电脑知识与技术 COMPUTER KNOWLEDGE AND TECHNOLOGY

目 次（第13卷第28期 2017年10月）

数据库与信息管理

1 云模式下智慧城镇大数据系统设计 …… 黄 炼
3 基于HTML5/.NET的新型高校图书馆数字资源系统设计与实现 …… 龙德应，唐嫦燕
5 高职专业教学资源库资源更新机制研究 …… 孙 祎
7 浅析“大数据”时代的计算机信息处理技术 …… 梁剑波，柴 群
8 基于NPOI数据导出方法的研究与实现 …… 连俊光
10 基于数据挖掘的课堂教学评价体系构建研究 …… 张继成，杜松江，张 路
12 基于solr的汽车数据检索组件的设计与实现 …… 陈 辰，唐兰文，张 超
15 基于Hadoop的油气信息分布式数据仓库的探究 …… 鲁帅帅，彭甲勇
18 基于MemCache的分布式扩展算法 …… 齐建军

网络通讯及安全

20 中国地区Internet特性分析 …… Dmytro Lande，李 晶，杨子江，Boris Berezin，周晓明，董 婷
24 云计算网络机房工程设计思路概述 …… 张 欣，徐振宇
26 软件定义存储的应用与分析 …… 徐华宇
28 基于改进朴素贝叶斯算法入侵检测系统研究 …… 曾国斌，冉兆春
30 关于计算机信息安全技术分析及防护策略探讨 …… 李 珊
32 基于证据推理算法的入侵检测系统 …… 王伍柒，周立萍
35 计算机网络安全防范技术的研究和运用 …… 王 锋
36 基于移动终端的智慧校园一站式服务平台的研究与实现 …… 罗金玲
38 大数据在应急管理中的应用 …… 胡淑新，王小可
40 数据结构中遍历操作的非递归算法 …… 詹泽梅
43 基于复杂网络理论的微信网络分析研究 …… 迪丽努尔·库尔班，阿布力米提·艾西丁
45 基于计算机云平台的商品跟踪系统分析与研究 …… 韦建国，宋丽萍
48 新时期高校网络安全分析与防范措施 …… 唐 旭，陈 蓓

软件设计开发

50 北关街道办事处网络服务平台系统设计 …… 宁 蕊
52 基于微信平台的高职院校移动学习平台的设计与实现 …… 全丽莉
54 网络课程学习管理平台的设计与开发 …… 邝嘉伟
57 基于HBuilder的极限运动网站分析与设计 …… 吴妍妍
60 药店销售管理系统的分析与设计 …… 周 波
62 ExamQA的构建及其在考试统计与质量分析中的应用 …… 姚秋阳，吴发明，何芋岐，杨建文，聂绪强
65 基于多维图形数据结构的连连看游戏程序设计方法探析 …… 宋兰霞，潘承毅，周作梅，洪 保，孟万堃
67 多肉绿植店管理系统分析与设计 …… 客美玲

期刊基本参数：CN34-1205/TP * 1999 * *16 * 288 * zh * p * ￥35 * 9000 * 122 * 2017-10

ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No.28, Oct. 2017

E-mail: jslt@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

中国地区Internet特性分析

Dmytro Lande[1,2], 李晶[3], 杨子江[1], Boris Berezin[2], 周晓明[4], 董婷[1]

(1.山东省科学院情报研究所,山东 济南 250014; 2.乌克兰国家科学院信息记录问题研究所,乌克兰 基辅 03056;3.山东广电网络有限公司青岛分公司,山东 青岛 266001;4.青岛市技术转移中心有限责任公司,山东 青岛 266001)

摘要: 文本通过与全球Internet资源对比的方法,评估中国地区Internet的独特特性,并在此基础上分析了采用RSS源采集中国地区Internet信息的可能性。

