基于突发事件热度的站点地图构建算法*

陈 翰, 韩永峰, 李弼程
(信息工程大学 信息工程学院, 郑州 450002)

摘 要: 为确保及时准确地获取监测网站的突发事件网络舆情数据,提出了一种基于突发事件热度的站点地图构建算法。该算法利用突发事件主题词典和改进 Shark search 算法采集样本网页,在此基础上对目标网站的超链接结构进行数据挖掘,完整构建出含有网站各版块突发事件热度的站点地图。以该站点地图为指导的网页采集器能够及时调整更新频率,准确采集所需网页,较好地适应监测网站的动态变化。实验表明,在站点地图的指导下,突发事件相关网页的采集有效性和效率均有明显提高。

关键词: 站点地图; 突发事件; 网页采集; 网络舆情; 数据挖掘
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)08-2943-05
doi:10.3969/j.issn.1001-3695.2012.08.037

Method of generalized sitemaps with emergency hot degree

CHEN Han, HAN Yong-feng, LI Bi-cheng
(*Institute of Information Engineering*, *Information Engineering University*, *Zhengzhou* 450002, *China*)

Abstract: In order to get the data about the emergency Web public opinion from the target sites in time, this paper proposed a method to generalize the sitemap based on the emergency hot degree. Utilizing emergency topic dictionary and advanced Shark search algorithm, the method collected enough sample Web pages to produce a sitemap containing the emergency hot degree of every board that was related to emergency in the target Web site. Under the guidance of the sitemaps generalized by this method, the Web clawer was intelligent enough to adapt well to the dynamical changes of the target sites, collected the needed Web page precisely and adjusted its update frequency when necessary. Experiments show that the Web clawer produces an outstanding performance both in the effectiveness and the efficiency with the help of the sitemap.

Key words: sitemap; emergency; Web clawer; online public opinion; data mine

0 引言

突发事件是指突然发生,造成或者可能造成严重社会危害、需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件[1]。突发事件一旦被网络媒体或者网民披露,短时间内就会引起大量关注,相关报道被重复转载、迅速传播,形成突发事件网络舆情。对“未然态”的舆情信息进行挖掘与分析,及时掌控突发事件网络舆情发展态势,避免突发事件事态扩大,有助于提高政府监管和处理网络突发事件的能力。

当前各舆情监测系统获取舆情数据主要通过网页采集技术。在突发事件网络舆情数据获取中,采用广度优先策略的通用采集器会存在以下三个问题:

a)采集器无法适应网站的动态变化。突发事件发生时,新闻网站大多会迅速设立专题版块予以报道,各大论坛也会开辟专门版块方便网民集中讨论。发现并添加新的专题超链接地址到采集系统的更新队列中,是及时获取突发事件网络舆情数据的必要条件,采用人工方法费时费力,无法满足时效性要求。

b)采集器效率低。通用网页采集器所抓取的网页中含有大量的无关内容,例如网站中含有的像“星座”“游戏”等休闲娱乐版块中,一般不会含有与突发事件相关的内容,因此采集时应避免采集这些版块的网页。

c)采集器更新频率不够灵活。常见网页采集器一般采用固定频率的采集方法,即每隔相同的时间对网站进行采集,但突发事件不同阶段的舆情数据更新频率差别很大,需要对采集器更新频率进行动态调整,才能做到在获取新事件最新动态的同时,保持对已有事件的持续关注。

目前,专门研究突发事件网络舆情数据获取的文献相对较少,但一些相关文献中提到的概念和方法可以借鉴到本文的研究当中。

1)基于主题的采集技术 该技术增加了对网页主题相关度的分析处理,仅需采集与用户预先定义好的某一主题相关的网页,大幅提高了系统资源和网络带宽的利用率。Bra 等人[2]提出了 Fish search 算法,它将在网络上遍历的网络网页采集器比喻成海里的一群鱼,当它们发现食物(相关信息)时,这些鱼就继续繁殖,寻找新的食物;当没有食物(没有相关信息)时,它们会就死掉。Shark search 算法[3]在 Fish search 算法的基础上进行了改进,能更好地保证采集中正确的搜索方向,提高相关信息的发现率。然而现有主题采集技术无法获取网站中各版块的更新频率,无法满足事件不同阶段和不同版块的区分性需求。

2)站点地图构建技术 该技术通过对网站中数量有限的

收稿日期: 2012-01-14; **修回日期**: 2012-02-29 **基金项目**: 国家社会科学基金重大项目(09&ZD014)

作者简介: 陈翰(1985-),男,硕士研究生,主要研究方向为文本信息处理(71482088@qq.com);韩永峰(1984-),男,硕士研究生,主要研究方向为文本信息处理;李弼程(1970-),男,教授,博导,主要研究方向为海量信息筛选、数据融合、数据挖掘.

