

网络舆情信息提取技术研究与实践

刘华春, 王星捷

(成都理工大学 工程技术学院, 四川 乐山 614007)

摘要:网络舆情信息提取是舆情分析系统中最为关键的部分,是实现舆情分析、舆情统计的数据基础。为此,设计并实现了一个基于话题线索的舆情信息提取方案。该方案将舆情页面以话题为线索进行逻辑划分;采用基于DOM树的广度优先搜索方法,设计了舆情信息提取算法;通过设置最低重复话题阈值 θ ,用户定制提取格式,信息去重去噪措施,实现了舆情信息的有效提取。通过对多个论坛舆情信息的提取实验,结果表明,所设计的方案有很好的提取性能,召回率、准确率、 F 指数都较高,能够很好地提取出论坛、评论等舆情信息。

关键词:舆情信息;Web信息提取;话题线索;DOC树

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2016)09-0008-04

doi:10.3969/j.issn.1673-629X.2016.09.002

Research and Implementation of Information Extraction Technology in Network Public Opinion

LIU Hua-chun, WANG Xing-jie

(Engineering & Technical College of Chengdu University of Technology,
Leshan 614007, China)

Abstract:Internet public opinion information extraction is the most critical part of public opinion analysis system, which is also a data base of the public opinion analysis and statistics. For this reason, a public opinion information extraction method based on clues topic is designed and implemented. In the method, pages of public opinion as one topic clue is divided to logical region, and the breadth-first search methods based on DOM tree is applied to design extraction algorithm of public opinion information. By setting a minimum repeat topic threshold θ , customized extraction format, removed duplicate and noise of information, public opinion extraction is realized effectively. By experiment of the public opinion of multiple forums, the results show that this scheme has good extract performance, and the recall, the correct rate and F measure are higher, which is able to well extract forum and reviews and other public opinion information.

Key words: public opinion information; Web information extraction; topic clues; DOC tree

0 引言

网络舆情系统是对网络中的舆论信息进行采集、检测、监控的互联网信息系统。用户针对所关注的舆论话题,能够快速检索所关注网站、论坛及以微博为代表的自媒体上的言论,对舆论观点分类,做出分析和预测预警。通过对舆情信息的过滤、提取、分类、聚类、主题监测、专题聚焦、自测等技术,使用户即时掌握网络舆情状态。

网络舆情系统通常包括数据采集、网页信息抽取、数据统计分析、舆情数据处理和系统管理等。网页信息抽取是网络舆情系统中极其关键的部分^[1]。网络舆

情信息主要来源于新闻报道、各种论坛、微博等,这些信息是非结构化或半结构化的。需要将其抽取、转换为结构化的信息,存入数据库中,使得采用成熟的基于数据库的各种查询和统计、分析舆情信息成为可能^[2]。结构化、规范化的各种舆情数据是网络舆情系统数据处理、舆情分析模块的基础和前提。

1 网页信息抽取技术分析

网页信息抽取是从采集到的网页中提取相关数据信息的过程,其研究内容是针对需要抽取信息的网站,研究其页面信息的分布规律,通过构造抽取规则,寻求

最为高效和准确的抽取方法,抽取网页中的信息,以供网络舆情分析使用。

传统的网页信息抽取方法是构造一个具有特定规则和针对性的包装器 Wrapper^[3]。包装器从采集的网页中提取所需要的数据信息,并将这些数据转化成恰当的格式,如 XML、表格等^[4]。目前,出现了很多采用不同技术而改进的包装器,如基于 HTML 文档、统计方法、DOM 文档、视觉的技术等等。

(1) 基于 HTML 文档的提取。

该类提取技术主要根据抓取的 HTML 文档的结构特点,制定一套正则表达式,过滤出需要的数据信息。也可采用 HTML 解析工具,如 HtmlParser 解析器,通过匹配 HTML 标签,抽取出网页中所需的信息。该类抽取技术优点是技术简单,抽取准确率高;缺点是通用性差,需要针对各类待抽取网页的特征单独制定抽取模板^[5]。

(2) 基于统计特征的提取。

该类提取技术是基于网页的文本信息与标签信息的比率关系。如网页中某块中文与 HTML 代码的比例,正文信息与周围超链接的比例,逗号、句号使用频率等文本特征,判别出该信息是文本信息还是广告导航之类的信息,从而抽取出需要的文本信息^[6]。该类抽取技术缺点是准确率不高,而且无法抽取 BBS 论坛信息。由于论坛类网页中各人语言的随意性,使得各个楼层正文信息长短不一,风格各异,所以难以采用该类方法。