关键词: 中国地区Internet;网络资源特性;RSS源;信息采集

中图分类号: TP311 **文献标识码:** A **文章编号:** 1009-3044(2017)28-0020-04

1 概述

随着Internet的发展,中国已成为全球网络用户数量最多的国家,目前超过6.88亿占全国总人口一半以上的用户使用Internet。而Internet发源地美国的用户量仅为2800万排名第三。同美国相比,中国地区Internet的发展有其自身的特色[1-3]。首先,通过移动客户端访问Internet的用户远超美国,中国约有90%的用户通过智能手机等移动客户端访问Internet,而美国仅有40%的用户通过移动客户端连入Internet;其次,在线发布内容具有高活性和高稳定性,中国在线发布内容超过美国大约20%-50%;最后,用户年龄段的构成不同于美国,中国20-29岁的用户占比最高大约30%左右,10-19岁的用户次之约占22%,另外还有24%的用户年龄段处于30-39岁。

Internet用户量排名前37位的国家用户使用Internet比例如图1所示。图中橙色和灰色部分分别表示使用和未使用Internet的用户量。该图可以反映某国用户对Internet的贡献情况,以及国内Internet的使用程度。

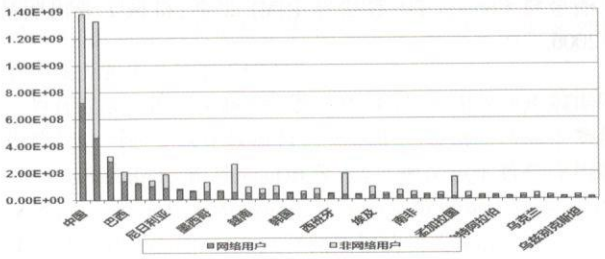

图1 全球Internet网用户比例分布

中国地区Internet用户增长状况如图2所示[4]。其中横轴表示年份,纵轴表示用户量,单位为百万。

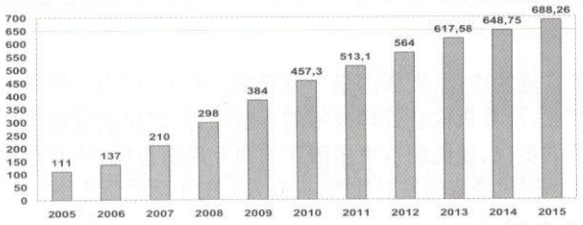

图2 中国地区Internet每年用户增长情况

中国拥有423万个网站和2123亿个网页,其每年的增长情况如图3所示。其中横轴表示年份,纵轴左侧表示网站的增长情况(单位:百万),右侧表示网页的增长情况(单位:十亿)。绝大多数网站使用汉语,仅有少量使用英语,这为欧美国家的用户访问增加了难度,幸好随着Google翻译等软件的逐渐成熟,对解决因语言障碍问题而引起的用户访问困难起到了很大的帮助。

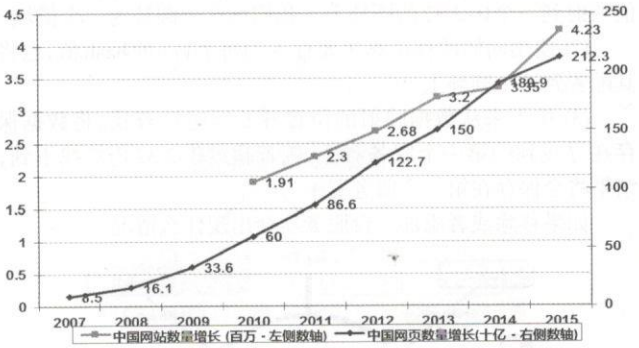

图3 中国地区Internet网站数量和网页年增长情况

目前大多数文献仅从用户量、网站和网页数量等统计性特征分析了对中国地区Internet的情况,很少有从采集角度对内容层面的特征进行深入分析。本文首先使用对比分析的方式对中国地区Internet资源特征进行总结,然后探讨了利用RSS和网络资源监控软件采集中国地区Internet网站的可能性。