样本网页进行分析,描绘出整个站点的层次结构,及时发现突发事件相关专题版块,这对于快速获取突发事件舆情数据尤为重要。现有方法中,Cai 等人[4]采用对样本网页进行多次聚类实现了站点地图的构建,Yang 等人[5]应用图的数据挖掘完成站点地图的构建,李魁等人[6,7]验证了以站点地图为指导的网页采集器在采集网页时的准确率和覆盖率上显著优于采用广度优先策略的通用网页采集器。这些方法虽然能够描绘出站点的层次架构信息,显著提高网页采集器进行网络舆情数据采集的能力,但无法区分出突发事件网络舆情数据较多的版块,无法满足突发事件舆情数据获取的有效性需求。

3)热点事件发现技术　突发事件一经报道或者曝光往往迅速成为热点事件,因此可以借鉴热点事件的研究理论和方法。刘星星[8]采用两层聚类和事件热度计算等策略发现热点事件。李恒训等人[9]对候选词集进行多重过滤,采用启发式规则的方法实现了基于主题词的网络热点话题发现算法。然而,热点事件的发现需要对一段时间内数量较多的网页进行聚类才能发现热点。据统计,只需 4 h 左右,突发事件就可能通过网络传播、发酵为有重大舆论影响的事件。因此,若直接应用热点事件发现方法检测突发事件,往往会错过处置突发事件的最佳时机,容易引发舆情危机。

因此,本文提出一种基于突发事件热度的站点地图构建算法,在该算法生成的站点地图的指导下,网页采集器在采集突发事件相关网页的有效性和效率上均有明显提高,确保了舆情监测系统快速准确地获取突发事件网络舆情数据。

1　基于突发事件热度的站点地图构建算法

本文提出的基于突发事件热度的站点地图构建算法,以监测网站的首页地址为输入,以基于突发事件热度的站点地图为输出。地图中包含该网站与突发事件相关的版块首页地址、版块间的结构关系、各版块的突发事件热度信息和相应的更新频率。算法基本流程图 1 所示。

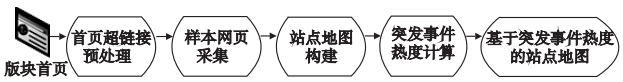

图1　突发事件热度站点地图构建算法流程

1.1　首页超链接预处理

网站首页是网站的入口网页,通常被设计成一个站点的目录页面,包含着数以千计的超链接,主要有以下几类:a)操作型超链接,一般位于页面的最顶部,用于完成登录或进行站内搜索等;b)导航型超链接,一般位于页面的次顶端或次底端,用于指向 1 级版块首页和重要的 2 级版块首页;c)内容型超链接,它是网站首页所含超链接的主体,位于页面的中心主体区域,用于指向网站各级版块中最新或最热的网页;d)广告型超链接,一般位于页面的两侧,以及页面不同区域的交接处,用于指向广告页面;e)文书型超链接,一般位于页面的最底端,用于指向网站的版权、许可证等各种法律文书等页面。本文将首页所含超链接指向的网页称为二层网页,将二层网页所含超链接所指向的网页称为三层网页。表 1 展示的是新浪网、搜狐网和凤凰网三个门户网站前三层网页的数量。由表 1 可知,网页的层数每增加一层,其所包含的超链接数便以指数级进行增加。然而并不是所有的网页对于构建突发事件热度地图都有帮助,因此必须对首页所含超链接进行过滤,以减少系统资源和网络带宽的浪费。

表1　首页超链接过滤前后不同层数网页数量

网站名称	层数					
	过滤前			过滤后		
	1	2	3	1	2	3
新浪网	1	1 461	674 532	1	80	45 489
搜狐网	1	1 109	514 785	1	74	34 574
凤凰网	1	906	313 210	1	67	29 217