(3) 基于 DOM 的提取。

该类抽取技术是采用 DOM 文档对象模型,即将 HTML 或者是 XML 这类文件理解或者说解析成一种文档对象,把 XML 文档里的各个标签视为节点对象,即 DOM 树,根据 XML 的节点信息,解析出所需的文本信息^[7]。将该技术用于 BBS 论坛网页抽取,具有明显优势。由于 BBS 论坛每一层的样式相同,反映在 HTML 代码上,各层都具有相同的兄弟节点,所以,可以制定通用的抽取模板。

(4) 基于机器学习技术的提取。

将目前非常流行的机器学习技术应用于网页信息的提取。机器学习是采用某种学习算法(如 BP 神经网络、SVM 支持向量机、关联、聚类等)进行数据模型训练学习,得到一种模型,再用此模型进行实际检测提取^[8-11]。其优点是自动化程度高,缺点是提取准确性较差。

在当前的 Web 网页中,绝大多数是新闻类网页,少部分是 BBS 论坛类网页。目前几乎没有一种通用模板可以包含这两种类型的网页。而网络舆情系统除了正文信息提取外,还需要统计作者名称、发帖时间、

回帖人名称、回帖时间等内容。因此,网络舆情信息抽取技术越来越趋向于算法的复杂化,是多种提取技术的交叉和综合应用。

2 网络舆情系统信息抽取

网络舆情是民众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和。在网络环境下,舆情信息的主要来源包括新闻评论、社区论坛、博客、微博等。网络舆情主要以话题的形式存在和传播。

2.1 网络舆情信息抽取特点

从信息资源特点来看,每一种信息资源特点都不一致,如论坛的文本通常较短,且用语多非书面化,在信息抽取时需要较多的样本和词典支持。新闻评论是跟帖的较多,各条评论之间关系复杂^[12]。为此,文中提出一种独立于舆情信息源的信息抽取方法,即面向话题的信息抽取方法。

2.2 面向话题的舆情信息抽取

(1) 话题线索抽取。

网络舆情信息抽取就是将基于某一话题的信息进行抽取,分析,统计。这些半结构化的信息主要分布于各类评论、论坛中。在论坛中,其结构为标题页加内容页面形式,标题页即为话题,标题链接内容页面,内容页面即为某一话题的评论内容。在各类评论中,其结构为话题加评论,话题为新闻、口碑等,评论为对该新闻或口碑的评价^[13]。为了便于浏览,通常一个页面所显示的内容是固定的,当内容超出一页时,采用多页显示,如图1所示。信息内容页主要显示话题内容及对

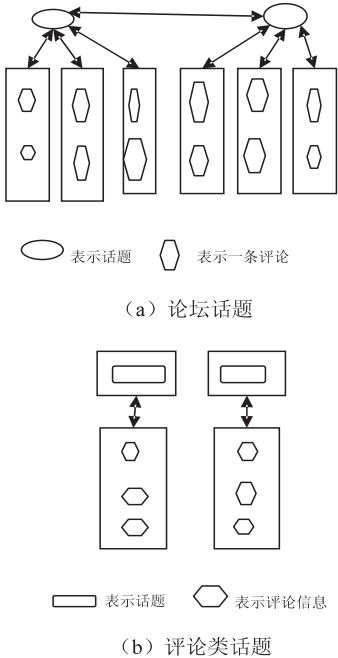


图1 论坛话题和评论类话题的组织结构

该话题的各种评论和链接。

由于论坛或评论在 Web 页面中大都采用同一功能的 CGI 模块来生成统一格式的 HTML 页面,发帖人传入的参数也是具有规律的,其 URL 具有相似的结构^[14-15]。因此,可以根据用户选定来生成特定 URL 类的匹配模式,实时地提取舆情信息。

话题线索抽取算法描述如下:

