基于互联网的数字媒体内容舆情分析系统①

王晓艳，梁晋春，郭晓霞，姚颖颖，汪 瑛

(国家广播电影电视总局 广播科学研究院信息技术研究所，北京 100045)

摘 要：快速发展的互联网已成为反应社会舆情的重要载体之一，如何对互联网数字媒体内容舆情进行分析监控，及时有效地发现危害社会安全的各类有害虚假的舆情信息，已成为促进我国数字媒体内容舆情信息安全和内容监管健康发展迫切需要解决的问题。提出了一个基于互联网的数字媒体内容舆情分析模型，设计并实现了基于互联网的数字媒体内容舆情分析系统，并就涉及的实用关键技术进行了探讨。

关键词：数字媒体内容；基于互联网；舆情分析；文字信息抽取；网页信息抽取；关键词提取；热点动态检测；文本倾向性分析

Public Opinion Analysis System for WWW-Based Digital Media Content

WANG Xiao-Yan, LIANG Jin-Chun, GUO Xiao-Xia, YAO Ying-Ying, WANG Ying

(Film & Television Academy of Broadcasting Science Information Technology Research Institute, State Administration of Radio, Beijing 100045, China)

Abstract: The rapid developing Internet has become an important reflection of public opinions. It has been an urgent problem to be solved how to analyze and monitor public opinions in digital media and on Internet, and how to timely and effectively spot all kinds false information harmful to social security, in order to promote the healthy development of information security and content supervision. This paper proposes a www-based digital media content public opinion analysis model. A digital media content public opinion analysis system based on the Internet is designed and realized, and the practical key technologies are discussed.

Key words: digital media content; www-based; public opinion analysis; characters information extraction; web information extraction; keyword extraction; hotspot's dynamic detection; text opinion analysis

1 引言

文化是国家和民族的灵魂。当今世界，文化与经济、政治相互交融，与科技的结合日益紧密。发展数字媒体内容服务，建设现代文化市场体系，发展新兴文化业态，大力推动文化产业升级，已成为我国文化事业和文化产业发展的必然选择。近年来，以数字电视、宽带网络、视频娱乐、网上生活消费等融合业务为特色的数字媒体业务与数字生活应用正在成为市场需求的热点，数字化、网络化、融合化已成广播电视事业发展的必然趋势。这种趋势必然造成数字内容信息（即舆情信息）的开放性和海量化，数字媒体内容更为多元化，传播对象更为广泛化，传播途径更为便捷化，传播终端更为多样化。在这种趋势下，社会上的每个人都可以成为信息的生产者和传播者，进而成为舆论的制造者，舆论的制造和传播将会变得更加直接。为了促进和保证数字媒体内容相关事业和谐、健康发展，面对如此海量的数字媒体内容，开展针对多种数字媒体内容的监管技术研究尤为重要。数字媒体内容舆情分析技术研究是数字媒体内容监管技术的基础性技术研究，舆情分析的技术手段与舆论的传播渠道有密切的关系，近年来随着互联网的快速发展，互联网上丰富、海量的数字媒体内容信息已经成为人们

① 基金项目:财政部中央级公益性科研院所基本科研业务费专项
收稿时间:2011-04-15;收到修改稿时间:2011-06-16

获取信息的重要来源，如何对互联网数字媒体内容舆情进行分析监控，及时有效地发现危害社会安全的各类有害虚假的舆情信息，已成为促进我国数字媒体内容舆情信息安全和内容监管健康发展迫切需要解决的问题。

针对互联网海量数字媒体内容管理的监管需求，我们提出了一个基于互联网的数字媒体内容舆情分析模型，并就‘网站发布违规药品广告’这一舆情需求设计并实现了基于互联网的数字媒体内容舆情分析系统。

2 模型设计

2.1 分析模型

基于互联网的数字媒体内容舆情分析模型主要由舆情关键词规划、舆情信息采集、舆情分析、舆情发布四部分组成。本系统建立的基于互联网的数字媒体内容舆情分析模型如图 1 所示。