定义　网页 URL 节点长度。以“.”或“/”号对网页 URL 进行分割,分割后得到的块数称为该 URL 的节点长度。例如“news. sina. com. cn”的节点长度为 4,“club. ent. sina. com. cn”的节点长度为 5。本文将“新闻”“体育”“论坛”等 URL 节点长度与网站首页 URL 节点长度相同的版块称为 1 级版块,将“国内新闻”“中超”等 URL 节点长度为网站首页 URL 节点长度 +1 的版块称为 2 级版块。一般而言,网站首页包含若干个 1 级版块,而每一个 1 级版块又有若干个 2 级版块。

通过长期观察,本文得出以下结论:

a)突发事件由于其自身具有的危害性、紧迫性和高关注性等特点,与其相关的网页常常毫无争议地占据着网站首页、各级版块首页等显著位置。

b)若网页 w 属于版块 A_1,而版块 A 是版块 A_1 的父版块,则如果网页 w 出现在版块 A 的首页上,则网页 w 必然出现在版块 A_1 的首页上。

c)在父版块首页中,总是包含指向其子版块的导航超链接。

根据 a),若某 1 级版块中不含有与突发事件相关的网页,则其下属的所有 2 级版块也不含有与突发事件相关的网页;根据 b),若某 1 级版块首页中含有与突发事件相关的网页,则该网页必定出现在其下属的某个 2 级版块的首页中;根据 c),从网站任何一个版块入口,都可以一个不漏地访问到该版块的每一个子版块。

本文在预处理时仅保留网站首页所含导航型超链接中的 1 级版块超链接,提出了首页超链接过滤算法(index page hyperlink filter,IPHF)。其主要思想如下:

a)由于导航型超链接一般位于网站首页的次顶端和次底端,因此超链接队列中前 15% 和后 15% 的超链接都需要被滤除。

b)根据观察,1 级版块名称一般不超过 4 个字,因此超链接锚文本长度在 4 以上(不含 4)的超链接需要被滤除。

c)依据 1 级版块首页的 URL 规律,URL 中不以“HTTP”开头,或含有“%”“&”“#”等传参符号,或 URL 节点长度大于网站首页节点长度的超链接都需要被滤除。

d)1 级版块首页一般为本版块所含主要内容的超链接目录,极少含有标点符号,本算法在滤除了〈Script〉脚本标签和〈! --……--〉注释标签后网页的〈body〉源码内进行检索,若超链接所指向网页源码中含有 5 个以上“。”“?”“!”等正文常见标点符号,则将该超链接滤除。

e)URL 重复的超链接需要被滤除(URL 对应锚文本不同的以首个 URL 对应锚文本作为该 URL 锚文本)。

本文将 IPHF 算法用于新浪网、搜狐网和凤凰网,经算法过滤后不同层数网页数量情况如表 1 所示。可以看出,过滤后网上第二层网页数量大幅减少,在保证召回率 100% 的前提下,精确率在 97.7% 以上,有效减少了系统资源和网络带宽的浪费。

1.2　样本网页采集

在构建站点地图之前,系统需要采集一定数量的样本网页,文献[5]将采集样本网页的网页采集器采集深度设定为5。从表1的实验结果可以看出,随着网站层数的增加,其网页数量呈指数级增长,虽然经过了超链接过滤,第三层网页的数量仍然数以万计,随着监测网站数量的增加,有限的网络带宽和系统资源将面临巨大挑战。由于本文研究的是突发事件网络舆情,因此,本文在采集样本网页时,采用了主题网页爬虫相关技术,使网页采集器只采集与突发事件相关的网页。

主题词具有代表性、简洁性、时效性等特点,能够最大程度地以最小的信息量代表突发事件的主题和内涵。因此,本文将建立突发事件主题词典,将版块内网页内容与突发事件主题词词典的相关度作为版块内容的突发事件相关度。选取主题词一般有两种方法:特征提取和人工挑选。特征提取是指给定一个关于突发事件的网页集合,由程序抽取出网页的标题和正文内容后,进行中文分词并自动提取这些网页里的主题词并赋予权值。特征提取的优点是权值量化定义精确,但对中文分词质量依赖程度高,而且会出现不少停用词和无意义词串。人工挑选的好处是实现简单,同时人的经验一般比较准确,不会与实际情况出现较大偏差;缺点是人工挑选主题词的工作量较大,而且可能出现遗漏,对主题词权值的量化不够客观。因此,本文采取了特征提取粗选和人工挑选细选两个步骤,保证主题词典的全面性和客观性。

