一种基于社交媒体短链接的网页舆情热点数据挖掘方法

郭 林

(上海市公安局,上海 200025)

摘 要:近年来,网络舆情热点数据挖掘受到广泛关注。获取社交媒体数据并从中找出舆情热点较容易实现,但网页舆情热点数据挖掘现有方法存在系统开销巨大、时效性差等问题。从社交媒体中日益流行的短链接入手,分析其对网页舆情热点挖掘的价值,并在此基础上提出一种网页舆情热点数据挖掘方法。

关键词:社交媒体;短链接;舆情热点;数据挖掘

DOI:10.11907/rjdk.151894

中图分类号:TP391 **文献标识码**:A **文章编号**:1672-7800(2015)011-0139-02

1 网络舆情热点及现有数据挖掘方式

1.1 网络舆情热点

舆情是指公众对于现实社会以及社会中各种现象、问题所表达的信念、态度、意见和情绪表现的总和,相对具有一致性、强烈程度和持续性,对社会发展及有关事态的进程会产生一定影响[1]。

随着个人电脑、智能手机等设备的普及和新兴市场内不断增长的互联网访问量,网络已成为人们获取与发布信息的主要渠道。根据英国《经济学人》杂志2010年的估算,全球网络信息总量的复合年增长率已经达到60%,且正日益加快[2]。而美国国际数据公司(IDC)更是提出了“数字宇宙”的概念,并预计到2020年,“数字宇宙”规模将超40ZB[3]。通过网络这一新兴传播载体,传统舆情有了新的表现形式,即网络舆情,它具有传播迅速、信息多元、方式互动等特点。

网络舆情热点是网民思想情绪和利益诉求在网上的集中反映,是网民热切关注、集中议论的聚焦点。它是社会舆情热点在网络上的映射,反过来也会对现实社会产生深刻影响。因此,越来越多的政府部门和企事业单位都愈加重视网络舆情热点。

1.2 网页舆情数据挖掘方式

想要从浩如烟海的网络舆情中甄别筛选出热点舆情的前提是尽可能多地获取网络舆情数据。网络舆情数据的主要来源有:网站、论坛、博客、社交媒体等[4]。其中,微博等社交媒体一般都内嵌搜索引擎,对外开放了API接口,针对性的数据爬虫技术也相对成熟,且大多自带热门话题排行等功能,因此获取社交媒体数据并找出其中的舆情热点并非难事。本文将网络舆情数据挖掘的范围重点放在除社交网络之外的网页数据上。目前,网页舆情数据的挖掘主要依靠搜索引擎来实现,主要分为以下3类[5-6]:

(1)目录式搜索引擎(Search Index/Directory)。以人工或半自动方式搜集信息,人工形成摘要并分类,实现目录浏览和检索功能。这一方式费时费力、信息量少、更新慢,已不适应当前网络发展。

(2)全文搜索引擎(Full Text Search Engine)。这种引擎一般拥有自己的检索程序(Indexer),俗称“蜘蛛”(Spider)、“机器人”(Robot)或“爬虫”(Crawler),通过链接从互联网上提取各个网站的信息并建立数据库,实现整个互联网公开网页数据采集。对于互联网出现的新数据,全文搜索引擎通过一系列优化算法来提高采集效率。常用策略有:①依赖于已抓取网页中所含的外部链接,凡采集系统识别为未记录的站点或内容,就会被安排抓取任务;②识别各站点的网页更新率来调整采集频率;③依据各站点的分类属性来确定采集或解析的精度。

目前,全文搜索引擎的采集策略已经是最优化的网页舆情数据采集机制,但它的不足之处在于投入大。由于其目标是尽可能快速地采集全世界所有的互联网公开网页数据,面对的数据体量和采集频率要求极高。目前世界范围内主流的全文搜索引擎Google每年投入超过百亿美元用以维持其庞大的采集系统和搜索产品的维护与开发。同时,由于采集面过于广泛且缺乏针对性,大大降低了其数据采集的时效性。数据即便采集回来,也只能通过相同话题的聚类算法[7]来粗略判断其中的网络热点。因此,全文搜索引擎拓扑全网的采集策略并不适用于那些有明确

作者简介:郭林(1981—),男,上海人,硕士,上海市公安局工程师,研究方向为互联网技术。

数据需求，只关注互联网中极小部分网页舆情热点数据的需求。

(3)元搜索引擎（META Search Engine)。这类引擎在接到用户关键词查询请求时，同时在其它多个引擎(一般是全文搜索引擎)上进行搜索，并将结果汇总排序后反馈给用户。由于直接使用其它引擎的搜索结果，元搜索引擎在系统开发维护方面花费极低，但在时效性和网络舆情热点算法上也存在全文搜索引擎的缺点，不再赘述。