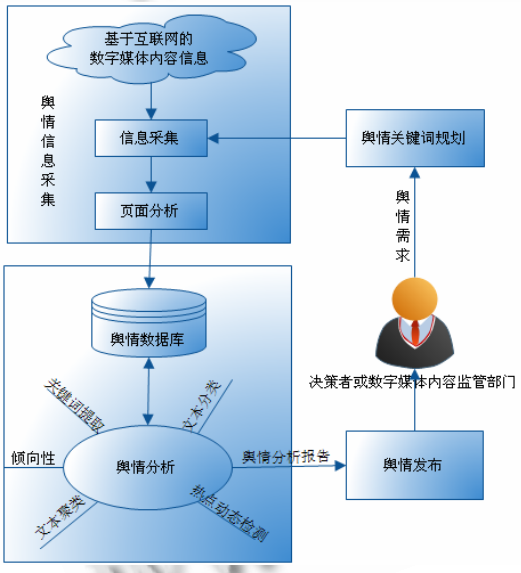

图 1　基于互联网的数字媒体内容舆情分析模型图

其中，(1) 舆情关键词规划模块是指决策者或者数字媒体内容监管部门根据自身舆情需求选择合适的舆情关键词，并确定舆情信息采集任务。

(2) 舆情信息采集模块是指根据确定的舆情关键词，从互联网自动采集相关舆情信息，并将主题相关的信息存入数据库。

(3) 舆情分析模块是指对采集的舆情信息进行热点检测、主题追踪、倾向性分析等工作，生成舆情分析报告。

(4) 舆情发布模块是指将生成的舆情分析报告提供给决策者或数字媒体内容监管部门，为决策和管理提供支持。

2.2 评价模型

针对‘网站发布违规药品广告’这一舆情需求，又设计了基于互联网的数字媒体内容舆情评价模型，如图 2 所示。

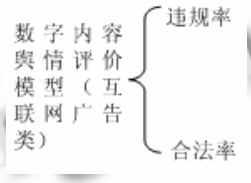

图 2　基于互联网的数字媒体内容舆情评价模型

其中，违规率=违规条数/广告总条数，合法率=合法条数/广告总条数。

3 模型实现

3.1 系统描述

基于互联网的数字媒体内容舆情分析系统根据上述舆情分析模型和评价模型，利用网页舆情信息提取、文字信息提取、关键词提取、热点动态检测与分析、文本倾向性分析等关键技术对互联网上的数字媒体内容进行分析，提取数字媒体的语义信息，对用户感兴趣的特定内容进行动态发现与跟踪，为掌握互联网数字媒体内容舆情提供有效的分析依据。

3.2 系统流程

根据基于互联网的数字媒体内容舆情分析模型，我们针对‘网站发布违规药品广告’这一舆情需求，设计并实现了基于互联网的数字媒体内容舆情分析系统，并实现了对图 2 中违规率和合法率两个指标的分析与评估。系统工作流程如图 3 所示。

该流程包括舆情数据源、系统分析、应用三个部分，其中，

(1) 舆情数据源部分是由互联网上各种不同格式类型的音视频文件库构成，相关信息记入数据库，原始内容存放到磁盘阵列。本系统数据源部分来自互联网药品网站中的部分广告类内容。其中，采用相关音视频特征抽取技术对原始数据进行处理，将数据存入数字媒体内容元数据库。数字媒体内容元数据、音视频样本库和敏感知识库（敏感词和敏感图像）组成一