本文选取了新浪网、凤凰网、中国新闻网、网易网、腾讯网这五家门户网站近十年来(2001 年—2011 年)关于突发事件的专题报道版块中包含的网页集,作为突发事件主题词典的训练语料。其中自然灾害类 93 个、事故灾难类 162 个、公共卫生事件 49 个、社会安全事件 87 个。选择的部分案例如表 2 所示。

表2　用于构造突发事件主题词典的部分突发事件案例列表

自然灾害	事故灾难	公共卫生事件	社会安全事件
泰国洪水肆虐	甘肃幼儿园校车被撞	思念水饺检出高致病球菌	山东泰安持枪袭警案
土耳其 7.2 级地震	西安建筑物爆炸	甲型 H1N1 流感	英国伦敦骚乱
甘肃舟曲突发特大泥石流	云南曲靖师宗县煤矿事故	打击“地沟油”专项行动	香港游客菲律宾遭劫持
中国多地遭暴雨洪涝灾害	“7.23”温州动车追尾事故	问题奶粉重现市场	菲律宾巡逻舰撞击我渔船
四川道孚草原火灾	上海地铁列车相撞事故	双汇瘦肉精事件	尼日利亚联合国大楼爆炸
⋮	⋮	⋮	⋮

在构造突发事件主题词典时,首先由 ICTCLAS 中文分词系统对训练语料进行中文分词和词性标注,将人名、地名、组织机构名等与突发事件无关的命名实体滤除,其次计算主题词权重:

$$W_t = 1 + \log_2(tf(t)) \times \text{weight}(\text{POS}(t)) \times \text{weight}(\text{position}(t)) \quad (1)$$

其中:W_t 表示词 t 的权重;$tf(t)$ 表示词 t 在训练语料集中的频率;weight(POS(t))为词 t 的词性权重,动词和名词取 2,其他词性取 1;weight(position(t))为词 t 的位置权重函数,若词 t 出现在标题中,函数值为 3,若词 t 出现在正文首句话中,函数值为 2,其他情况下,函数值为 1。

权重归一化公式如式(2)所示。

$$W_t = \frac{W_t}{\max_T(W_t)} \quad (2)$$

以人工方法剔除了与突发事件无关的无意义词后,以主题词权重为依据排序,由高到低选取了 350 个主题词,构造出突发事件主题词典。

在突发事件主题词典构造完毕后,本文对主题网页采集器中经典的 Shark search 算法进行改进,并用于样本网页的采集。具体算法如图 2 所示,主要改进如下:

a)取消了 Shark search 算法中计算 ChildNode 的 AnchorContextScorre 项,因为现有网页对大部分超链接都采用集中放置的布局策略,若仍然考虑超链接周围的锚文本,则必然影响超链接所对应网页主题相关度的精确计算。

b)增加了选取网页 ChildNode 数量的继承超链接宽度 Width,其理由根据如下假设:若某网页突发事件相关度高,则该页面中包含指向突发事件相关网页的超链接数量也较多。

在改进的 Shark search 算法中,网页的突发事件相关度由式(3)计算得出:

$$\text{Cor}(\text{CurrentNode}, D) = \sum_{T_{\text{Web}}} W_t \times \text{weight}(\text{position}(t)) \quad (3)$$

其中:T_{Web} 为网页中突发事件主题词的集合;W_t 为词 t 在突发事件词典中的权重;weight(position(t))为主题词 t 的位置权重函数,若词 t 出现在标题中,函数值为 3,若词 t 出现在正文首句话中,函数值为 2,其他情况下,函数值为 1。

算法中定义的其他参数如下:继承超链接宽度 Width,继承超链接深度 Depth,相关度衰减因子 δ,锚文本权重系数 β。为了得到使算法效果最优的一组参数值,本文进行了若干组实验,得到一组最优化参数:Width = 80,β = 0.8,δ = 0.5,Depth = 3。

