基于知乎的热点话题信息抽取方法研究

武丙帅
（成都信息工程大学 软件工程系，四川 成都 610225）

摘要：互联网的高速发展，导致网上应受限制的数据信息与日俱增，这给数据工作者带来了巨大的挑战和压力。为了响应国家加大对于网络舆情的监督号召，文章通过分析研究热点话题的信息抽取方法，并将这些方法应用在知乎上，以知乎为基础分析热点话题的信息抽取，以此推广到其他的网站。为网络舆情的监督提供支持。

关键词：知乎；热点话题；信息抽取
中图分类号：TP393.094 文献标识码：A 文章编号：1673-1131（2015）12-0022-02

0 引言

知乎是以分享彼此的专业知识和经验见解为理念，以实名制为基础的网络问答社区。在 2013 年 3 月开放公众注册，在两年时间的时间内注册用户数量迅速攀升。社区氛围友好与理性，连接各行各业的精英。为互联网的用户提供了大量的高质量信息。

知乎主要的产品服务包括：①首页页面。首页页面可以分为两个板块，主要是最新动态和当前热点话题的精选内容，当前热点话题占到版面的 70%，在这一板块中，用户可以关注话题下问题点击查看，并能发表评论，分享话题到其他媒体，收藏话题，举报等。②话题页。话题页面主要包括两个板块，已关注的话题动态和其他人关注的话题动态，在其他人关注的话题上也能进行关注。③发现页，发现页面主要分为：编辑推荐，今日最热和本月最热，热门圆桌，热门收场等版块。其中占据版面最大的就是今日最热，在这个板块可以查看最新的热门话题。④消息页：消息页是关于会员账号收到发出的相关消息的记录。在网页主页的右上角有会员提问的快捷按钮。以上是知乎提供给用户的各项产品[1]。随着智能手机的发展和自媒体的兴起，知乎的移动终端客户端也已上线，在移动终端上下载相应的 App(Application)，也能体验知乎的相关服务。本文主要以网页知乎来作为研究对象。

1 热点话题特点

热点话题的出现是随着互联网的社区、论坛、BBS，以及最新的自媒体发展起来的。而热点话题的研究也是基于有着庞大粉丝群的用户群体来展开的，主要是为了针对突发事件或某些普遍现象。热点话题集中点主要包括突发事件和关系每个人的切身利益的问题热点话题，或者是那些拥有比较大争议的话题。

从这些话题来看其特点主要有五个方面：①话题内的主

本系统中要检测的总共有五道脉冲信号，其中编码器的最大脉冲频率为 11.22KHz，DF16 光电传感器的最大脉冲频率为 2.244KHz，依据香农定理计算得：

$$T_s \le \frac{\pi}{2\pi \times 11.22} \quad (3)$$

也即数据采集卡的频率满足 f_s：

$$f_s \ge 2 \times 11.22 = 22.44\ \text{KHz} \quad (4)$$

因此，要求采集卡通道采集频率应大于 22.44KHz，且最好有余量。

根据测量要求，本故障诊断系统总共需要 2 个数字量输出、5 个模拟量输入及一个高频脉冲量输出，所以选择的数据采集卡至少应具有 2 个数字输入通道和 5 个模拟输入通道和 1 个计数器输出，还应留有一定的余量，为以后系统升级或扩充做准备。

软件使用的是美国 NI 公司的 LabVIEW，由于这款数据采集卡没有与 LabVIEW 相配套的驱动，所以，只能采用 LabVIEW 调用底层 DLL 的方法实现数据采集，其中最主要的是通过 LabVIEW 中调用函数节点来调用数据采集卡 DLL 中封装的函数来实现数据的采集。

3 DF16 光电传感器故障诊断试验

为验证系统能否达到设计要求，就必须进行 DF16 光电传感器正常诊断试验。试验步骤：接入一个正常的四通道 DF16 光电传感器，然后运行诊断系统，其诊断结果如图 4 所示。

