

基于褒贬值的媒体社会责任监测

浙江传媒学院受众满意度研究团队^①

用户对媒体产品和明星的接受和认可程度直接体现出媒体和明星的社会效益和经济效益。对经济效益的评价,国内外主要通过收视率、发行量、票房、点击率、百度指数、微博阅读人数等用户行为指标进行评价,但对社会效益的评价却没有公认的评价指标。2015年党中央和国务院在《关于推动国有文化企业把社会效益放在首位、实现社会效益和经济效益相统一的指导意见》中,明确要求文化企业社会效益指标考核权重应占50%以上,形成对社会效益的可量化、可核查要求。但谁来评价媒体的社会效果、如何评价、指标的信度效度等问题,一直困扰学界(刘燕南,2015)^[1]和业界(田进,2016)^[2]。现有评价体系无法满足对媒体社会责任进行量化考核和规范的需要。因此,媒体产品和明星的社会效益评价成为迫切需要解决的问题。

浙江传媒学院管理学院“视频节目受众满意度研究”团队长期关注媒体的用户效果测评,与上海易步信息技术有限公司共同研发了“浙江传媒学院视频监测与分析系统”。研究团队通过该系统能够自动抓取网络新闻、微博、论坛、弹幕、微信、视频网站中的评论等大量数据,并进行语义分析,计算出网民对于娱乐视频和明星的褒贬值(-5到+5)。

一、系统概况

“浙江传媒学院视频监测与分析系统”是应用大数据和人工智能技术,收集用户的评论信息,并进行语义分析,计算出褒贬值,对娱乐视频(电视剧、综艺、电影、明星等)进行态度测评的系统。该系统是2011年浙江省财政资助项目“媒介测评与分析实验室”的主要成果,并于2012年8月开始运行;2015年、2016年受浙江传媒学院资助,系统不断更新,测评工作稳步发展,2016年6月开始监测上星频道黄金时间(19:30—23:00)的电视剧和综艺节目,2017年1月开始监测100位娱乐明星,2017年4月开始监测主要院线上映的电影。图1为系统的运行过程框图。

第一步:设置监测对象。目前是人工设置监测对象,课题组成员完成关键词的录入。随着机器学习技术的发展,系统计划拓展关键词计算机自动收集和设置功能。

第二步:计算机自动收集数据。计算机自动抓取与监测对象相关的新闻、论坛、微博、弹幕、微信公众号、视频网站评价等信息,形成大数据。

第三步:褒贬值计算。基于娱乐褒贬词库,对收集到的信息,利用语义分析技术,计算出每一条新闻、论坛、微博、弹幕、微信公众号、视频网站评价的褒贬值,褒贬值在正5到负5之间。

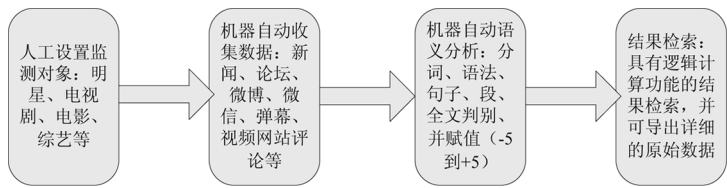


图1 流程简图

第四步:数据输出。根据收集和计算的数据,系统生成结果。一是简单的统计图表,二是原始数据,三是月度汇总数据,以供进一步分析使用。

^① 团队成员:葛进平、许蔚蔚、龚立群、陈杰、章霖轩、楼彩霞、张鹏、明旭。通讯作者葛进平,男,研究员,浙江传媒学院管理学院,邮箱:1105908203@qq.com。本团队受浙江传媒学院管理学院、浙江传媒学院协同创新中心资助。

二、系统功能

“浙江传媒学院视频监测与分析系统”由热点趋势、褒贬分析、数据导出、后台管理、月度汇总、在线褒贬打分等六个功能模块组成。

(一) 热点趋势

热点趋势是监测对象行为指标的汇集。图2是湖南卫视《真正男子汉第2季》2016年11月20日到12月20日的热点趋势。系统可以显示“10天”、“1个月”、“3个月”、“今年”、“全部”等不同时间段,也能任意设定开始和结束时间,每天的信息量和折线;显示鼠标悬停所在日期的新闻、论坛、微博、弹幕、微信、视频网站下评论的数量和总篇数;能按 excel 数据文件格式导出每天的具体数据。

