基于数据流的网页内容分析技术研究

王佰玲[1,2],曲 芸[1],张永铮[3],田志宏[1]
(1.哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001;2.北京大学信息科学与技术学院,北京 100871;
3.中国科学院计算技术研究所,北京 100190)

摘 要: 提出针对网络数据流中活跃信息进行话题相关数据采集与分析方法.首先给出面向论坛话题的定义;然后对网络数据流进行分析、对用户访问行为进行分类;并给出基于数据流的用户行为识别方法及话题相关数据抽取、存储算法;最后给出实验分析,结果表明,所提出的基于数据流的论坛话题数据采集方法能够很好地反映用户行为,并对基于数据流的网络舆情热点话题发现、突发事件检测与实时跟踪等应用提供有利的数据资源.

关键词: 网络舆情; 热点话题; 突发事件; 网络数据流
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2013)04-0751-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.04.021

Research on Network-Traffic Based Web Traffic Computing Technology

WANG Bai-ling[1,2], QU Yun[1], ZHANG Yong-zheng[3], TIAN Zhi-hong[1]
(1. *School of Electronics Engineering and Computer Science*, *Peking University*, *Beijing* 100871, *China*;
2. *School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*;
3. *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: In this paper, a network-traffic based topic extracting and analyzing method is introduced. The new topic definition for web2.0 and the classification of user behavior is given; the detecting method of user behavior, topic extracting method, and data storage algorithm is also proposed. At last, a prototype of topic collector based on network traffic is implemented; the testing results show that the user behavior and the hot topic can be collected and detected effectively and correctly, and the new method provides a new data channel for analyzing public opinion.

Key words: public opinion; hot topic; emergent event; network traffic

随着 Web 2.0 等互动性网络媒体的大量涌现,大量网络论坛、博客、播客、维基百科等低门槛应用被开发出来,打破了传统单向式媒体在信息上的垄断.这些应用不仅方便了人们在网络上以更便捷和更丰富的方式表达观点,更重要的是它们改变了网络的秩序.由此网络舆情研究已经成为舆情研究的重要领域,网络舆情的爆发也以"内容威胁"的形式逐渐对社会公共安全形成威胁,其研究方法还不尽完善.特别是面对一些突发事件和公共事件,舆情管理部门越来越难以全面掌握话语权,这就对网络舆情监管提出了新的更高要求.

网络舆情信息作为社会万象的映射,最直接、最快速地反映了社会舆情的状况和发展态势.对网络舆情进行快速获取、有效分析、持续跟踪、及时预警、有效调控,有助于对形成群体问题的因素及时进行判断与处置,有助于对社会舆情进行有效的掌控.

本文主要针对网络数据流中话题相关数据的识别、提取与存储方面进行研究.基于网络数据流的信息挖掘主要是针对网络中真实的、活跃的数据流量信息进行分析,从而及时发现网络中的热点主题,主要优势在于:(1)网络数据流更真实地反映了用户所关心数据的热度,不会受到网站作弊等因素的影响;(2)基于网络数据流分析,可以提高数据的获取速度,及时反映出网络舆情真实情况;(3)网络数据流不仅仅在时间上反映了用户关心数据的变化,也在空间上反映了事件的扩散范围、扩散方向及用户的行为.

1 论坛话题研究基础

定义 1 观点(View) 观点是论坛中作者对某件事情发表的评论,其存在形式可以是论坛某个板块中关于某件事情的主帖,也可以是针对某个帖子的回复.

收稿日期:2011-09-10;修回日期:2012-09-20
基金项目:国家自然科学基金(No.61170262)

定义 2 事件(Event) 针对一件事情,在论坛中往往有一系列的帖子进行评论.我们把针对同一件事情的所有观点(包括主帖和回帖)定义为事件.