2 中国地区Internet特性分析

中国地区Internet内容的采集受一系列因素的影响,包括:网站及网页数量,区域分布,语言和编码,web文档数据格式、报纸、新闻机构、教育和科研机构门户网站,开放出版物,社交网络等。文献[1-4]已对这些因素进行了分析,在此不再赘述。然而网站内容的访问往往并不是通过直接输入URL地址实现的,而是依赖于搜索引擎及网站的索引。而不同搜索引擎在不同国家地区Internet覆盖情况取决于搜索引擎所属国家、托管网站搜索国家以及网站的类型(如商业、政府、组织、大学网站等)[5,6]。文献[7,8]提出了对搜索引擎索引的评价及其可视化方法。鉴于现有研究成果,在对中国地区Internet特性进行分析

收稿日期:2017-08-05

师，除合理使用文献提供的数据外，还应将中国地区Internet与其他国家地区Internet进行比较，使用对比的方法发现中国地区Internet不同于其他国家和地区的独特特性。

2.1 网站数量

由文献[4]提供的数据可知，2010年底中国地区网站总量为191万，到了2015年底网站数量达到423万。中国不同区域网站分布情况如图4所示。其中横轴表示区域，纵轴表示网站数量。由图可知，广东省的网站数最多约67.1万，占总量的15.9%，而西藏的网站数最少仅为1000。

截止2016年9月全球网站总数为10.8亿，根据Web服务器的监测显示其中活跃的网站数量约为1.73亿。由此可以看出，中国地区网站仅占世界活跃网站总数的2.4%左右。而国土面积远小于中国的乌克兰地区，其拥有网站532万，占世界活跃网站总数的30.7%左右。

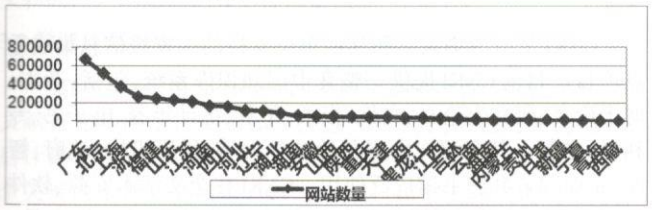

图4 中国各地区网站分布情况

2.2 网页数量

2006年5月搜索引擎baidu为用户提供了超7.4亿次网页访问，而到2015年底网页数量达到了2123亿。中国各地区网页的分布情况如图5所示，其中横轴表示区域，纵轴表示网页数量。由图可知，北京市网页数量最多远超850亿（其中静态网页500亿，动态网页340亿），青海网页数量最少约为3400万（其中静态网页2000万，动态网页1300万）。全国超2120亿的网页中，静态网页1310亿，动态网页800亿，两者的比值约为1.63。具体到中国某区域时，这一比值分别从重庆的4.3和江苏的3.19，到宁夏的0.37和新疆的0.5不等。

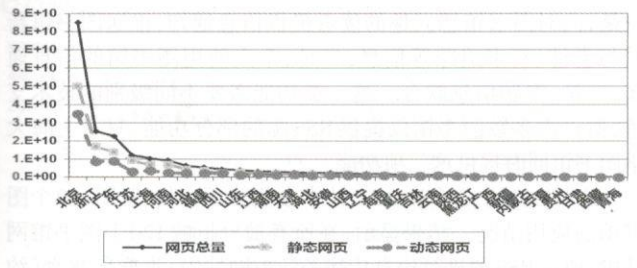

图5 中国各地区网页分布情况

2016年9月，全球被搜索引擎列入索引的网页数量不少于47.2亿[9]。2005年全球被编入索引的网页数量约为115亿，而2015年编入索引的网页数量超过3045亿[10]。

2.3 网页更新频率

中国各区域网页更新周期如图6所示，其中横轴表示区域，纵轴表示网页更新比例。不同颜色代表不同更新周期，其中青、红、黄、蓝、紫分别表示更新周期为周、月、3个月、半年以及半年以上。由图6可知，每周更新的网页占比最大的省份是甘肃省约为10.2%，而超6个月更新的网页占比最大的省份是海南约为22.6%。不同周期网页更新比例的平均值分别为：4.5%、24.4%、33%、27.6%和10.5%。