1.3　站点地图构建

构建站点地图需要对网站的超链接结构进行数据挖掘。根据功能的不同超链接可分为两种类型:结构性超链接和功能性超链接。结构性超链接从版块页面出发,指向该版块的子版块,用户可以依据自身兴趣随着这些超链接深入网站的具体版块。网站中,除了结构性超链接外,还存在大量超链接,用于在两个网页间建立快速通道,此类超链接称为功能性超链接。如图 2 所示,一个教授的首页可以链接到其所在院系首页和其所授的课程首页上。此外,在规模稍大的网站中,相同版块下的网页常采用模板生成,这些网页中通常含有相同的导航条,包含大量指向站内重要网页的超链接。这些功能性超链接为用户提供了便利,却给构建站点地图带来了麻烦。图 2 左侧显示的网站既包含了结构性超链接,也包含了功能性超链接;右侧显示的网站只含有结构性超链接。图中左侧所示的功能性超链接有 10 个,而结构性超链接只有 5 个,实际网站中,结构性超链接几乎淹没在功能性超链接中。

以下为改进的 Shark search 算法。

输入:种子 URL 队列 Q,继承超链接宽度 Width,继承超链接深度 Depth,主题词典 D,相关度衰减因子 δ。

输出:样本网页集 P。

while Q 不为空:

a)取出 Q 的第一个节点 Node,令其为 CurrentNode,令 ChildNode 为 CurrentNode 的孩子节点,采集 CurrentNode 对应的网页并存入样本网页集 P;

b)若 CurrentNode 的继承超链接深度 Depth≥0:

c)计算 CurrentNode 和 D 的相似度 Cor(CurrentNode,D)

若 Cor(CurrentNode,D)≥0

令 InheritedScore(ChildNode) = δ * Cor(CurrentNode,D);

否则

令 InheritedScore(ChildNode) = δ * InheritedScore(CurrentNode);

令 AnchorText 为指向 ChildNode 的超链接锚文本;

d)计算 AnchorText 和 D 的相似度 AnchorScore = Cor(AnchorText,D);

e)计算孩子节点的 PotentialScore:

PotentialScore(ChildNode) = β * AnchorScore(ChildNode) + (1 - β) * InheritedScore(ChildNode)β 为算法参数

f)对 CurrentNode 的每一个 ChildNode:

计算 ChildNode 的继承超链接深度 Depth:

若 Cor(CurrentNode,D) ≥0

令 Depth(ChildNode) = Depth

否则 Depth(ChildNode) = Depth(CurrentNode) - 1

若 ChildNode 已存在于队列中,则比较已有节点的 Depth 和新增节点的 Depth,取值较大者;

g)将 CurrentNode 孩子节点的前 Cor(CurrentNode,D) * Width 个加入 Q;

end while

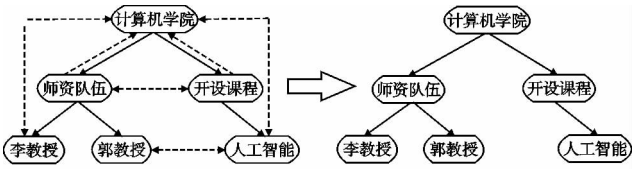

图2 超链接机构与站点结构图

本文将网站看成一个有向图 $G(V,E,r)$,其中:V 代表网站中的所有网页的集合;E 代表网站中所有从源网页指向目标网页的所有超链接集合;r 代表网站首页。本文需要构建的网站站点地图是对有向图 $G(V,E,r)$ 进行数据挖掘,得到 $G(V,E,r)$ 的一个子集 $T(N,V_s,r)$,其中:T 是一个有向树;r 为 T 的根,是网站的首页;N 为 V 的子集,是从 r 出发经过若干超链接能够到达的所有版块首页的集合;V_s 为 E 的子集,是网站中所有结构性超链接的集合。本文采用文献[5]中提到的站点结构树生成算法,该算法首先使用决策树算法进行机器学习,训练出识别站点中结构性超链接和功能性超链接的分类器;然后将统计学方法与训练好的分类器相结合,预测出新的超链接为功能性超链接的概率,将此概率作为超链接的权值;最后使用有向图最小生成树算法生成站点结构树。该算法生成的站点结构树的准确率平均值为91.9%。