2 短链接及其价值

2.1 短链接的概念

短链接也称短网址(URL shortening)，是一种应用于互联网的有效缩短网址长度，但仍然能访问原始网络地址的技术[8]。通常包含短链接生成和地址重定向两个过程。短链接服务提供商会提供用户一个包含脚本的界面，该脚本包含请求缩短的目标长地址，系统经过滥用预防、URL过滤验证等检查后会生成一个随机字符串，并将该字符串与目标地址以某种形式关联地存储在数据库中，并返回与该字符串相关的短链接。当用户访问该短链接时，服务提供商通过数据库匹配获得相应的目标长地址，使用 301、302 或 META 等域名重定向技术引导用户访问目标网站[9]。

短链接的广泛使用，得益于推特、微博等社交媒体在全世界范围内的风靡以及其对于发布内容长度的严格限制。推特的字数限制是 140 个字符，而新浪微博的字数限制是 140 个。受制于上述限制，用户想要在这些社交媒体中发布指向外部网页的链接，不得不借助短链接来实现。据统计，仅 2009 年，美国知名短链接服务提供商 Bitly 的短链接访问数就高达 2.1 亿次[10]。

2.2 短链接对网页舆情热点挖掘的价值

帕雷托法则(Pareto principle)，也称二八定律或 80/20 法则，该法则指出，在很多情况下，80％的结果取决于 20％的原因[11]，而这一法则目前正被广泛应用于各个领域。同样，对于互联网中的舆情数据，其中具有传播价值且会成为舆情热点的数据量占全部数据的比重极少。如果能够精确获知网页数据的传播价值，就能用至少小于全网采集 20％的代价获得超过全网 80％的网页数据价值，并从中找出热点；而放弃的超过 80％的这部分数据的传播价值极低，从获取舆情热点数据的角度是可以接受的。

那么，如何精确实时地获知网络中舆情数据的传播价值呢？现有技术已经能够实时跟踪拥有数亿用户的社交媒体数据并识别其热度，因此只要将所有实时从社交网络中识别出来的指向第三方网页的短链接都解析出来，并配合包含短链接的社交媒体数据的转评情况，来度量这个短链接所指向数据的价值。通过这样的方式，至少可以获得以下两个方面的有价值的数据：

(1)网页舆情热点数据。社交媒体作为目前世界范围内最佳的信息快速传播渠道，只要在其数亿乃至数十亿用户中有人对某一第三方网页内容产生了分享或传播的冲动，就可能将其作为短链接发布在社交媒体上。全球社交网络用户有数十亿之多，每天可能访问的网站基本涵盖了全部的公开网页内容。因此，该策略就相当于让数十亿网民为系统人工筛选有价值的信息。采集这部分数据就相当于采集了最精华的网页舆情热点数据部分。

(2)实时热门网站。除了页面内容本身，所有短链接指向的外部地址都会有其顶级域名。通过一定的网站热度算法，可以实时统计一段时间内短链接指向的顶级域名的频率高低，也就能获知哪些网站产出舆情热点数据的可能性最高，并且可以识别出一些新出现的网站信源，这样就实现了对热门网站列表的实时获取。

3 基于短链接的网页舆情热点挖掘方法

3.1 社交媒体数据采集

目前，采集微博等社交媒体数据有以下两种常用策略：①通过社交媒体官方提供的 API(Application Programming Interface，应用程序编程接口)进行采集[12]。但所有社交媒体服务商都不会无条件将完整 API 开放给普通用户，其所提供的接口在等级、权限、调用次数等方面都有所限制，且返回的内容是指定的[13]，因此使用 API 的方式永远只可以解决数据获取中的部分问题；②通过程序模拟浏览器行为登陆相关社交媒体[14]，使用爬虫技术获取数据并对其进行持久化的策略。两种策略相比较，通过 API 获取数据的效率高，但是受服务商限制较大；爬虫策略效率相对低，但获得数据比较完整，稳定性更好[15]。将上述两种策略进行有效整合，可基本实现对社交媒体数据的最优采集。

3.2 短链接的过滤与解析还原

在全面采集社交媒体数据的基础上，要对其中的短链接进行过滤抽取。目前，境内外提供短链接服务的网站较多，较为常用的有 goo. gl、bit. ly、adf. ly、t. co、t. cn、url. cn 等 300 多家。但其转换后的短链接命名都遵从一定规则，一般都以上述域名作为短地址的开头，只要依据这些规则，就可以较容易地将其提取出来。此后，将抽取出的短链接通过其服务提供网站逐一进行还原，获得原始 URL 地址。

3.3 获取网页舆情热点及热门网站数据

在对获得的 URL 地址对应的网页数据进行采集时，将相关 URL 地址对应的社交媒体信息的转发数、评论数作为重要指标，制定网站热度算法，就能生成一张实时的反映网站热门程度的列表。根据列表排名，形成采集任务序列，对相关网站进行网页深度采集，以获取其全站内容。