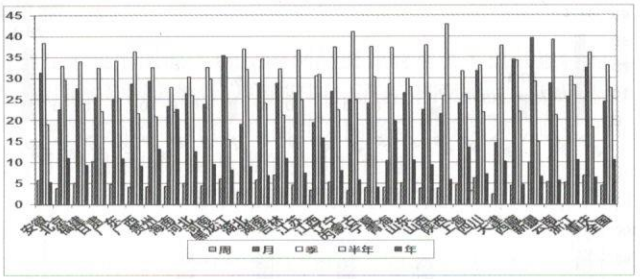

图6 中国各地区网页更新周期

利用Google系统高级搜索接口获取的数据，绘制全球网页更新周期如图7所示。仅有0.23%的网页更新周期为一天，1.5%的网页更新周期为一周，而80%以上的网页更新周期超过一年。

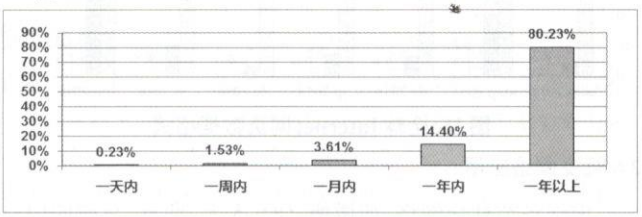

图7 全球网页更新周期

2.4 网页语言

中国各区域网页使用的语言字符集如图8所示，其中横轴表示区域，纵轴表示网页占比。不同颜色代表不同语言，其中青、红、黄、蓝分别表示中文、方言、英语和其他语言。

借助于Google和Bing对.com和.cn域名使用的语言进行评估，全球Internet存在约5亿.cn和约1亿的.com域名的网页使用中文，超700万.cn域名，超50亿应用.com域名的网页使用英语，超5000万.cn和超30亿.com域名的网页使用德语，超1000万.cn域名和约5亿.com域名的网页使用法语。

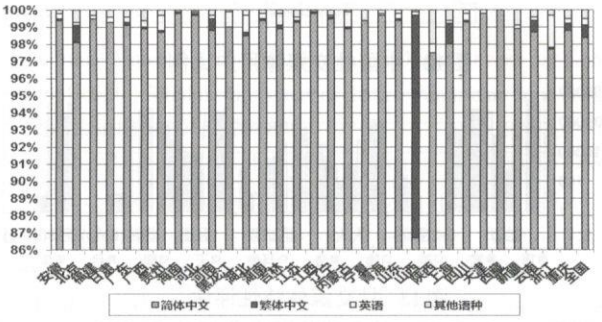

图8 中国各区域网页使用语言情况

2.5 网页格式

中国网站的网页和媒体应用的数据格式如图9和图10所示。由图可知，中国doc格式文件远超PDF格式文件，而2013年全球Internet中pdf格式文件是doc和docx文件的6倍。图11展示了pdf、doc/docx、rtf、txt等格式在Web文件中的占比。