定义 3 事件树(Event Tree) 在论坛中,访问者不仅会对主帖进行评论,而且还会对回帖进行评论,有的时候回帖表达的观点影响力也会大于主帖.事件树用来刻画主帖和回帖所表达观点的关联关系,事件树的树根是那些影响力比较大的观点(可以是主帖,也可以是回帖).如果事件树 A 的节点 b 所对应的观点影响力逐渐增大,甚至超过事件树 A 的根节点 a,在表述上则可以把 b 节点称为新的根节点,根节点 b 及其下所有关联事件构成事件树 B,事件树 B 称为事件树 A 的子树.

定义 4 观点块(View Block) 每个作者发帖子表述观点的时候,帖子内容在页面上的显示是独立的,并且是有明显界限划分的区域,我们把这个区域称为观点块.这种分割算法基于 DOM(Document Object Model)[1].DOM 提供了树形结构的页面模型,本文即采用基于 DOM 来建立 Block 树.在论坛上浏览帖子时,Web 中提供了大量可见的元素来划分页面,呈现给访问者的是一系列相关的并且是有限个的观点块集合,如图 1 所示.

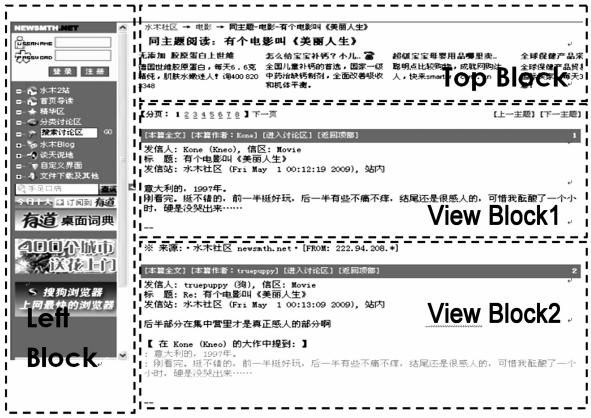


图1 用户浏览时页面上观点块分布

2 基于数据流话题数据采集原理

论坛帖子状态转换分析呈现了帖子在生命周期中不同阶段的展现形式.每种状态转换都是用户通过网络进行操作完成的,所以从数据流传输的角度可以对论坛帖子进行更深入的实时分析.相关数据获取技术及部署环境参照图 2.其中关于数据获取性能影响因素问题,本作者已在文献[2]中作了详尽分析,并在文献[3]中给出了最优的处理方法,采用该技术可以实现线性速度处理,可以满足本文的技术基础要求;关于系统的可扩展性问题,本作者在文献[4]的 5.5 节中给出了 MPMD(多程序多数据流)的解决方案,可以解决本文研究工作应用时的扩展问题;本作者在文献[5]中给出了可并行的数据分析方法,该方法可以对本文分析结果的进行快速分类.其他学者亦有相关研究,如文献[6]针对不平衡的网络数据流建立了集成分类模型、文献[7]给出了一种基于缓冲区紧耦合的数据流抽取算法,这些研究均奠定了本文研究的技术基础.

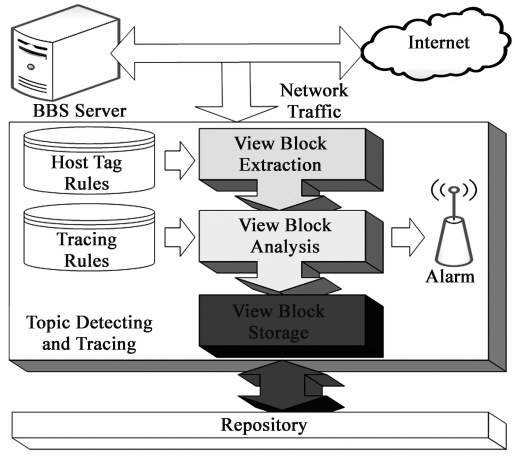


图2 数据流中话题分析系统部署环境及原理

2.1 网络数据流信息分析

根据 HTTP 规范,GET 命令用于信息获取,而且过程是安全的和幂等的;而 POST 命令用于用户向站点传送数据.从 HTTP 请求与响应交互过程中,可以根据协议固有信息对服务交互方法进行识别;进而解析出可以用来识别用户行为的有效信息,如图 3 所示.

