017] 图4是本发明舆情信息管理模块结构示意图。

[0018] 附图标记：

[0019] IP确定模块1、舆情信息采集模块2；舆情信息处理模块3；舆情信息管理模块4；移动客户端模块5；主题设置单元21；电力猫接入单元22；舆情信息采集单元23；正文提取单元31；特征项提取单元32；网页分类单元33；数据库单元41、用户登录单元42；信息检索单元43。

具体实施方式

[0020] 结合以下实施例对本发明作进一步描述。

[0021] 参见图1、图2、图3和图4，本实施例的电网工程物资合同卖方履约诚信舆情监控系统，包括IP确定模块1、舆情信息采集模块2、舆情信息处理模块3、用户管理模块4和移动客户端5；

[0022] (1) IP确定模块1：用于确定与电力行业相关的网页IP；

[0023] (2) 舆情信息采集模块2：设定所述IP确定模块1所确定的IP链接为种子链接，采用主题爬虫策略从此种子链接出发，搜集和电网工程物资合同履约主题相关的网页；

[0024] (3) 舆情信息处理模块3：用于提取采集的网页信息中的正文部分，从所述正文部分的中文分词结果中提取具有代表性的网页特征项并计算相应特征项的权重，从而进行网页分类；

[0025] (4) 舆情信息管理模块4：包括数据库单元41、用户登录单元42和信息检索单元43，所述数据库单元41用于存储舆情信息处理模块3确定的主题相关网页，用户通过用户登录单元42输入密码登录舆情信息管理模块4，并可通过信息检索单元43输入要检索的关键词进行电网工程物资合同履约情况的信息检索，信息检索单元43即显示包含所述关键词的相关网页；

[0026] (5) 移动客户端模块5：为安装有相关应用的智能手机或平板电脑，用户可以通过移动客户端模块5实时访问舆情信息管理模块4的数据库单元41，从而了解电网工程物资合同履约诚信情况。

[0027] 本优选实施例提出一种电网工程物资合同卖方履约诚信舆情监控系统，通过对互联网海量的信息的抓取和科学有效的分析，得到了反应“电网工程物资合同履约”的舆情热点话题，实现了电网工程物资合同卖方履约诚信的有效监控。

[0028] 优选地，所述舆情信息采集模块2包括主题设置单元21、电力猫接入单元22和舆情

信息采集单元23，所述主题设置单元21用于根据本系统的主题，设置主题初始关键词和主题爬虫的初始链接，所述电力猫接入单元22用于当判断电脑通过电力猫接入网络时，即令舆情信息采集单元23采用主题爬虫策略搜集主题相关的网页。

[0029] 本优选实施例构成了本系统的舆情信息采集模块，规定只有在判断电脑通过电力猫接入网络时即令舆情信息采集单元进行采集，此时的网络相对稳定，提高了爬虫的可靠性和效率，此外，避免了系统一直进行网页爬虫造成的电脑资源消耗。

[0030] 优选地，所述主题设置单元21采用主题爬虫策略搜集主题相关的网页，具体包括：

[0031] a.根据本系统的主题，设置主题初始关键词组G＝{“电网”、“电力物资”、“履约”、“物资合同”，“违约”}，设置关键词的权重分别为q_{g1}、q_{g2}、q_{g3}、q_{g4}和q_{g5}，则主题文档可以初步表示为$W_g=(t_{g1},t_{g2},t_{g3},t_{g4},t_{g5})$，其中，$t_{g1}$、$t_{g2}$、$t_{g3}$、$t_{g4}$、$t_{g5}$分别代表关键词电网、电力物资、履约、物资合同和违约；

[0032] b.设置IP确定模块1所确定的IP为种子链接，主题爬虫从此种子链接出发，搜集主题相关网页；

[0033] c.从舆情处理模块3确定的各个主题相关网页中，提取权重较高的前h个特征项加入关键词组G，形成新的关键词组G。

[0034] 本优选实施例在舆情信息采集模块通过设关键词和权值，可以确定爬虫的主题，在爬虫搜索的过程中，在通过添加主题相关网页中权重较高的特征项作为关键词，实现了最大程度的描述爬虫的主题范围。

[0035] 优选地，所述舆情信息处理模块3包括正文提取单元31、特征项提取单元32和网页分类单元33，所述正文提取单元31用于根据电力行业网站的网页特点，采用文本分割的方式提取网页的正文部分，所述特征项提取单元32用于从正文部分的中文分词结果中提取具有代表性的特征项并计算所述特征项在文本中的权重，所述网页分类单元23用于判断采集得到的网页是否为主题相关网页。