1.4 突发事件热度计算

本文根据网络舆情载体的不同特点将其分为新闻、论坛和博客三种类别,分别计算其突发事件热度。

在新闻版块中,单位时间内发布的与突发事件相关的网页数能够反映该版块的热度,因此定义新闻版块突发事件热度(news board emergency hot degree)如式(4)所示:

$$\mathrm{BEHD}_{\mathrm{news}} = (1+\omega)\sum_{A}\left(\sum_{T} W_t \times \mathrm{weight}(\mathrm{position}(t))\right) \tag{4}$$

其中:W_t 是主题词 t 在突发事件主题词典中的权重;A 为版块中所有文章的集合;ω 为网页采集器更新时刻 1 h 内版块中突发事件相关文章的数量;weight(position(t))为主题词 t 的位置权重函数,若词 t 出现在标题中,函数值为 3,若词 t 出现在正文首句话中,函数值为 2,其他情况下,函数值为 1。

论坛的浏览数和回复数最能够反映帖子的热度,某一版块中所有与突发事件相关的帖子热度的累加就可以代表该版块的突发事件热度,因此,定义论坛版块突发事件热度(BBS board emergency hot degree)如式(5)所示:

$$\mathrm{BEHD}_{\mathrm{BBS}} = \sum_{P}\left(\sum_{T} W_t\right) \times \mathrm{weight}(p) \tag{5}$$

其中:W_t 是主题词 t 在突发事件主题词典中的权重,weight(p)为帖子 p 的权重系数,由式(6)计算得出:

$$\mathrm{weight}(p) = \lambda \times \mathrm{reply\_count} + (1-\lambda) \times \mathrm{browse\_count} \tag{6}$$

其中:reply_count 为帖子 p 的回复数;browse_count 为帖子 p 的浏览数;λ 为回复数权重系数,一般取 0.8。

博客与论坛比较相似,本文采取与论坛相同的方法进行处理。

在网页采集器更新频率方面,本文设置了 8 级更新时间,分别是 1 min、2 min、5 min、10 min、15 min、20 min、30 min、60 min、2 h、4 h。更新频率变化规则为:若某版块的突发事件热度在更新时较前一次增长了 γ,则更新频率加快 1 级;若某版块的突发事件热度在更新时较前一次降低了 γ,则更新频率减慢 1 级;其他情况下,更新频率不变。经过大量实验,当 γ = 50% 时,采集网页采集器的更新频率与版块更新频率基本相同,采集效果最好。

有效时间的设置是为了使站点地图能够适应监测网站的动态变化,及时发现监测网站的新增版块,当站点地图生成时间超过有效时间后,站点地图需要重新生成,一般将有效时间设置为 4 h。

2 实验结果与比较分析

基于突发事件热度的站点地图主要用于指导网页采集器快速准确地获取突发事件网络舆情数据,本文以主题命中率 R_h 来衡量网页采集器的采集有效性,即 $R_h = \frac{N_h}{N}$。其中:N_h 为采集的网页中命中主题的数量,N 为采集的网页总数。以采集相同数量主题网页时,网页采集器实际采集的网页数量 N_p 来衡量网页采集器的采集效率。实验中,将基于突发事件热度站点地图采集策略的网页采集器分别与基于广度优先算法采集策略和基于 PageRank 算法采集策略的网页采集器进行对比,实验结果如表 3 和 4 所示 。

表 3 不同算法指导下网页采集器的有效性对比

采集网页数	主题命中率		
	基于广度优先算法/%	基于 Page Rank 算法/%	基于突发事件热度站点地图算法/%
1 000	54.00	61.67	88.77
2 000	49.50	35.17	67.68
3 000	36.67	25.89	71.54
4 000	29.33	23.67	65.87
5 000	24.60	28.53	64.98
6 000	21.61	31.39	69.78
7 000	21.23	34.62	71.41
8 000	19.50	33.42	68.94
9 000	18.70	32.19	72.43
10 000	18.07	33.23	74.15

从表 3 可以看出,在采集网页的初始阶段,不同采集策略指导下的网页采集器均可保持较高的主题命中率;随着采集的推进,广度优先和 PageRank 算法指导下的网页采集器的主题命中率开始下降;而本文算法生成的站点地图能够较好地防止主题漂移的发生,始终保持较高的主题命中率,达到了预期的效果。在网页采集数量足够多的情况下,站点地图指导下的网页采集器采集有效性分别是广度优先和 PageRank 算法指导下