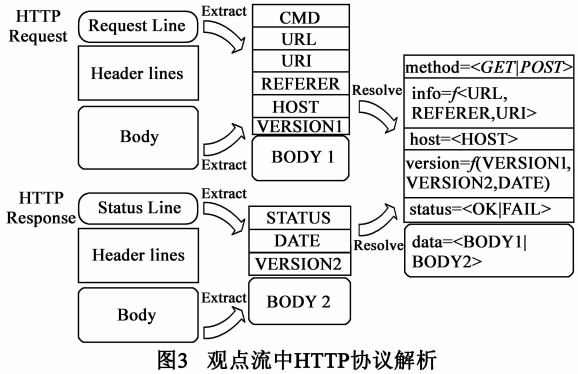


图3 观点流中HTTP协议解析

通过对 HTTP 协议的简要分析,基于数据流的话题检测技术所涉及到的必要内容特征范式可以表述为:

$$\text{HTTP Stream} ::= <\text{Request}, \text{Response}> \quad (1)$$

$$\text{Request}:: = <\text{method}, \text{host}, \text{URL}, \text{URI}, \text{REFERER}, \text{VERSION1}, \text{BODY1}> \quad (2)$$

$$\text{Response}:: = <\text{status}, \text{VERSION2}, \text{DATE}, \text{BODY2}> \quad (3)$$

这里"::="表示从左侧数据项中可以提取出所有右侧表达式所包含的数据属性值.其中 method 唯一标示了该数据流是在向服务器提交数据,还是从服务器获取数据;host 用来表示论坛主机地址;URL/URI/REFERER 均取自 HTTP 请求头中所包含的相应字段,具体可参照

HTTP 协议规范;用 status 表示论坛回应;VERSION1 和 VERSION2 分别表示协议的版本及服务器端数据版本;BODY1 和 BODY2 分别表示数据流向服务器传送的数据内容以及从服务器端接收到的数据内容;DATE 表示数据最后更新时间.

2.2　用户访问行为分类

论坛板块中事件的演化过程是受用户行为驱动的.论坛中一个事件和用户相关的主要因素包括:1)用户发帖行为(Post a New Thread Behavior,PNTB);2)用户浏览行为(Get a Thread Behavior,GTB);3)用户修改行为(Edit a Thread Behavior,ETB);4)用户回复行为(Reply a Thread Behavior,RTB).除此之外,还有一些行为,比如作者删帖(Delete a Thread Behavior,DTB)、论坛管理员将帖子置为只读、隐藏、删除等.本文所研究的话题检测与跟踪将这些看作论坛正常管理行为,所以不会检测这些行为.以上访问行为对事件演化过程影响如图 4 所示:

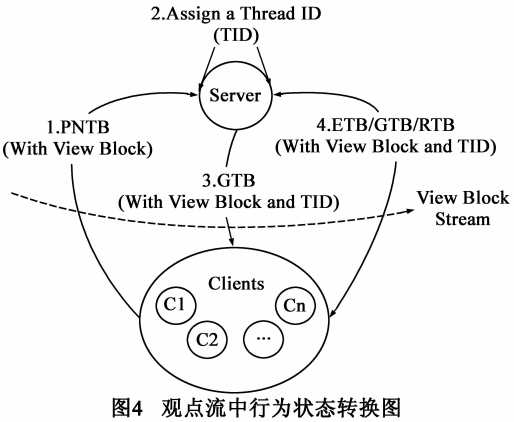


图4　观点流中行为状态转换图

基于该状态图,针对论坛话题的用户行为可以表示为:

$$\text{User Behavior}:=<\text{PNTB},\text{GTB},\text{ETB},\text{RTB},\text{DTB}> \quad (4)$$