通过以上 3 个步骤，实现了基于社交媒体短链接的网页舆情热点数据挖掘，但完全依赖社交媒体短链接可能会遗漏一些有阅读价值但是缺乏传播价值的网页数据，由此导致数据采集不够全面。因此，若对部分特定关键词的搜索引擎实时搜索结果作为数据补充，就有望实现以能够承

LVM技术在新疆地震数据存储中的应用

李亚芳,陈述新,刘杰超

(新疆维吾尔自治区地震局 网络数据中心,新疆 乌鲁木齐 830011)

摘 要:为解决新疆地震数据存储系统在使用过程中遇到的32位Linux服务器挂载单个分区空间受限和无法扩展的难题,利用Linux系统中提供的LVM逻辑卷管理机制,实现对磁盘的动态管理,使地震数据存储系统磁盘空间在Linux服务器上的挂载更加灵活方便。

关键词:数据存储;LVM;逻辑卷管理;Linux

DOI:10.11907/rjdk.151795

中图分类号:TP392 **文献标识码:**A **文章编号:**1672-7800(2015)011-0141-03

0 引言

新疆地震数据存储系统是新疆地震信息支撑平台的重要组成部分,承担着新疆地震局各子系统的数据库和各类地震数据的存储工作。现有的存储系统是一套华为S5500T磁盘存储,硬盘裸空间71T,RAID后可用空间为47T,应用模式为IP SAN架构,即以现有的IP网络为基础,存储设备双线上联至核心交换机,各子系统服务器通过接入交换机上联,使用存储系统上的磁盘空间,网络拓扑如图1所示。

存储系统上的磁盘空间通过划LUN的方式,可以分割成不同大小的逻辑磁盘,通过映射关系将若干逻辑磁盘提供给对应的服务器使用。各业务子系统服务器磁盘空间多作为数据存储或备份之用,目前有3台,分别是:前兆台网数据库热备份服务器,操作系统为SUSE Linux Enterprise Server 11 SP1(i586);测震数据存储服务器和信息网络数据库备份服务器,操作系统均为SUSE Linux Enterprise Server 10(i586)。

1 使用中常见问题

在给服务器分配和挂载存储系统上的磁盘空间过程中,遇到以下问题:

(1)新疆地震局现有的服务器多为“十五”期间配置的一批浪潮服务器,安装的操作系统多为32位的SUSE

受的代价挖掘几乎全部的网页舆情热点数据及用户特定需求数据的目标。

参考文献:

[1] 柳虹,徐金华.网络舆情热点发现研究[J].科技通报,2011(3):421-425.
[2] MANY. All too much[J]. The Economist,2010(2):6-7.
[3] GANTZ J,DAVID R. The digital universe in 2020: big data,bigger digital shadows,and biggest growth in the far east [EB/OL]. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf,2013.
[4] 石彭辉.基于社会网络分析的网络舆情实证研究[J].现代情报,2013(2):27-31.
[5] 王晔.垂直搜索引擎若干问题研究[D].上海:复旦大学,2011.
[6] 褚建立,刘彦舫.计算机网络技术实用教程[M].北京:华大学出版社,2007.323-324.
[7] 孙学刚,陈群秀,马亮.基于主题的Web文档聚类研究[J].中文信息学报,2003(3):12-16.
[8] WILLIAM H,DUTTON. The Oxford Handbook of internet studies[M]. Oxford: Oxford University Press,2013:87.
[9] 薛富,高一男.基于内容提取的短链接生成算法研究[J].网络技术安全与应用,2014(2):114-115.
[10] JENNA WORTHAM. Googl challenges bit.ly as king of the short [EB/OL]. http://bits.blogs.nytimes.com/2009/12/14/googl-challenges-bitly-as-king-of-the-short/,2009.
[11] BUNKLEY NICK. Joseph juran pioneer in quality control dies [EB/OL]. http://www.nytimes.com/2008/03/03/business/03juran.html,2009.
[12] 卢体广.微博舆情系统中数据采集技术研究[D].湘潭:湘潭大学,2014.
[13] 廉捷,周欣,曹伟,等.新浪微博数据挖掘方案[J].清华大学学报:自然科学版,2011(10):1300-1305.
[14] 纪伟.微博数据采集系统的设计与实现[D].石家庄:河北科技大学,2013.
[15] 俞忻峰.社交网络挖掘方案研究[J].现代电子技术,2015(4):25-29.

(责任编辑:陈福时)

作者简介:李亚芳(1985—),女,山东枣庄人,硕士,新疆维吾尔自治区地震局网络数据中心工程师,研究方向为网络数据。