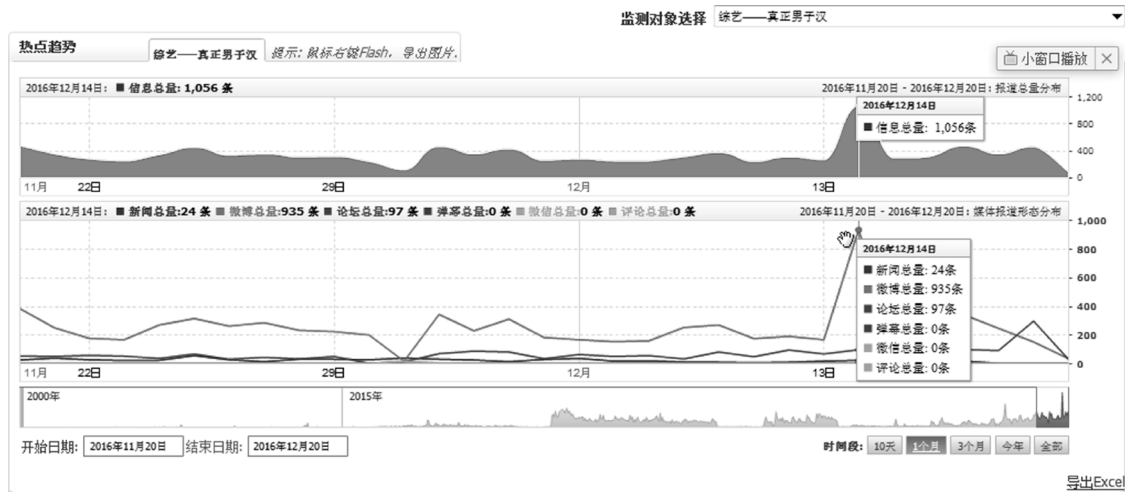


图2 《真正男子汉第2季》热点趋势(2016年12月20日截图)

图2显示2016年11月20日到12月20日,每天与《真正男子汉第2季》相关信息数量变化不大,为300条左右,12月14日高达1056条,研究者有兴趣可以单独分析。

系统提供信息来源结构饼图,显示24小时10个实时热词等功能。

(二) 褒贬分析

褒贬分析是系统最具特色的功能,汇集监测对象的态度指标,实现对监测对象满意度的分析。满意度分析以褒贬值为主,踩顶数为辅。褒贬值基于语义分析技术,对收集到的新闻、论坛、微博、弹幕等数据进行评分,分值在正负5之间,正数、零、负数分别对应正面、中性、负面评价。

图3是浙江卫视和安徽卫视播出的电视剧《美人私房菜之玉蝶传奇》2016年11月20日到12月20日的的数据。系统可以显示“10天”、“1个月”、“3个月”、“今年”、“全部”等不同时间段,也能任意设定开始和结束时间,每天正面、负面、中性信息量和折线;显示鼠标悬停所在日期的正面、负面、中性评价的数量;下部分的折线则显示每天具体的褒贬值。

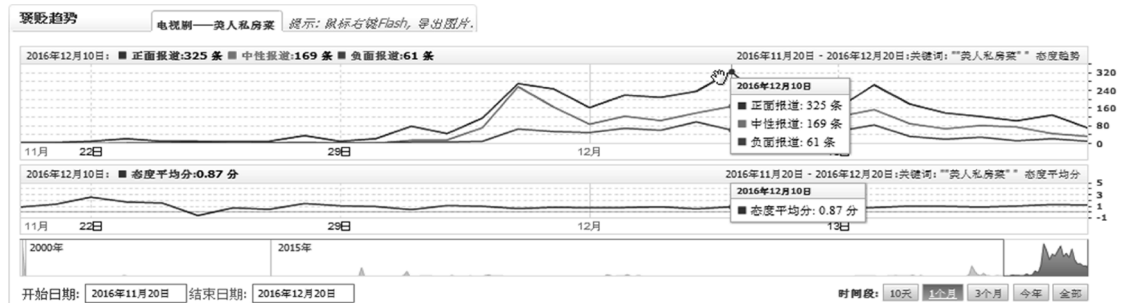


图3 《美人私房菜之玉蝶传奇》褒贬分析(2016年12月20日截图)