[0036] 优选地，所述正文提取单元31用于根据电力行业网站的网页特点，采用文本分割的方式提取网页的正文部分，具体包括；

[0037] a.将采集得到的电力行业网站的网页进行滤波处理，去除网页中的噪声部分；

[0038] b.从网页源文件中按顺序提取文本块，得到文本块集合$A=\{a_1,a_2,\cdots\cdots a_n\}$，对文本块集合中的每个文本块的字符数进行统计，并将统计结果存入数组B_i对应的位置处，数组$B_i=\{b_1,b_2,\cdots\cdots b_n\}$，对数组$B_i$进行处理，其计算公式为：

[0039] $$b_i=\frac{\alpha_1 b_{i-1}+\alpha_2 b_i+\alpha_3 b_{i+1}}{\alpha_1+\alpha_2+\alpha_3}\qquad(i=1,2,\ldots\ldots,n)$$

[0040] 式中，b_{i-1}、b_i、b_{i+1}分别为文本块i-1、i、i+1的字符数总数，α_1、α_2、α_3分别为b_{i-1}、b_i、b_{i+1}的权重，且α_1、α_2、$\alpha_3>0$；

[0041] c.定义文本块分界阈值f_1和f_2，则f_1和f_2分别为：

[0042] $$f_1=\frac{\min\limits_{1\le i\le\frac{n}{v}} b_i+\min\limits_{\frac{n}{v}\le i\le 2*(\frac{n}{v})} b_i+\cdots+\min\limits_{(v-1)*(\frac{n}{v})\le i\le n} b_i}{10}$$

[0043] $$f_2=\rho_1\frac{\sum_{i=0}^{n} b_i}{n}+\rho_2 f_1$$

[0044] 式中，n为文本块的总数，b_i为文本块i中的字符总数，ρ_1和ρ_2分别为文本块中的平均字符数和f_1的权重，ρ_1、ρ_2>0，且$\rho_1+\rho_2=1$，v为数组B_i的分组数；

[0045] d.根据文本块字符数与分界阈值之间的关系进行正文部分提取，定义文本块子集$C=\{b_i,b_{i+1},\cdots\cdots,b_{i+m}\}$，其中$i+m\leqslant n$，且$C\in B$，则当文本块子集C满足$\{b_i,b_{i+1},\cdots\cdots,b_{i+m}\}$中的值全部大于$f_1$且$\{b_i,b_{i+1},\cdots\cdots,b_{i+m}\}$中的值大于$f_2$的个数$k\geq\frac{m}{2}$时，则判断文本块子集C为网页正文部分。

[0046] 本优选实施例根据电力行业网站的网页特点，采用文本分割进行网页正文部分的提取，具有较高的提取精度并且有效减少文本块的遗漏，提高了本系统的监控精度。

[0047] 优选地，所述特征提取单元32用于从正文部分的中文分词结果中提取具有代表性的特征项并计算特征项在文本中的权重，具体为：

[0048] a.采用一种改进的信息增益计算方法进行特征项的选择，定义采集得到的网页中类别为C_i($1\leqslant i\leqslant m$)的文本有$\{w_{i1},w_{i2},\cdots\cdots w_{ie}\}$，则改进的信息增益$IG(C_i,t_j)$的计算方法为：

[0049]
$$IG(C_i,t_j)=\frac{u(t_j,C_i)}{\sum_{i=1}^{m}u(t_j,C_i)}\Big(\sum_{i=1}^{m}\Big[P(t_j|C_i)ln\frac{p(t_j|C_i)}{p(t_j)p(C_i)}+p(\overline{t_j}|C_i)ln\frac{p(\overline{t_j}|C_i)}{p(\overline{t_j})p(C_i)}\Big)ln\left(1+\sum_{k=1}^{e}\frac{q_{ik}(t_j)}{\max\limits_{1\leq k\leq e}q_{ik}(t_j)}\right)$$

[0050] 式中，$u(t_j,C_i)$为C_i类文本出现的特征词t_j的次数，$p(t_j)$为特征词t_j出现的概率，则$p(\overline{t_j})$为t_j不出现的概率，$P(t_j|C_i)$为特征词t_j存在的文本属于C_i类的概率，$p(\overline{t_j}|C_i)$为t_j不存在的文本属于C_i类的概率，m为类别数，$p(C_i)$为C_i类文本出现的概率，$q_{ik}(t_j)$为特征项t_j在文本w_{ik}($1\leqslant k\leqslant e$)中出现的频率，$IG(C_i,t_j)$为特征词t_j的信息增益值；