式(4)中每个用户行为所包含的内容特征范式分别可以表述为:

$$\text{PNTB}:=<\text{method},\text{action},\text{host},\text{data},\text{version}> \quad (5)$$

$$\text{GTB}:=<\text{method},\text{action},\text{host},\text{tid},\text{data},\text{version}> \quad (6)$$

$$\text{ETB}:=<\text{method},\text{action},\text{host},\text{tid},\text{data},\text{version}> \quad (7)$$

$$\text{RTB}:=<\text{method},\text{action},\text{host},\text{fid},\text{data},\text{version}> \quad (8)$$

以上公式中“:=”左侧表示用户具体行为,右侧表示用户行为的组成元素.其中,method 属性可以取自式(2)的输出,一般情况下 method:和 = <“POST”,“GET”>;host 属性可直接取自式(2)的输出,用来表示用户行为所对应的论坛地址.

上述范式右侧除了可以从式(2)~(3)中直接获得的数据项外,其他数据项需要进行一些中间计算才能获得.为了表述方便,直接给出中间变量 info 的范式,以便于后续论述可以直接引用该数据项提供的元素集合:

$$(\text{info}=f(\text{URL},\text{URI},\text{REFERER}))::=<\text{action},\text{tid}|\text{fid}> \quad (9)$$

$$\text{action}:=<\text{“newthread”},\text{“replythread”},\text{“editthread”},\text{“browsethread”}> \quad (10)$$

$$\text{version}:=f(\text{VERSION1},\text{VERSION2},\text{DATE}) \quad (11)$$

这里“=”表示左侧数据项 info 由右侧变量 URL,URI,REFERER 进行计算所得,最终输出项可以用“::=”右侧的所有元素来表示.tid 为话题索引 ID,fid 为用户回帖 ID,action 的每个元素由不同的 tag 来标示,包括 PNTB_Tag、GTB_Tag、ETB_Tag 及 RTB_Tag 来表示 action 的不同元素.

3　话题数据抽取与存储算法

论坛话题在网络数据流传输过程中,只有用户 PNTB、RTB、ETB 三种行为会导致观点块内容变化或者增加新的观点块,而数量最多(远超过 PNTB、RTB、ETB 三种行为之和)的 GTB 行为不会导致观点块内容变化.PNTB、RTB 与 ETB 三种行为不同的是,PNTB、RTB 导致新的观点增量,而 ETB 导致已有观点数据变化.PNTB 与 RTB 行为区别是,RTB 生成的观点块在数据流传输过程中,存在 TID 与其他事件显性相关联;而 PNTB 则无关联.这些不同点将有助于研究基于数据流的观点块提取与分析.

3.1　行为识别与观点块抽取算法

观点块抽取的过程就是从网络数据流中识别用户行为,再根据用户行为的不同采用不同的模式提取观点块的过程.主要分为以下几种情况:

(1)PNTB 行为中观点块抽取　对于某一个事件的生命周期而言,PNTB 行为发生且只发生一次,并且这个行为生成的观点块将作为本事件的主节点,后续相关的观点块将以此节点为根进行增量存储.

(2)GTB 行为中观点块抽取　对于某一个帖子 GTB 行为可以发生多次,表示有很多访问者关心这个帖子;也可能不发生,即没有任何人关心过这个帖子.对于每次 GTB 行为,其承载的观点块都有一个唯一的 TID 与其相对应,表示该用户浏览的是编号为 TID 的帖子.此时抽取出来的观点块不是孤立的,需要在存储过程中建立关联关系,以便于做话题检测与跟踪.

(3)RTB 行为中观点块抽取　对于某一个帖子 RTB 行为可以发生多次,表示有很多访问者关心这个帖子并加以评论;也可能不发生,即没有任何人对这个帖子发表评论.对于每次 RTB 行为,其承载的观点块都有一个唯一的 TID 与其相对应,表示是编号为 TID 帖子的回帖.所以,此时抽取出来的观点块不是孤立的,其观点块内容虽然没有在网络流中出现过的,但是其对应事

件的父节点已经在网络流中出现过的.