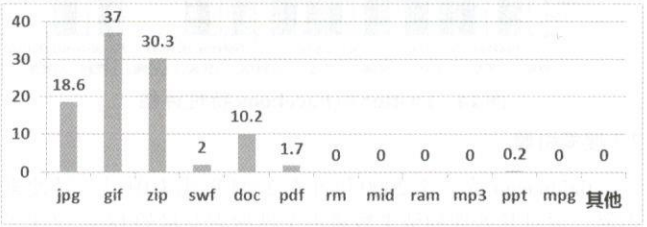

图9 中国网页媒体数据格式

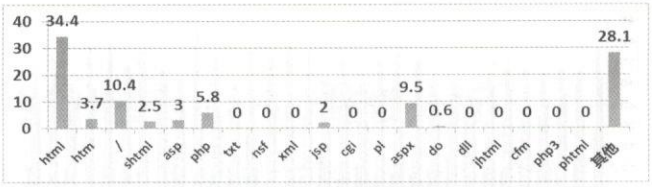

图10 中国网页编辑语言

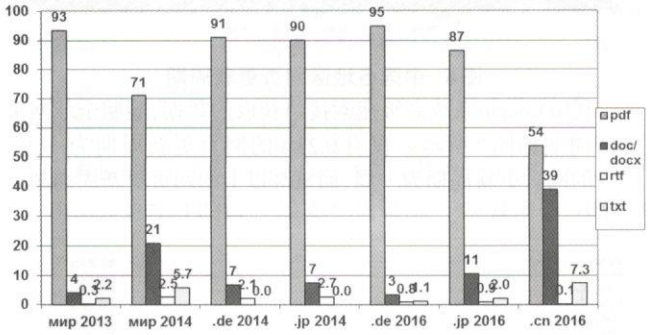

图11 全球Internet网页数据格式

2.6 社交网络应用

中国各类社交网络，如微博、QQ、人人、朋友、豆瓣的用户比例如图12所示。借助Google和Bing对各类社交网络特性的评估如图13所示。图例中不同的颜色依次代表简体中文、繁体中文、英文、德文、法语、其他语言网页的数量，以及近24小时、一周、一个月、一年内的网页数量。

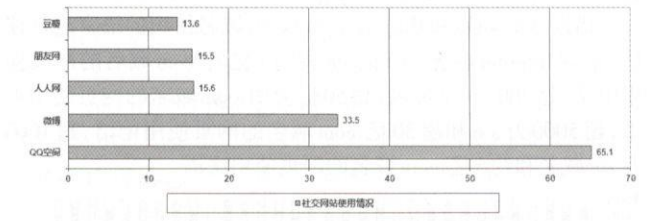

图12 中国社交媒体用户比例

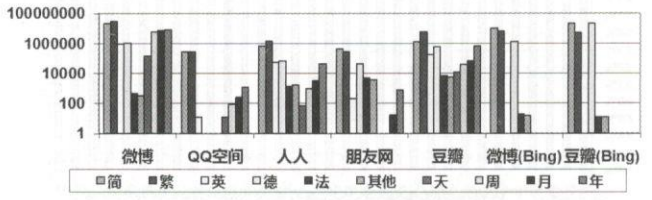

图13 社交媒体特性评估

借助Google和Bing对全球范围内典型社交网络Twitter和Facebook的评估如图14所示。通过图13和图14的对比可知，中国的社交网络被搜索引擎评估的网页数量以千万计，而在全球社交网络被搜索引擎评估的网页数量以亿计，远超中国的数量。

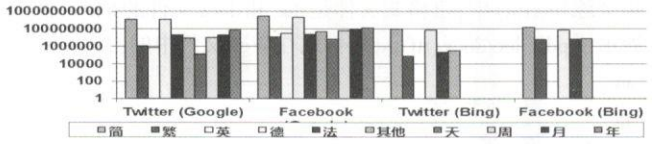

图14 Twitter和Facebook特性评估

2.7 搜索引擎

Baidu.com成立于2000年并在2004年成为中国头号搜索引擎。通过其处理的请求数量占全球搜索总量的18%，仅次于Google。2006年百度向用户提供超过7.4亿网页、8000万图像和1000万媒体文件的检索。

2015年12月搜索引擎拥有5.66亿用户，跻身中国Internet第二大常用的基本应用。其中Baidu搜索系统用户使用量约为91.2%，手机端用户90.3%。其后为Soso/Sogou搜索系统用户使用量约为45.8%，360搜索系统用户使用量约为38.6%，Google搜索系统用户使用量约为27.4%。

2.8 科技文献资源

中国地区各类提供科技文献资源中，最突出的是Baidu学术和CNKI。

Baidu学术创立于2014年，以百度搜索系统为基础。至2014年底Baidu学术收录了数十万科学网站并索引上亿出版物，提供对国际和国内资源的免费访问。截至2014年底，Baidu学术服务每天的访问量达到800万。其中约20%的请求为英文。