起，向数字媒体内容舆情分析工作提供知识信息和特征信息。

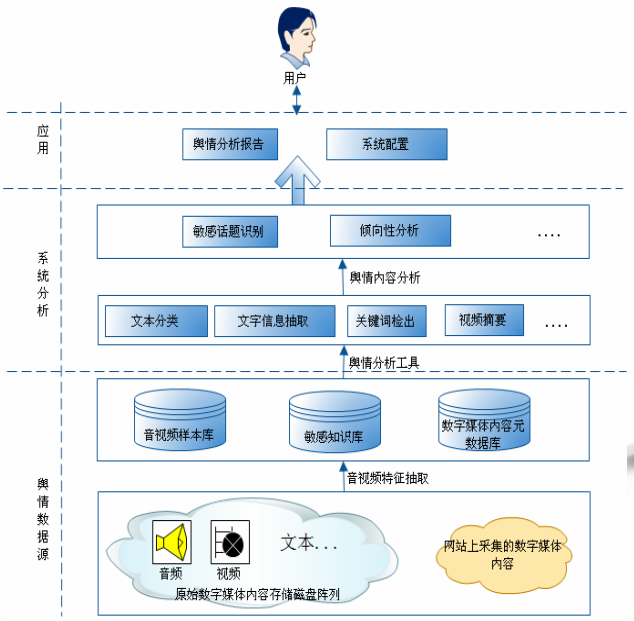

图 3 基于互联网的数字媒体内容舆情分析系统工作流程图

（2）系统分析部分是指对数据库中的数据进行处理分析。根据数字媒体内容舆情分析的不同需求，采用文本分类、文字信息抽取、关键词检出、视频摘要等舆情分析工具，进行敏感话题识别、倾向性分析等工作，并生成舆情分析结果。针对本系统研究的舆情需求，系统可以分析出广告是否合法广告、疑似广告以及违规广告，为下一步的舆情评估提供基础数据。

（3）应用部分实现对基于互联网的数字媒体内容舆情分析系统的可视化操作和对舆情分析结果的可视化展示。用户可对该系统进行系统配置，舆情分析结果可从不同角度可视化展示给用户。

3.3 系统架构

基于互联网的数字媒体内容舆情分析系统分为两个子系统，分别为网络数据采集子系统和数字媒体内容舆情分析子系统。两个子系统之间通过 FTP 协议以文件形式进行数据交换。整个系统的所有任务均由调度服务器进行调度。网络数据采集子系统支持多个采集节点的分布式采集，所有采集的数据均通过 FTP 上传到存储服务器。数字媒体内容舆情分析子系统是对数字媒体内容进行舆情分析，其处理结构保存到数据库中。用户通过客户端对数据进行操作。系统架构如图 4 所示。

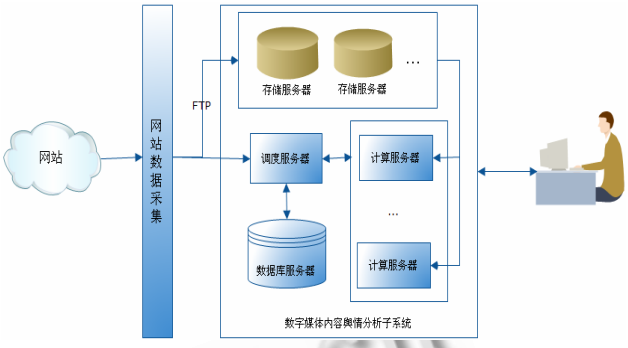

图 4 基于互联网的数字媒体内容舆情分析系统架构图

该系统的软件实现采用 Visual C++6.0，c/s 结构，系统数据库采用 Oracle 10g 数据库。

3.4 系统界面

针对‘网站发布违规药品广告’这一舆情需求，实现的软件系统部分界面如图 5 所示。

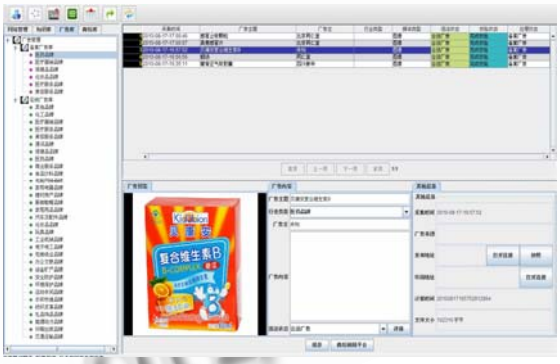

图 5 软件系统部分界面

网站发布违规药品广告情况统计表如表 1 所示。

表 1 网站发布违规药品广告情况统计表

网站发布违规药品广告情况统计表					
网站数量	违规率	合法率	违规条数	合法条数	总发布条数
具体：					
行业类型	违规率	合法率	违规条数	合法条数	总发布条数