(4)ETB 行为中观点块抽取 对于某一个帖子 ETB 行为可以发生多次,表示用户多次修改该帖子内容;也可能不发生,即表示该用户从未进行观点的修正与更新.对于每次 ETB 行为,其承载的观点块都有一个唯一的 TID 与其相对应,表示是对编号为 TID 的帖子进行更新.所以,此时抽取出来的观点块不是孤立的,其更新后的观点块内容虽然没有在网络流中出现过的,但是其修正节点已经在网络流中出现过的.算法 1 描述了针对 PNTB 行为观点块抽取算法,其它行为算法类似.

算法 1 观点块抽取算法 Extrace_VB_From_HTTP().

Input: 1)HTTP Stream 2) Host Tag Rules = < Host, PNTB_Tag, RTB_Tag, ETB_Tag, GTB_Tag >

Output: ViewBlock = < TID, version, content >

```
1.  http = HTTP_INIT(HTTP Stream)/* According to formula (1) ~ (3) to initialize http object */
2.  tmp = TEMPLATE_SET(http.host)/* Select the host template according to the input configures */
3.  If http.method = "POST"
4.    switch http.action
5.    case tmp.PNTB_Tag:/* This a new thread topic, so extract and save the post data */
6.      Generate PNTB object according to formula (4),(5),(9) and (10)
7.      ViewBlock = Extract_VB(PNTB);
8.      Analyze_and_Save(ViewBlock);
9.      Break;
10.   case tmp.RTB_Tag:/* Reply a view block, so extract and save the reply data */
11.     Generate RTB object according to formula (4),(8),(9) and (10)
12.     ViewBlock = Extract_VB(RTB);
13.     TID = Extract_ID(RTB);
14.     Analyze_and_Save(ViewBlock, TID);
15.     Break;
16.   case tmp.ETB_Tag:/* Update a view block, so extract and save the new data */
17.     Generate ETB object according to formula (4),(7),(9) and (10)
18.     ViewBlock = Extract_VB(ETB);
19.     TID = Extract_ID(ETB);
20.     Analyze_and_Save(ViewBlock, TID);
21.     Break;
22.   default: Break;
23.   End of switch
24. End
```

3.2 观点块存储算法

对于论坛话题,本文亦借鉴 Delta 存储[8]思想,研究以观点块为粒度的论坛话题存储方法.在存储结构中,每个话题存在两个查找入口,一个是观点块内容散列后的关键字,一个是 URL 散列后的关键字,二者分别作为不同用途的入口.图 5 以一个论坛数据为例,介绍论坛话题存储结构示意图.

依据以上结构,针对不同行为的观点块存储算法可以描述如下:

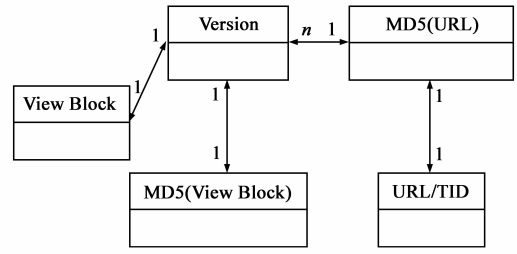


图5 论坛话题存储示意图

(1)PNTB 行为中观点块存储方法:这是一个事件树的第一个观点块,需要向库中增加的存储信息包括 < ViewBlock, MD5(ViewBlock), version > 三元组;

(2)GTB 行为中观点块存储方法:

(a)首先生成该观点块对应 URL 的散列关键字,根据该关键字快速搜索 URL 散列表.如果存在相同的值,则执行步骤 b,否则执行步骤 c;

(b)将该行为作者信息等进行统计,然后执行步骤 e;

(c)生成该观点块内容的散列关键字,根据该关键字快速搜索 ViewBlock 散列表,如果存在相同值,执行步骤 d,否则执行步骤 e;