CNKI是由清华大学和其他单位支持的国家级信息集成重点项目。目前CNKI提供一整套中国知识库系统，包括：杂志、博士论文、研究生论文、文献、报纸、年鉴、统计年鉴、电子书、专利、标准等。其资源在中国各地被各大学、科研机构、政府、智库、企业和公共图书馆广泛使用。CNKI在全文学术资源、软件数字化和知识管理领域整合新的内容并开发了新的产品。CNKI目前成为中国规模最大、应用最广的在线数字图书馆。

3 Internet资源采集

RSS (Rich Site Summary，丰富站点摘要)，用于频繁变动信息的发布，是一项用户定制感兴趣网页更新的技术。2004年RSS源数量仅为30.7万，到2016年Feedage.com目录收录的RSS源超过31亿。2005年约有30%的用户采用RSS源[11]获取内容，截至2008年这一比例增至50%。

文献[12]研究了Web 2.0技术，如社交网络、wiki技术、博客、RSS、即时通讯和编目功能在中国顶尖38所大学图书馆的应用。结果显示，RSS应用频率排名第二，约有55%的大学图书馆使用该技术。大学图书馆最常采用RSS的三个基本功能：一是，向对图书馆感兴趣的读者提供信息通知，推送图书馆新闻与事件、新书追踪等信息；二是，个人使用图书馆的信息通知；三是，专题信息联合。这三类功能需要不同级别的技术支持，所以大多数图书馆仅提供RSS源的部分功能，只有上海大学图书馆同时提供这三项功能。

文献[13]研究了Web 2.0技术在北美、欧洲和亚洲120个图书馆的应用情况。结果显示，在所有被分析的120个图书馆网站中，通过RSS源进行信息传播的学校网站中，北美有28个(约占70%)，欧洲与亚洲分别为17和15个(占比分别为43%和37%)。RSS源在三个地区大型图书馆中的平均应用率约为50%，在Web 2.0应用排行榜中，紧跟微博之后位列第二。美国使用RSS技术的比例最大，在100所科技图书馆97%使用了该项技术。

中国及全球其他地区RSS源的使用情况如图15所示。由图可知全球过半的图书网站使用RSS源，该比例超过亚洲各国平均使用率，却低于欧美平均使用率。

3.1 数据源分类

为评估利用RSS源采集中国地区Internet网站信息，将Internet网站资源分为以下几类：报纸门户网站、新闻门户网站、

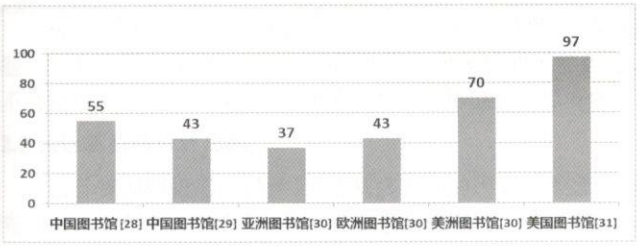

图15 RSS资源使用情况

高校和院所网站、国家机构网站、法律信息网站。对每类网站我们根据Alexa排行榜选出前20位的网站进行评估。结果表明，多家顶尖报社及新闻门户网站以及文献[12,13]中列出的大型图书馆网站，均利用RSS源进行信息传播。此外中国报社网站分析表明，约有40 %的中文网站和50 % 英文网站使用RSS进行信息传播。约60 % 的中文新闻门户网站和约70 % 的英文新闻门户网站应用RSS传播信息。

3.2 微博应用分析

文献[2]对微博和Twitter两个社交媒体进行了对比分析，该文献首先从两个社交媒体中挑选50个热门话题的关键词，然后计算包含关键词的热门话题出现频率。结果表明，微博中每个关键词出现的平均时间约为6小时，每个主题出现的时间分布符合幂规则，这表明这些热门话题中只有少数主题具有长期流行的特点。而Twitter中推特每个关键词出现的平均时间约为20-40分钟，其主题时间分布与微博相似。两者在关键词出现时间上的区别说明微博上具有竞争力的话题要少于Twitter。