4 关键技术应用分析

4.1 网页舆情信息抽取

初步统计，在互联网上，经食品药品监管部门

批准，可向个人消费者提供互联网药品交易服务的网站约有几十家，且网站信息繁杂。如何从医药类终端网页中提取标签内容，就涉及到文本分类技术中的网页文本信息提取问题。网页文本信息提取面临网页格式多样化、写作风格多样化、大量和主题无关的噪音信息、结构信息丰富等问题。目前，在网页分析领域基于 HTML 的结构有大量的开源代码。其中比较有名的开源类库有 NekoHtml，HtmlParser 和 HtmlCleaner 等，经过反复对比试验后，本系统使用了 HtmlCleaner 这个类库来提取网页中某个标签内文本的方法，它的源代码十分精简适于阅读和修改，并且处理速度很快。

此外，在网上药房信息提取方面，本系统使用以下方法提取药品名称、类型、生产企业、批准文号等内容信息。

(1) 设计一个描述药品信息的关键词表，该词表中包含了药品名称、生产企业、批准文号、功能主治等词语，这些词是药品描述页中描述药品信息的常用短语。同一信息可能采用不同短语进行描述，所以将这种不同短语按近义词处理，例如生产企业和厂商就是一对近义词。

(2) 对该药品描述页扫描 form、div、table 等标签，如果某个标签内出现了两个以上的关键词表中的非近义词，将其文本内容块进行返回。

4.2 文字信息抽取

在采集的药品发布终端网页中，包括丰富多样的文字、图像、音频、视频等信息，本系统中，文字信息提取技术主要应用于图像和 flash 的内容分析。

针对图像和 flash，它们除了在终端描述页中有相应的描述内容外，其自身往往还携带着丰富的信息。广告主在制作一个 Flash 或图片类型的广告时，往往在其上包含了自身的商标和广告语。利用商标检索方法可以检测出该 Flash 或图片中是否含有本地所建商标库中的商标。利用文字识别方法(OCR)可以检测出该广告中所包含的厂家、产品名称以及功效等重要信息。所建商标库中的商标和各个广告主是相对应的，检测出了商标，自然也就可以判定出广告主的名称。OCR 和商标检测方法提取出的内容可以和网页中提取出的元数据内容一起用来提取广告主的名称、广告内容等信息。

此外，利用文字信息提取技术对图像进行处理，得到图像中的文字信息。图像的内容分析还包括颜色直方图特征提取、特定目标识别等其他处理。

对于 Flash，首先可以把 Flash 分解成 Shape 关键帧，然后利用图像内容分析的方法对 Shape 关键帧进行内容分析，最后综合利用全部 Shape 关键帧的信息得到 Flash 的内容分析结果。

由于从视频和图像中提取的文字信息不可能百分之百正确，所以在必要时系统还要以人工的方式对文字提取的结果进行校正，对漏提取的文字信息进行标注。

4.3 关键词提取

关键词提取技术就是自动从文档或文档集合中摘取精要或要点，其目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的内容描述。人们希望从海量文本中快速准确地获得自己感兴趣的内容，这是信息检索领域目前迫切需要解决的问题。然而现在的信息检索系统只能提供给用户检索到的文档全文，因此，人们提出了通过关键词和摘要为用户提供简明扼要的内容描述。关键词是简要描述一篇文档内容的重要元数据，用户可以通过关键词迅速了解文档的内容，从而判断文档是否是自己感兴趣的话题。在多文档的自动关键词提取方面，本系统在比较多种关键词提取算法后，最终采用参考文献[6]中所介绍的自动关键词提取方法，该方法结合词性规则和统计信息，有效利用多个文档所反映的全局性的重要信息，同时又尽可能的过滤掉信息冗余，该方法首先基于词性标注结果找出名词、动词、名词短语作为候选关键词；然后过滤部分候选关键词，并计算其余候选关键词的权重；最后根据用户所需个数给出最终关键词列表。