(d)将 < URL/TID, MD5(URL) > 而元组添加到库中,并和该观点块的 version 建立连接,执行步骤 e;

(e)完成操作,退出;

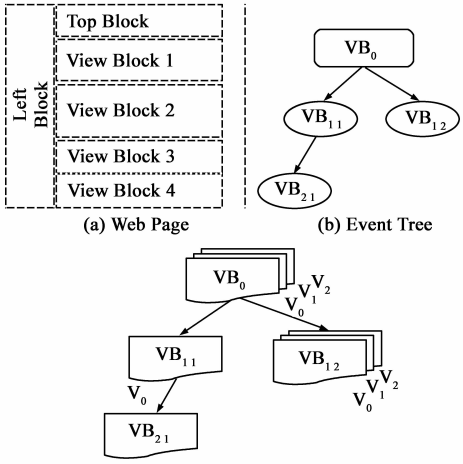
(3)RTB 行为中观点块存储方法:该行为向论坛添加了新的观点,完成观点块抽取后,需要向库中增加的存储信息包括 < ViewBlock, MD5(ViewBlock), version > ;

(4)ETB 行为中观点块存储方法:该行为更新了论坛已有的观点,完成观点块抽去后,需要根据 MD5(URL)搜索 URL 散列表,并将观点块内容以增量存储形式更新找到后的库表信息;

假设论坛某板块中 t_1 时刻用户 1 发表主帖 A; t_2 时刻用户 2 浏览主帖 A,并且针对 A 回帖 B; t_3 时刻用户 3 浏览帖子 A、B,并且针对 A 回帖 C; t_4 时刻用户 4 浏览帖子 A、B、C,并且针对 B 回帖,生成帖子 D.按照定义 1 ~ 定义 6,图 6(a)可以等价变换为图 6(b)中的事件树.按照上述存储结构及算法,数据存储模式如图 6(c)所示.

4 实验分析

本实验中选择某高教论坛实际访问数据流作为数据源.该论坛 2003 年建站,共有注册账号 7 万余人,属于匿名注册论坛,主要面向高校在校生及毕业生.通过为期近一个月的话题相关数据采集并存储,对其进行用户行为统计及热点话题分析.由于篇幅有限,我们只给出 PNTB、RTB 两种用户行为及热点话题的结果,如下

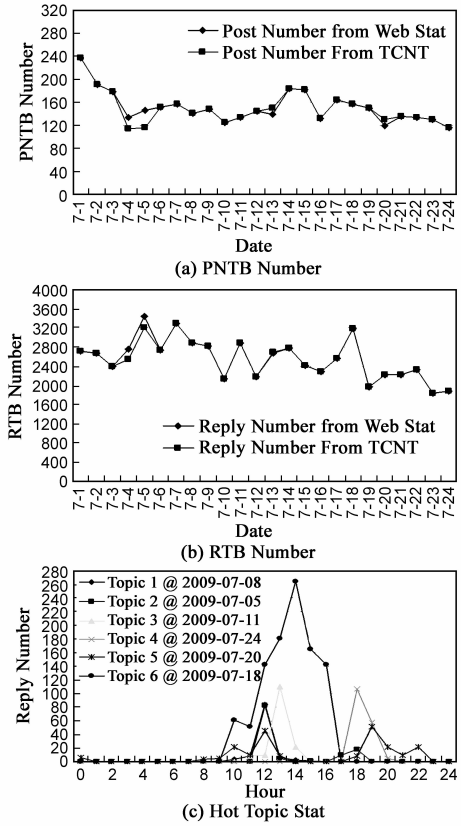


(a) Web Page　(b) Event Tree　(c) Data Repository

图6　基于观点块的话题存储模式

图 7 所示.