[0051] 将计算所得的特征词的信息增益值按从大到小顺序排列，选取前n个特征词作为文本的特征项；

[0052] b.定义文本W_i的特征项为$\{t_1,t_2,\cdots\cdots t_n\}$，则对应特征项的权重$\{w_{i1},w_{i2},\cdots\cdots w_{in}\}$的计算公式为：

[0053]
$$w_{ij}=\frac{\sqrt{e^{\frac{p_{ij}}{\sum_{r=1}^{b}p_{rj}}}}*ln(\frac{H}{s_j}+1)}{\sqrt{\sum_{j=1}^{n}(p_{ij}*ln(\frac{H}{s_j}+1))^2}}\quad(j=1,2\ldots\ldots n)$$

[0054] 式中，w_{ij}表示特征项t_j在文本中W_i中的权重，而p_{ij}表示特征项t_j在文本W_i中出现的频率，H表示全部文本集中的文本数量，s_j表示文本集中包含t_j的文本数，b表示文本集中文本的数量，p_{rj}表示特征项t_j在文本W_r中的出现的频率。

[0055] 本优选实施例提出在信息增益的计算过程中引进了词频和集中度这两个参数，增加了特征项的分类能力，有助于选出较有效的特征项，采用一种改进的权重计算方法，相较于传统的TFIDF计算权重方法，综合考虑了特征项对不同文本的影响程度大小，加大了文本之间的差异性，因此具有更优的分类效果。

[0056] 优选地，所述网页分类单元33用于判断采集得到的网页是否为主题相关网页，定

义采集得到的文档为采集得到的文档$W_i=(t_{i1},t_{i2},t_{i3},\cdots\cdots t_{in})$，主题文档为$W_g=(t_{g1},t_{g2},t_{g3},\cdots\cdots t_{gv})$，则文档$W_i$和文档$W_g$之间的文档相似性系数$\omega_{ig}$的计算公式为：

[0057] $$f_{(t_{i1},t_{gl})}=\max_{1\le l\le v} sim(t_{i1},t_{gl})$$

[0058] $$f_{(t_{i2},t_{gl})}=\max_{1\le l\le v} sim(t_{i2},t_{gl})$$

[0059] ……

[0060] $$f_{(t_{in},t_{gl})}=\max_{1\le l\le v} sim(t_{in},t_{gl})$$

[0061] $$\rho_{ig}=\frac{\sum_{k=1}^{n} f_{(t_{ik},t_{gl})} e^{(w_{ik}+q_{gl})}}{n*v}$$

[0062] 式中，ρ_{ig}为文档W_i和文档W_g之间的文档相似性系数，$sim(t_{i1},t_{gl})$为特征项t_{i1}和特征项t_{gl}的概念词语相似度，$sim(t_{i2},t_{gl})$为特征项t_{i2}和特征项t_{gl}的概念词语相似度，$sim(t_{in},t_{gl})$为特征项t_{in}和特征项t_{gl}的概念词语相似度，$f_{(t_{i1},t_{gl})}$、$f_{(t_{i2},t_{gl})}$和$f_{(t_{in},t_{gl})}$表示文档W_i中的特征项t_{i1}、t_{i2}、t_{in}分别和文档W_g中所有特征项之间的概念词语相似度的最大值，w_{ik}和q_{gl}分别为特征项t_{ik}和t_{gl}的权重；

[0063] 定义主题阈值为μ，判断系数为r，根据判断系数r的大小判断采集得到的文档W_i是否为与电网工程物资合同履约相关的网页，具体为：

[0064] $$r=ln^{\frac{\rho_{ig}}{\mu}}$$

[0065] $$\begin{cases} r>0, & \text{则为主题相关网页} \\ r<0, & \text{则为非主题相关网页} \end{cases}$$

[0066] 式中，ρ_{ig}为文档W_i和文档W_g之间的文档相似性系数，μ为主题阈值；

[0067] 当判断为主题相关网页时，即将网页送入数据库单元进行存储，并将主题相关网页中包含的链接加入主题爬虫的等待队列，当判断为非主题相关网页时即舍弃。

[0068] 本优选实施例提出一种改进的文档相似性系数的计算方法，引进了特征项的权重进行文档相似性系数的计算，解决了不同特征项对文档的影响程度不同而造成的相似度系数差异较大的问题，此外，通过计算得到采集得到的文档和样本文档之间的文档相似性系数，按照设定的主题阈值来判断当前的网页是否为主题相关网页，能够较为有效的进行主题相关网页的判别。