图 3 显示,《美人私房菜之玉蝶传奇》的信息量在 12 月初开始增加,10 日的信息量最大,正面、中性、负面的信息分别为 325、169、61 条,平均褒贬值为 0.81。

(三) 数据导出

数据导出是本系统的重要部分,主要提供用于进一步研究的原始数据,并提供简单的分析。

图 4 是数据导出的条件设置和部分检索结果。设置条件是 2016 年 11 月 1 日到 12 月 19 日,东方卫视《金星秀》的论坛评论数据,共有 938 条,导出的 excel 数据文件包括标题、来源、来源类别、态度值、阅读数、回复数、转发数、评论数、发布日期等内容。



图 4 统计报告检索结果

点击标题,可链接到具体的网址,看到具体的内容。

数据导出的设置灵活,可以选择不同的“来源类别”、“态度”和“时间范围”。

数据导出可以多关键词检索,用“A B”(AB 之间为空格)表示“A”和“B”同时存在(和运算),用“A-B”表示有“A”但没有“B”(非运算),用“A|B”表示有“A”或有“B”(或运算)。需要特别注意的是,检索的第一个关键词需要通过后台管理预先设定。否则监测的结果不全,检索的结果当然不全。

(四) 后台管理

后台管理主要用于设置监测的对象,一般用户不能使用。后台管理有“节目文本监测管理”、“节目视频监测管理”和“广告过滤模块”三个部分。图 5 是文本内容关键词设定的截图。



图 5 被监测对象的设置

“节目文本监测管理”是设置和管理被监测的文本。“设置”是创建新的监测对象，其中“关键词名称”是被监测内容的称谓，“关键词”是被监测内容的关键词，可设置多个关键词，其逻辑关系与上面介绍的数据导出的关键词设定一样。系统对“一般”、“重点”、“研究”三个等级的监测对象，抓取的信息依次增加。图5是对明星“鹿晗”的监测设置，系统自动收集含有“鹿晗”或“鹿哥”的信息，监测等级是“重点”，监测时间是2016年12月14日到2017年底。

“管理”是对被监测的文本的编辑，包括查询、编辑、关闭、删除、导出被监测对象数据等功能。设置组负责监测对象的设置。监测对象的设置是系统最为基础的工作，是一个不能间断的长期工作，有两个内容，一是在系统中设计监测对象的名称、关键词、类型、等级、时间等信息，如图5，重播的节目不重复出现，编辑其信息即可。二是定期提供监测对象清单的 excel 文件，包括监测对象的名称、关键词、播出平台、分类等信息；重播的节目不改变原来的信息，新增一条记录。

“节目视频监测管理”是设置和管理被监测视频，仅需要设置节目的名称。“广告过滤模块”是为了提高系统监测的准确率，过滤无用信息，由系统管理员设定，管理员不用设定。这里不介绍这两个功能。

（五）月度汇总

月度汇总生成综艺、电视剧、明星、电影等类型的月度汇总数据的 excel 文件。具体的汇总原则是，以汇总“2016年12月”为例，对于12月开始监测、12月份未结束的监测对象，则汇总“开始时间”到“2016-06-30”的数据；对于12月以前开始监测，12月份未结束监测的对象，则统计“2016-12-1”到“2016-06-30”；对于12月以前开始监测，12月份结束监测的对象，则统计“2016-12-1”到“结束日期”。