图 7(a)、图 7(b)中方形结实线为 TCNT 原型系统测试数据,菱形结实线为论坛网站官方提供数据.从中可以看出在 7 月 4 日、7 月 5 日 TCNT 系统输出的访问者行为数量少于网站提供的数量,这主要是因为在这期间原型系统处于更新改进中,数据获取中断了一段时间.在 7 月 13 日、7 月 20 日 TCNT 给出的用户行为多于论坛官方提供的数据,经过分析,网站在这两天均有管理员将部分不良主帖删除(隐藏),而 TCNT 能够实时将这些进行记录.图 7(c)中展示了对存储后数据经过简单分析后得出的典型热点话题随时间变化曲线.这些话题仅仅发生于当日,在页面流量上、回帖数量上有明显变化,之后便很快结束.这些特征可以作为后续研究基础.



(a) PNTB Number　(b) RTB Number　(c) Hot Topic Stat

图7　用户行为、热点话题检测结果

5　相关工作对比与分析

在网络信息获取方面,国内外研究主要通过网络爬虫(机器人)技术,比如中科天玑 Golaxy 系统[9]等.通过网络爬虫获取舆情数据缺点是该机制导致数据抓取、分析检测等具有一定的周期性,对于突发事件无法达到实时预警、跟踪与控制.本研究只针对网络中活跃信息进行截取,可以弥补网络爬虫的缺陷,对舆情实时分析、预警及管控,与传统舆情系统形成功能互补.

在 Web 页面分析方面,目前主要研究工作集中在 HTML 解析、DOM 树分析、视觉特征分析等方向,李[10]和林[11]分别给出了基于多特征的 HTML 网页内容提取方法,以及将 HTML 文档转换为 DOM 树方法.和本文相比,主要是信息获取方法不同,信息处理方法是值得在后续工作中借鉴研究的.

在网络信息抽取方面,国内外研究主要集中在建立针对各类网站的全自动化信息采集与抽取工具,并将这些信息按照一定的格式进行整合.和本文的数据采集与抽取工作相比,本文只关注网络中活跃信息流,同时可以将冗余数据块进行在线过滤,从而有效地采集论坛话题数据,进而进行数据实时分析与预警.

在网页的存储研究方面,相关研究内容包括页面的转化和处理方式、页面版本比较算法、网页的存储方式的研究等.本文根据网络数据流中蕴含话题数据的行为特征,参照文献[9]中的页面块存储方式,给出了基于观点块的存储方法,不仅可以增量存储数据流中数据,而且可以快速在线搜索及版本更新.在研究过程中,本文还参考了文献[13]中的数据存储技术,文献[14,15]中的主题抽取技术,文献[17]中的 Web 页面分割技术.

6　结论及下一步研究工作

本文提出的新型话题数据采集方法适用于网络舆情等应用,其特点是能够针对网络流中活跃的数据信息进行在线采集与分析.本研究主要贡献如下:(1)对用户访问论坛行为进行形式化定义及分类;(2)给出针对网络数据流的话题相关数据采集与存储算法,该方法将网络流中大部分信息(约占页面流量 85%)排除在抽取过程之外,节省大量时间与空间资源.下一步工作主要包括:(1)完善数据抽取算法,使之扩展为面向更多的 Web2.0 服务;(2)结合网络舆情等应用,进一步研

究在线热点话题发现算法、突发事件检测算法、话题检测与跟踪技术等.

参考文献

[1] Gupta S, Kaiser G, Stolfo S. Extracting context to improve accuracy for HTML content extraction[A]. Ellis A, Tatsuya H, eds Proc of the 14th Intl Conf. on World Wide Web—Special Interest Tracks and Posters[C]. New York: ACM Press, 2005. 1114 – 1115.

[2] 王佰玲,方滨兴,云晓春.传统报文捕获平台的性能影响因素分析[J].计算机工程与应用,2003,39(22):151 – 152.
WANG B L, FANG B X, YUN X C. The analysis of the performance factor in traditional packet capture plat[J]. Computer Engineering and Applications, 2003, 39(22): 151 – 152. (in Chinese)

[3] 王佰玲,方滨兴,云晓春.零拷贝报文捕获平台的研究与实现[J].计算机学报,2005,28(1):46 – 52.
WANG B L, FANG B X, YUN X C. The study and implementation of zero-copy packet capture platform[J]. Chinese Journal of Computers, 2005, 28(1): 46 – 52. (in Chinese)

[4] 王佰玲.基于良性蠕虫的网络蠕虫主动遏制技术研究[D].黑龙江哈尔滨:哈尔滨工业大学.2006.