图中的①所指示为波形显示区，显示 DF16 光电传感器的四道脉冲波形，②显示每个通道的各种参数值，其中上面的表格为 DF16 光电传感器的标准值，下表格为测试值。③为故障判断区，分别显示哪个通道的频率、占空比、波峰、波谷及相位差出现问题，故障判断区左侧显示每个通道的判断结果，指示灯变绿说明正常，指示灯显示红色则说明此通道不正常，④区显示故障原因分析及处理方法。从图 4 中可以看出，系统运行于 80km/h 的情况，各项指标均符合要求，所以最终判断所测试的 DF16 无故障，③区中显示“传感器正常”。

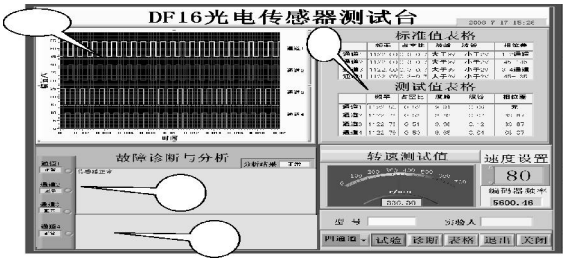


图 4 DF16 光电传感器正常试验界面图

4 结语

本文对 DF16 光电传感器故障诊断试验台的各项功能进行了具体的试验，并对试验中得出的数据进行了详细的记录，并且分析了试验的诊断结果，试验说明了 DF16 光电传感器故障诊断试验台满足用户要求，可以交付使用。随着计算机技术、电子测量技术、信号处理技术、软件技术、以及人工智能技术的发展，传感器故障诊断技术由离线监测到在线监测、由现场诊断到远程诊断是现代化生产和技术发展的必然趋势，传感器故障诊断系统可以进一步往这趋势靠拢。

参考文献：

[1] 罗长洲.一种新型光学编码器[J].光学精密工程,2003,11(l):104-108
[2] 叶盛祥.光电位移精密测量技术[M].四川科学技术出版社,2003

题数,或者主题内的回帖数,都会达到一定的数量,而赞或者反对也是衡量热点话题的一个重要标准。②话题的参与人数在经过一定的时间后会逐渐形成一定的规模。③话题作者的权威程度对于话题的热度也会有很大的影响,一般而言热点话题作者的平均权威度比较高[2]。对于话题作者的权威,一般除了大咖之外,也有各个领域的权威人士。除此之外一些争议性的话题同样会引起大量的关注。④热点话题经常被媒体引用和转载。媒体的参与为热点话题的形成起到了推波助澜的作用,甚至有些被推上头版头条。⑤热点话题帖子回复时间间隔短,但整帖的持续讨论时间长。

基于热点话题的特点,当下对于热点话题的舆情监督不容忽视。如果监督不利,舆论方向错误就会在社会上产生不良的影响。因此对热点话题信息的抽取的研究显得尤为重要。

2 热点话题信息抽取研究现状

目前对于热点信息的研究主要集中在两个方面:一是话题的识别,也就是话题的信息抽取;二是对于话题的持续追踪。截至目前国内外已有大量关于话题信息抽取的研究,但其主要集中于常规的新闻资讯。而对于网络社区、微博、知乎这样的问答社区研究的相对较少。这主要是因为问答社区和新闻资讯之间存在一定的差异,研究的难度会有大幅提升。目前所使用的TDT(Topic Detection and Tracking,话题检测与跟踪)技术并不适用于网络社区的热点话题的研究。主要表现在以下几个方面:

①网络社区主题涵盖内容广泛,表现形式多样。其次网络社区的帖子存在着很强的交互性,主题和主题之间相互的交叉引用。②网络社区话题发起人的不确定性和多元性,发起者是注册者而并非类似于媒体新闻资讯的编辑,社区话题发起者往往不具备对话题舆论的导向能力和责任。③网络社区的管理人员对于社区网络话题的处理态度往往带有个人因素,且不能及时进行正确引导和有效利用。

热点话题的信息抽取是为及时中断或屏蔽不符合当下价值观的热点话题提供技术依据。

3 知乎热点话题的信息抽取方法

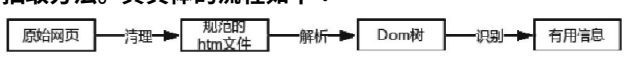
基于人们对知识获取和分享的渴望,知乎的影响力与日俱增,而相应的知乎中热点话题和它的评论对于知乎用户的影响也越来越大。所以急需提出一种适用于知乎热点话题信息抽取的方法,用于监督知乎的热点信息,并对主题舆论进行正面引导。