①判别是论坛类信息源转②;如果是评论类信息源转⑤。

②论坛类信息源:从标题页中提取每个指向消息内容页面的链接,初始为未处理,表示该链接为某一话题的起始位置,下载该 URL 指向的消息页面。

③提取同话题的消息页面内容。将话题线索中指向该消息页链接处理标志置位已处理。

④递归转②处理,判别下一话题链接处理标志,若未处理转③,全部已处理转⑥。

⑤评论类信息源:从消息话题页提取话题和评论,置处理标志为已处理。若全部话题页标志为已处理,转⑥。

⑥结束话题线索抽取。

(2) 信息内容提取。

信息内容提取的目的是将半结构化的 HTML 形式话题,提取其属性值,如发帖人、发帖时间、话题内容、点赞数、转发数等信息,将其转换为结构化的信息内容记录,存入数据库表中,重构结构化的话题线索,为舆情分析、统计提供数据基础。

通常一个信息页面中包含多条信息,每条信息即是一个话题内容或一个评论内容块的信息。在 HTML 结构中,每一个信息块是 DOM 树的一个相对独立的子树,子树之间有相同的父节点,子树呈兄弟节点关系,其内部结构特征相同,如图 2 所示。div 下都是相同的结构,代表了一条信息,因此,用户指定一个信息节点的处理方式,系统能够自动处理其他节点。

```
<div class="x-reply font14" xname="content">
  <div class="w740">
  </div>
</div>
<div class="x-reply font14" xname="content">
  <div class="w740">
  </div>
</div>
```

图 2 论坛信息的树状结构

(3) 舆情信息提取算法。

论坛页面由于其具有重复子树的特点,由前两节可知,论坛舆情信息提取的算法核心是基于重复模式的 DOM 树遍历。文中采用广度优先搜索算法遍历

舆情论坛 DOMDocument。广度优先遍历算法是从树的根节点开始,依次遍历下一层的子节点。由于舆情论坛回帖信息大部分是从属于某一个话题节点,即父节点,回帖节点信息大都是平行的,因此采用广度优先搜索算法是最合适的。具体算法流程如图 3 所示。

输入: docnode, HTML 页面转换成的 DOM 树根

节点: θ , 重复子树模块的阈值

输出: 子树节点信息

流程:

(1) 初始化队列;

(2) 将 DOM 树根节点加入队列;

(3) 遍历队列;

(4) 重复计数器初值为 0;

(5) 遍历每个子节点;

(6) if 如果重复子节点数 = θ then

(7) 返回当前节点元素;

else

图 3 论坛信息提取算法

该算法采用一个队列来实现 DOM 树的广度优先搜索过程,循环测试是否找到符合条件的节点,如果找到,并且总数大于设定的阈值 θ ,退出循环,算法结束。重复子树模块阈值 θ ,是具有相同子树的节点统计值,预先设定,如果页面中相似的节点出现的次数大于 θ ,这些节点就为同一话题节点。

(4) 信息去重去噪。

网络舆情信息提取需要处理的数据量巨大。在海量数据提取的过程中,最主要的是不再保存重复的提取信息,这样可减轻数据存储时的负担,并且为分析数据提供方便。文中的舆情信息自动抽取技术在存储数据时对数据库进行了优化,为了避免重复数据的采集,采用 HashCode(哈希值)作为表的索引。以论坛为例,通过对作者、时间、标题这 3 个字段组成的字符串进行哈希运算,由于重复的对象具有相同的哈希值,这样有效避免了重复信息的存储,极大提高了数据库的查询效率。

3 系统实现及实验结果分析

3.1 系统实现

网络舆情的信息源站点具有不同的页面格式,因此,文中所提出的抽取系统可以根据用户设定的抽取规则定制抽取模块。如图 4 所示,舆情信息抽取系统分为规则定制部分和信息抽取部分。规则定制部分流程:抽取样本页面,定制舆情话题线索抽取规则,生成 XML 格式的抽取规则模块。信息抽取部分工作流程:启动舆情话题线索抽取引擎,系统根据生成的 XML 抽

取规则,从舆情信息源站点抽取合乎规则的预期信息结果文件,保存在数据库和 XML 文件中。

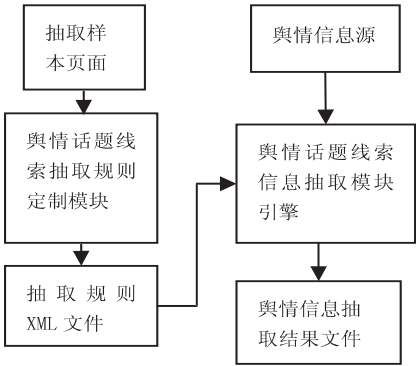


图4 舆情信息抽取系统实现图

3.2 实验结果分析

(1)性能评价指标。

MUC(Message Understanding Conference,消息理解会议)为信息检索和信息提取领域内的算法性能测试提供评估参数,主要有召回率 R (Recall)、正确率 P (Precision)和 F 指数。召回率是指正确抽取的记录占被抽取页面中所有记录的比例;正确率是指所有抽取出来的记录中正确抽取的评论记录所占的比例。

$$F = \frac{2 \times P \times R}{P + R}$$

(1)