为更进一步分析微博的特征，本文对近1小时的关键词进行抽取，所得到5天内的关键词变化曲线如图16所示。其中横轴表示关键词，纵轴表示信息量。我们以口袋妖怪和快乐大本营两个关键词为例说明微博的特征。微博中关键词口袋妖怪在2016年7月22日14:00和2016年7月23日4:00出现的次数分别占据top-50排行榜的第18位和第9位，信息量分别为1万和16万。关键词快乐大本营在在2016年7月22日15:00和2016年7月23日9:00出现的次数分别占据top-50排行榜的第47位和第2位，信息量分别为2万和20万。

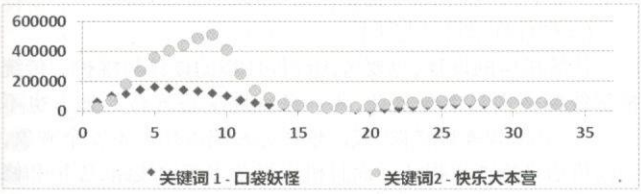

图16 微博关键词变化情况

基于百度和Google的搜索服务，对两个关键词的搜索量变化情况进行分析。图17绘制了两个关键词的搜索变化情况，由图可知两个关键词的搜索变化极大。

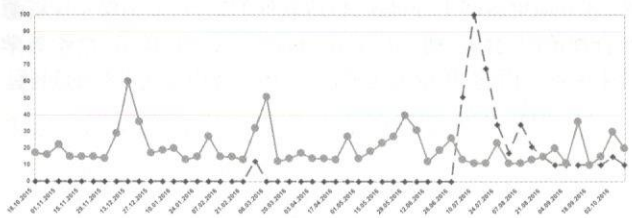

图17 关键词搜索量变化情况

4总结

在采集中国地区Internet网络资源时需要考虑其独特的特性，如网站及网页数量；更新周期与语言；网页格式；报纸、新闻门户网站的流行性评估；中国社交网络应用数据等。通过研究总结中国地区Internet资源的主要特性有：

(1)网络资源与用户数量增长速度超互联网全球范围的水平；

(2)拥有自己的社交媒体，更新总量超全球范围内同类别社交媒体；

(3)拥有自己的搜索引擎百度、搜狗等，这些搜索引擎在中文搜索领域拥有绝对优势，并显著占据中国市场；

(4)目前RSS源应用相对较少，但RSS源应用呈上升趋势，尤其是在移动端。

参考文献：

[1] Deans P.C., A framework to understanding social media trends in China, The 11-th Internation. DSI and APDSI Joint Meeting, Taipei, Taiwan. July 2011:12-16.

[2] Yu L., Dynamics of trends and attention in chinese social media, arXiv preprint arXiv:1312.0649, 2013:1-17.

[3] Bolsover G., Social Foundations of the Internet in China and the New Internet World: A Cross-National Comparative Perspective, Oxford Internet Institute, University of Oxford, 2013: 1-22.

[4] 37次中国互联网络发展状况统计报告,2016

[5] Vaughan L., Equal representation by search engines? A comparison of websites across countries and domains, Journal of Computer-Mediated Communication, 2007:888-909.

[6] Vaughan L., Search engine coverage bias: evidence and possible causes, Information processing & management,2004:693-707.

[7] Orduña-Malea E., The dark side of Open Access in Google and Google Scholar: the case of Latin-American repositories, Scientometrics,2015:829-846.

[8] Orduña-Malea E., Methods for estimating the size of Google Scholar, Scientometrics,2015:931-949.

[9] Bosch A.,Estimating search engine index size variability:a 9-year longitudinal study, Scientometrics,2016:839-856.

[10] Gulli A., The indexable web is more than 11.5 billion pages, Special interest tracks and posters of the 14th international conference on World Wide Web. ACM, 2005:902-903.

[11] Ma D., Use of RSS feeds to push online content to users, Decision Support Systems,2012:740-749.

[12] Han Z., Web 2.0 applications in top Chinese university libraries, Library Hi Tech,2010:41-62.

[13] Si L., An investigation and analysis of the application of Web 2.0 in Chinese university libraries, The electronic library, 2011: 651-668.