表1是2016年12月汇总文件的部分数据。汇总所有数据来源的数据条数和褒贬值，褒贬值为简单算术平均数，研究者可以进一步进行各种权重计算，得出不同的褒贬值。

表1 2016年12月电视剧汇总数据（2016年12月20日生成）

| 名 称 | 开始 时间 | 结束 时间 | 新闻 | | 论坛 | | 微博 | | 弹幕 | | 合计 | |
|-----------|----------|----------|-----|------|------|------|------|------|-------|-------|-------|------|
| | | | 条数 | 褒贬值 | 条数 | 褒贬值 | 条数 | 褒贬值 | 条数 | 褒贬值 | 条数 | 褒贬值 |
| 胭脂 | 12/1 | 12/31 | 92 | 1.34 | 571 | 1.65 | 2561 | 0.38 | 21027 | -0.08 | 24251 | 0.02 |
| 三国 | 12/1 | 12/24 | 142 | 1.5 | 3742 | 1.46 | 5078 | 0.57 | 7347 | 0 | 16309 | 0.52 |
| 美人私房菜 | 12/3 | 12/31 | 165 | 0.45 | 854 | 0.87 | 5207 | 0.82 | 0 | 0 | 6226 | 0.81 |
| 麻雀 | 12/1 | 12/31 | 198 | 1.15 | 671 | 1.83 | 4005 | 0.52 | 0 | 0 | 4874 | 0.73 |
| 兰陵王妃 | 12/1 | 12/31 | 50 | 0.75 | 972 | 0.64 | 2624 | 0.6 | 0 | 0 | 3646 | 0.61 |
| 咱们相爱吧 | 12/1 | 12/21 | 231 | 0.4 | 549 | 0.93 | 2574 | 0.63 | 0 | 0 | 3354 | 0.66 |
| 北上广依然相信爱情 | 12/12 | 12/31 | 64 | 0.14 | 359 | 0.57 | 1679 | 1.48 | 0 | 0 | 2102 | 1.29 |

（六）在线褒贬打分

在线褒贬打分是一个对非系统抓取的数据进行语义分析的接口，研究者上传需要评价的文本文件，系统自动计算出文本的褒贬值。这是无对象测评，需要对褒贬值的计算结果进一步分析。

三、系统数据收集

“浙江传媒学院视频监测与分析系统”就是利用大数据和人工智能的理论和 技术，评价视频节目和明星的社会效益和社会责任，为媒体的发展提供坚实的数据支撑。

（一）监测对象

系统的负载决定监测内容的数据。我们研究的领域是娱乐内容，目前监测的内容主要有两个方面。

一是上星频道黄金时间播出的电视剧和综艺节目。监测的时间是节目播出前 7 天至节目播出后 7 天。黄金时间是指 19：30—23：00 之间播出的电视剧和综艺节目。2016 年 10 月 18 日，考虑系统容量，重播综艺节目的播出不稳定等原因，暂停监测重播综艺节目。

二是明星。考虑到系统容量，第一批监测对象设定为 100 人，2016 年 12 月 20 日开始监测。100 人的确定依据是来源艺恩泛娱乐指数排行榜、《第一财经周刊》品牌价值明星榜、新浪微博势力月榜。2016 年艺恩泛娱乐指数排行榜的权重为 0.4，2015《第一财经周刊》品牌价值明星榜的权重为 0.2，2016《第一财经周刊》品牌价值明星榜的权重为 0.3，2016 年 11 月新浪微博势力榜的权重为 0.1。学生团队提高了 55 人有昵称，如杨颖、杨幂、吴亦凡的昵称分别为 Angelababy、幂幂、凡先生。收集数据时，昵称和明星的姓名一样作为抓取数据的关键词。

本研究团队计划 2017 年上半年开始监测院线上映的主流热门电影，2017 年下半年监测在网络上首播的综艺和网剧。

（二）数据来源

根据设置的监测对象的关键词，利用爬虫技术，从互联网上收集监测对象的信息。目前主要的数据来源有四个方面：一是网络新闻。网络新闻主要是传统媒体内容在互联网上的再现，获得的信息比较全面。二是论坛。收集论坛和贴吧上信息，获得的信息也比较全面。三是微博。收集新浪微博的信息，获得的信息不全面，采用随机技术取得 20% 左右的信息，视为随机数据，可以代表总体。四是弹幕。收集视频网站上弹幕的信息。目前能收集优酷土豆、腾讯和搜狐视频网站的弹幕抓取，华数和芒果 TV 网站没有弹幕。爱奇艺的弹幕加密不能抓取。