网页采集器采集有效性的 4.1 倍和 2.2 倍。

表 4　不同算法指导下网页采集器采集的采集效率对比

主题相关网页数	实际采集网页数量		
	基于广度优先算法	基于 Page Rank 算法	基于突发事件热度站点地图算法
1 000	1 851	1 621	1 126
2 000	3 872	4 464	2 604
3 000	6 599	8 327	4 001
4 000	10 008	12 552	5 520
5 000	14 073	16 057	7 058
6 000	18 701	19 242	8 492
7 000	23 411	22 131	9 892
8 000	28 539	25 123	11 342
9 000	33 887	28 230	12 723
10 000	39 421	31 239	14 072

从表 4 可以看出,同样是因为主题漂移的发生,广度优先和 PageRank 算法指导下的网页采集器,在采集相同数量的主题相关网页时,实际采集的网页数量逐渐增多,且随着采集地深入,累加效应越发明显,在网页采集数量足够多的情况下,站点地图指导下的网页采集器采集效率分别是广度优先和 PageRank 算法指导下网页采集器效率的 2.8 倍和 2.2 倍。

在时间复杂度方面,基于广度优先算法的时间复杂度为 $O(|E|)$,而 PageRank 算法的复杂度在最坏的情况下为 $O(|V|\times|E|)$,一般而言,其时间复杂度为 $O(100\times|E|)$,其中数字 100 指 PageRank 算法一般需要迭代 100 次才能达到可接受的 PageRank 值,而本文提出的基于突发事件热度站点地图算法所需时间复杂度为 $O(|V|^3)$。可以看出,为了在克服主题漂移的基础上获得较高的采集有效性和采集效率,本文所提出的方法在时间复杂度上远高于其他算法。然而,网站版块的变化并不是经常发生,因此方法的时间复杂度仍在可以忍受的范围内。

3　结束语

本文提出了一种基于突发事件热度的站点地图构建算法,该算法利用突发事件主题词典和改进 Shark search 算法采集样本网页,在此基础上对目标网站的超链接结构进行数据挖掘,完整构建出含有网站各版块突发事件热度的站点地图,以该站点地图为指导的网页采集器能够及时调整更新频率,准确采集所需网页,较好地适应监测网站的动态变化。实验表明,在站点地图的指导下,网页采集器在采集突发事件相关网页的有效性和效率上,均优于广度优先算法和 PageRank 算法指导下的网页采集器。下一步工作的重点是突发事件主题词典的改进,如将主题词典训练语料从新闻网页集扩展到论坛和博客网页集,并在此基础上从网站和版块的突发事件热度入手,分析突发事件网络舆情的演化规律。

参考文献:

[1] 中华人民共和国主席令第六十九号中华人民共和国突发事件应对法[EB/OL]. (2007-08-30)[2011-12-03]. http://www.gov.cn/ziliao/flfg/2007-08/30/content_732593.htm.

[2] BRA P D, HOUBEN G J, KORNATZKY Y, *et al*. Information retrieval in distributed hypertexts[C]//Proc of the 4th RIAO Conference. 1994:481-491.

[3] MICHAEL H, MICHAL J, YOELLE S M, *et al*. The shark-search algorithm-an application: tailored Web site mapping[J]. Computer Networks and ISDN Systems, 1998, 30(7):317-326.

[4] CAI Rui, YANG Jiang-ming, LAN Lai, *et al*. iRobot: an intelligent crawler for Web forums[C]//Proc of the 17th International World Wide Web Conference. 2008:447-456.

[5] YANG C C, LIU Nan. Web site topic-hierarchy generation based on link structure[J]. Journal of the American Society for Information Science and Technology, 2009, 60(3):495-508.

[6] 李魁,程学旗,郭岩,等. WWW 论坛中的动态网页采集[J]. 计算机工程,2007,33(6):80-82.

[7] 蔡欣宝,郭若飞,赵朋朋,等. Web 论坛数据源增量爬虫的研究[J]. 计算机工程,2010,36(9):285-287.

[8] 刘星星. 热点事件发现及事件内容特征自动抽取研究[D]. 武汉:华中科技大学,2009.

[9] 李恒训,张华平,秦鹏. 基于主题词的网络热点话题发现[C]//第五届全国信息检索学术会议. 2009:134-143.