(2)结果分析。

利用网络爬虫分别从汽车之家论坛、天涯社区论坛、新浪论坛、猫扑社区、网易论坛各抓取 100 个页面,共计 500 个页面。文中算法将每个页面基于信息块的子树,从每个信息块中提取出“作者”、“正文”、“时间”、“其他”。“其他”为链接或按钮等非文本信息。测试结果如表 1 所示。

表1 舆情信息抽取结果

类别	R	P	F
正文	0.992	1	0.996
作者	0.986	0.995	0.990
时间	1	1	1
其他	1	0.994	0.997

经过测试可以看出, R 、 P 、 F 指数都较高,可以比较满意地提取出所需信息的内容,抽取效果较好。

4 结束语

网络预期信息抽取是网络舆情系统中最重要的部分,是进行后续的舆情分析、舆情统计等的基础。文中采用面向舆情话题的信息提取方法,将话题线索转换

为对文档的 DOM 树的广度优先搜索,并采取设置重复子树阈值 θ 、去重去噪等方法以实现舆情信息的提取。在提取系统设计中,采用了基于用户制定格式,即标注提取方式。实验结果表明,召回率、正确率都较高,可以较为满意地提取舆情信息内容。

参考文献:

[1] 王 权,施韶亭. Web 信息抽取技术在统一检索系统中的应用研究[J]. 计算机应用与软件,2010,27(10):120-122.

[2] 王全民,王 莉,曹建奇. 基于评论挖掘的改进的协同过滤推荐算法[J]. 计算机技术与发展,2015,25(10):24-28.

[3] 姬 鑫,钟 诚. 基于分块的新闻网页信息抽取算法[J]. 计算机应用与软件,2015,32(4):317-322.

[4] 张 昕,鄂海红,宋美娜,等. 基于视觉特征的就业信息页面抽取方法[J]. 软件,2014,35(9):16-20.

[5] 张 奇,郝志峰,温 雯,等. 基于互信息度量的 Web 信息抽取[J]. 计算机应用与软件,2013,30(12):15-18.

[6] 吴 秦,胡丽娟,梁久祯. 基于分块重要度和二维条件随机场的 Web 信息抽取[J]. 南京大学学报:自然科学版,2014,50(1):79-86.

[7] 王志华,魏 斌,李占波,等. 基于本体的 Web 信息抽取系统[J]. 计算机工程与设计,2012,33(7):2634-2639.

[8] Madhavan J, Ko D, Kot L, et al. Google's deep web crawl[J]. Proceedings of the VLDB Endowment,2008,1(2):1241-1252.

[9] Stevanovic D, An Aijun, Vljajic N. Feature evaluation for Web crawler detection with data mining techniques[J]. Expert Systems with Applications,2012,39(10):8707-8717.

[10] 顾韵华,高 原,高 宝,等. 基于模板和领域本体的 Deep Web 信息抽取研究[J]. 计算机工程与设计,2014,35(1):327-332.

[11] Liu X, Gong D. A comparative study of a-star algorithms for search and rescue in perfect maze[C]//Proc of ICECICE. [s. l.]:IEEE,2011:24-27.

[12] 丁艳辉,李庆忠,董永权,等. 基于集成学习和二维关联边条件随机场的 Web 数据语义标注方法[J]. 计算机学报,2010,33(2):267-278.

[13] Cali A, Martinenghi D. Querying the deep web[C]//Proceedings of the 13th international conference on extending database technology. [s. l.]:[s. n.],2010:724-727.

[14] 赵 涛,张太红,陈燕红. 中文农业网页去重及相似度判断研究[J]. 计算机技术与发展,2015,25(1):191-194.

[15] 房 勇,李银胜. 基于 DOM 状态转换的隐网页信息抽取算法[J]. 计算机应用与软件,2015,32(9):17-21.