本研究团队计划 2017 年增加对微信公众号、视频网站上的评论的数据收集。

（三）数据处理

平权原则：一条新闻、论坛、微博、弹幕、微信均代表 1 个人的观点或态度，其权重不应该有差异。因此褒贬值计算简单算术平均。

去零原则：由于中性评价较多，褒贬度为 0 的不计算（后续需进一步讨论）。分子不变，分母变小，平均值会增高。

新闻褒贬值=新闻褒贬值之和/新闻条数

论坛褒贬值=论坛褒贬值之和/论坛条数

微博褒贬值=微博褒贬值之和/微博条数

弹幕褒贬值=弹幕褒贬值之和/弹幕条数

褒贬值=（新闻褒贬值*新闻条数+论坛褒贬值*论坛条数+微博褒贬值*微博条数+弹幕褒贬值*弹幕条数）/（新闻条数+论坛条数+微博条数+弹幕条数）

四、系统褒贬值计算

无论是幕后还是前台，人工智能不断融入娱乐产业。2016 年 11 月 3 日，VOGUE 11 周年庆典上，明星大秀衣服，借助 Watson 的力量，李宇春穿上“认知裙子”。这款时尚的“认知裙子”从设计到穿在身上，只用了一个星期。^[3]2017 年 1 月江苏卫视的《最强大脑》第 4 季开播，“人机大战”博得网友的热烈关注，百度机器人“小度”在前 3 集以 2 胜 1 平，直接进入最终脑王的角逐。

1. 人工整理创建褒贬词库、程度词库、否定词库。研究团队对所有词汇（每年一次）和新增的词（每月一次）进行褒贬值评分，褒贬程度由低到高依次为-5、-4、-3、-2、-1、0、1、2、3、4、5 等

11个分值。截止到2016年11月,娱乐领域的褒贬词库有21303条。

2. 对收集到的信息首先进行分句。

3. 对每一句话进行分词,得到词汇的相互修饰关系。根据中文语法和词汇修饰关系计算该句的褒贬分值:以褒贬词的值为基准,以程度词为权重,以否定词进行词性反转,综合计算得到该句的褒贬分值。

4. 将所有句子的褒贬分值进行加权、归一化处理,得到该文本的褒贬分值,并使得其在 $[-5, +5]$ 区间。

5. 以上步骤没有考虑特定对象,如果要进行特定对象的褒贬分析,则以特定对象为关键词对句子进行过滤,只计算包含特定对象关键词的句子的褒贬分值。

五、数据清理与整理

“浙江传媒学院视频监测与分析系统”以为教学和研究服务为目的,使用者需要具备数据整理和分析能力。系统主要通过“数据导出”提供最原始的数据,通过“月度汇总”提供大类监测对象月度汇总数据。

(一) 月度数据

该数据包含播出频道、节目名称及分类、播出时间(来自节目设置文件),监测的天数、新闻、论坛、微博、弹幕、微信、网站评价的条数和褒贬值(来自节目监测文件)。月度数据采用的原则为:每天新闻、微博、论坛、弹幕、微信、视频网站评论的信息量之和大于等于15条,或者每月新闻、微博、论坛、弹幕、微信、视频网站评论的信息量之和大于等于400条;每月监测的天数大于等于5天。

(二) 多月合并数据

每月数据基础上,合并相同的监测对象,监测天数和监测条数相加,褒贬值按照监测条数加权计算。以电视剧为例,2016年共生成13个数据文件:7个单月度数据文件(6月、7月、8月、9月、10月、11月、12月);6个多月度数据文件(6-7月、6-8月、6-9月、6-10月、6-11月、6-12月)。

根据本系统所生成的娱乐产品和明星褒贬值月榜,是对社会责任的重要评价指标。学报将连续发布。娱乐产品和明星褒贬值月榜作为刚出生的婴儿,期盼能备受关注并快速成长。

参考文献:

- [1] 刘燕南,张渤,刘双.电视评估体系的顶层设计与基层实践——来自电视台一线的报告及思考[J].现代传播,2015(8):1-6.
- [2] CTR洞察.田进:全面深入开展广播电视节目综合评价工作[EB/OL].<http://chuansong.me/n/1368121451063>,2016-12-21.
- [3] 吴晓波频道.人工智能不止是“下棋”[EB/OL].http://www.iqiyi.com/v_19rra8m5zg.html?vf_m=f_191_360y,2016-12-07.

[责任编辑:华晓红]