[0069] 基于上述实施例，根据采集得到的不同网页信息进行了一系列测试，以下是测试得到的评估结果：

[0070]

网页信息	判断系数r	网页分类情况	分类精度
关键词：电力、电力公司、物资公司、合同、履约、供应	$r>0$	主题相关网页	100%
关键词:供电、物资、履约、抽检、供应商、工程	$r>0$	主题相关网页	100%
关键词：电力物资、履约、合同、供应、预警、转变	$r>0$	主题相关网页	100%
关键词：电网建设、保障、电力、基础设施、改造	$r>0$	主题相关网页	92%

[0071] 从上述实施例可以观察到,网页筛选单元针对采集得到的不同网页信息进行网页分类具有较高精度,完全可以满足电网工程物资合同卖方履约诚信舆情监控系统的要求。

[0072] 最后应当说明的是,以上实施例仅用以说明本发明的技术方案,而非对本发明保护范围的限制,尽管参照较佳实施例对本发明作了详细地说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或者等同替换,而不脱离本发明技术方案的实质和范围。

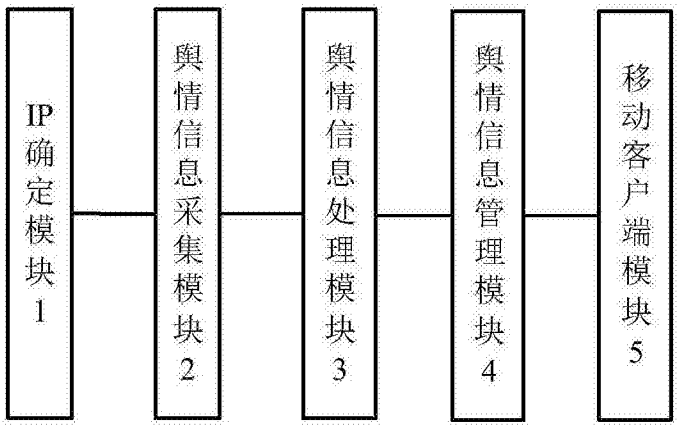

IP确定模块1
舆情信息采集模块2
舆情信息处理模块3
舆情信息管理模块4
移动客户端模块5

图1

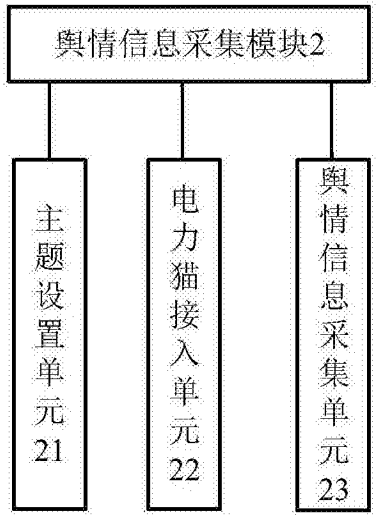

舆情信息采集模块2
主题设置单元21
电力猫接入单元22
舆情信息采集单元23

图2

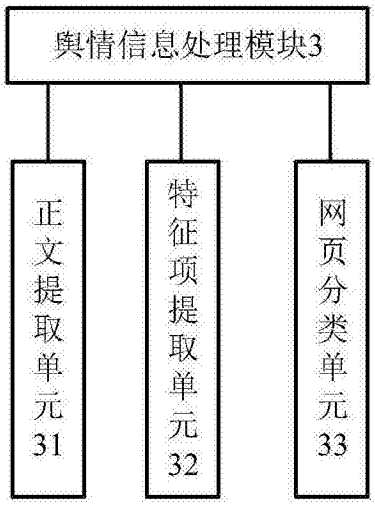

舆情信息处理模块3
正文提取单元31
特征项提取单元32
网页分类单元33

图3

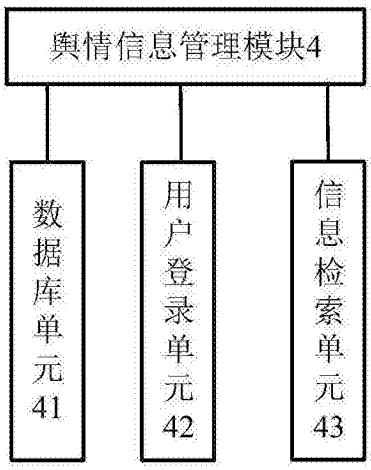


图4