3.1 基于自然语言的信息抽取

目前使用较多的是基于自然语言处理的信息抽取方法。此种方法适用于以话题类句子为研究对象,首先把话题中的一段话或整个帖子分割成多个句子,然后对于句子的每个词进行标记,然后将预处理好的句子与预先设定的带有关键字的句子进行比对,从而获取话题中的有用信息。知乎帖子相对较长,句子数量比较多时,采用这种方法,会比较快地抽取到话题中的有用信息。提高信息抽取的效率。但是对于在话题中语句太短,像帖子中经常出现"顶""谢谢",这样还不能形成完整的句子的帖子,此种方法就不再适应。

3.2 基于网页结构树的信息抽取

从整个网页的信息出发,通过对网页结构的分析来抽取热点话题所涉及到有用信息,也就是基于网页结构树的信息抽取方法。其具体的流程如下:



在分析之前将网页解析出具有层次结构的页面语法树来。并生成相应的抽取规则,将规则应用到语法树上。从而实现对有用信息的抽取。对于分析出来的不同的语法树结构使用不同的分类抽取规则,可以明显地提高抽取系统的效率。其主要的计算方法为:

$$\mathrm{Re}\,levancy(B_i)=\frac{Linkcount(b_i)}{Contentcount(b_i)}$$

$$Linkcount(B_i)=\sum_{j=1}^{N}Linkcount(b_j)$$

$$Contentcount(B_i)=\sum_{j=1}^{N}Contentcount(b_j)$$

其中,Bi 表示网页语法树中第 i 个分块结点,bj 表示 Bj 的第 j 个子树。Linkcount(b_j)为所有子树的链接之和。Contentcount(b_j) 为所有子树中非结点之和。但是这种方法也有着不足之处,对于有着明显结构特征的网页,信息抽取会比较的容易。但是类似于知乎的网站,网页的结构发生了很大的变化。有很大一部分网页都不适合此种方法来抽取信息[4]。

3.3 基于本体论的信息抽取

互联网的发展也带动了很多门户网站的发展,在门户网站上也会用到信息抽取,以用来分析门户网站的经营状况。门户网站大多采用基于本体的信息抽取方法。此种方法利用网络数据的本身所描述的信息来实现抽取。它对于帖子语句的数量和网页的结构依赖相对较小。但是需要事先构建一个完善的本体库,作为比对的参考,但是上文已经提到知乎的话题具有不确定性和广泛性,因此这种方法并不适用于知乎。

3.4 基于归纳式的信息抽取

基于归纳式信息抽取方法的主要理论是利用归纳式学习方法生成的抽取规则。在网页中标记出所要抽取的数据,然后依据系统在样本上归纳出的规则来进行分析。此种方法的抽取信息的准确度和效率取决于所获得样本的数量和质量[5]。且对于单个语句语法分词的复杂度依赖比较小,因此需要在分析前收集大量的相关话题信息来作为样本。而对于新的话题或者论点,会导致标本收集困难,因此这样的分析方法对于具有相似性的同类话题分析起来相对容易,但对于知乎,也并不能完全适用。

4 结语

针对目前网络的舆论监督形式越来越严峻的情况,为了更好地响应国家加强网络舆论监督的号召。本文结合知乎的实际情况,阐述了几种可以应用到知乎的信息抽取方法,每种方法都有各自的优缺点和适用的范围,目前没有完全适用所有环境的信息抽取方法,当下可以针对不同话题的特点和发展情况来选择不同的分析方法,以此达到最好的分析效果。

参考文献:

[1] 齐海风.网络舆情热点发现与事件监控技术研究[D].哈尔滨工业大学,2014(9)45-47
[2] 王娜.Web 文本挖掘技术研究[D].兰州理工大学,2013
[3] 胡静.Web 文本挖掘中数据预处理技术研究[J].现代计算机,2009(2):48-50
[4] 邹涛.基于 WWW 的文本信息挖掘[J].情报学报,1999,19(4):291-295
[5] 孙黎明.基于 BBS 的社会热点话题识别与跟踪算法研究[D].华南理工大学,2014

作者简介:武丙帅(1989-),男,硕士在读,研究方向为数据库与知识工程。