[5] 王佰玲,田志宏,张永铮.奇异值分解算法优化[J].电子学报,2010,38(10):2234 – 2239.
Wang B L, Tian Z H, Zhang Y Z. Optimization of singular vector decomposition algorithm[J]. Acta Electronica Sinica, 2010, 38(10): 2234 – 2239. (in Chinese)

[6] 欧阳震诤,罗建书,胡东敏,吴泉源.一种不平衡数据流集成分类模型[J].电子学报,2010,38(1):184 – 189.
OUYANG Z Z, LUO J S, HU D M. An ensemble classifier framework for mining imbalanced data streams[J]. Acta Electronica Sinica, 2010, 38(1): 184 – 189. (in Chinese)

[7] 詹英,吴春明,王宝军.一种与缓冲区紧耦合的环形循环滑动窗口的数据流抽取算法[J].电子学报,2011,39(4):894 – 898.
ZHAN Y, WU C, WANG B. An algorithm for data stream sampling based on ring circular sliding window tightly-coupled with buffer[J]. Acta Electronica Sinica, 2011, 39(4): 894 – 898. (in Chinese)

[8] MacDonald J. Versioned file archiving, compression, and distribution[OL]. http://www.cs.berkeley.edu/~jmacd/. UC Berkeley, 1999.

[9] Golaxy 中科天玑[OL]. http://www.golaxy.cn/, 2009.

[10] 李连霞.基于多特征的 HTML 网页内容提取的研究[D].山东济南:山东大学,2008.

[11] 林昌平,郑皎凌.基于 DOM 规范的网页分析技术研究[J].成都信息工程学院学报,2007,(S1):113 – 117.

[12] TSE SourceCode[OL]. http://sewm.pku.edu.cn/src/TSE/, 2009.

[13] Gomes D, Santos AL, Silva MJ. Webstore: A manager for incremental storage of contents[R]. Technical Report, DI/FCUL TR 04 – 15, Lisbon: University of Lisbon, 2004.

[14] Sekiguchi Y, Kawashima H, Okuda H, Oku M. Topic detection from Blog documents using users' interests[A]. Aberer K, Hara T, eds Proc of the 7th Intl Conf on Mobile Data Management(MDM 2006)[C]. Washington: IEEE Computer Society, 2006. 108 – 111.

[15] Wang XY, Xiong FY, Ling B, Zhou A. A similarity-based algorithm for topic exploration and distillation[J]. Journal of Software, 2003, 14(9): 1578 – 585.

[16] McCown F. Dynamic web file format transformations with grace[A]. Proc of the 5th Intl Web Archiving Workshop and Digital Preservation[C]. 2005. 22 – 23.

[17] Lampos C, Eirinaki M, Jevtuchova D, Vazirgiannis M. Archiving the Greek Web[A]. Proc. of the 4th Int'l Web Archiving Workshop[C]. 2004.

作者简介

王佰玲 博士,副教授,硕士生导师;哈尔滨工业大学(威海)网络技术研究所负责人;哈尔滨工业大学(威海)计算机学院院长助理;北京大学信息科学与技术学院博士后;北京邮电大学兼职副教授;计算机协会高级会员.研究领域包括计算机网络安全、信息内容安全、网络攻防对抗、信息穿透对抗、网络舆情技术等.
E-mail: wbl@hit.edu.cn

张永铮 博士,副研究员,硕士生导师.研究领域包括计算机网络安全、信息内容安全、物联网技术等.

田志宏 博士,副研究员,硕士生导师.研究领域包括信息内容安全、网络信息搜索技术等.