4.4 热点动态检测与分析

舆情热点检测技术是指从不断涌现的网上舆情中及时地获得新发生的热点信息，并对其进行持续追踪，主题检测与追踪技术是解决这一问题的基础。主题检测就是从数字媒体信息流中自动检测出各个主题将相应内容划归到相应的主题，并且能够实时地针对新到的数字媒体内容检测新的主题，主题检测技术可以取代人工完成自动专题生成、热点新闻生成等任务，在本系统中，我们采用参考文献[6]中所提出的互联网舆情热点的动态检测算法，用于解决舆情热点的自动发现。该算法在综合考虑舆情热点的特征和人们认知规

律的基础上，引入主题排序、主题合并和调整、报道淘汰以及主题描述等步骤。实验表明，在实际应用场景中，明显提高了舆情热点的检测效果。该方法在主题排序方面引入在某一时刻对主题计算得分值的机制，该机制综合考虑主题内文档的时间特性和数量特性，进而在某一时刻为每个主题给出一个较合理的得分值用于主题排序。在主题相似性方面，引入主题合并和调整的机制，用于克服同一个主题被误分为多个小主题的现象，每处理固定个数的报道，就对主题两两之间进行比较，若依据比较策略发现两主题相似度较高，则进行主题的合并和调整。在主题描述问题，采用将特征词和报道标题相结合的方法，用于克服两者单独使用的缺陷。首先，选择主题内部权重最高的若干个特征词作为主题描述的一部分；同时，根据报道选择策略，选取该主题内最具代表性的若干篇报道的标题作为主题描述的一部分。此外，引入主题内报道淘汰的机制，用于克服主题内容过于宽泛的现象，每处理固定个数的报道，就对各主题内的报道按照时间和相似度规则进行淘汰。

在本系统中，采用了这种互联网舆情热点的动态检测算法对医药类广告进行热点检测，记录广告发布网站，发布信息，用户可以查看不同时段内的热点信息．比如“最近 24 小时”、“最近周”等。并在热点自动发现任务的基础上，对自动发现的热点进行深入分析，从多方面、多角度综合分析和展现当前的舆情热点。针对‘网站发布违规药品广告’这一舆情需求，本系统研究关注的热点话题就是网站发布的药品广告是否违规，并从采集广告量、违规率、合法率、行业类型等角度进行分类统计与分析，还可从网站分布、时间分布等方面对热点进行进一步分析。

4.5 文本倾向性分析

近年来随着互联网的快速发展，互联网上丰富、海量的数字媒体内容信息已经成为人们获取信息的重要来源，如何判断海量的网络数字媒体内容是正面还是负面的，及时有效地发现危害社会安全的各类有害虚假的舆情话题信息，这就需要对这些信息进行倾向性分析。倾向性分析的目的是判断文本的情感类别，即该文本对某一主题是持支持还是反对态度。根据实现的方法可分为基于词的倾向性分析和基于机器学习的倾向性分析[7]。文本倾向性分析中的主要任务有以下三个：找出文档中能够体现情感的词或短语；判断所找出的词或短语的倾向性极性以及强度；找出所抽取的词或短语与主题的关系。

针对‘网站发布违规药品广告’这一舆情需求，本系统需要判断采集的广告类数字内容是正面还是负面的，即网站发布的药品广告是否违规，这就需要对这些信息进行倾向性分析。主要步骤包括：

(1) 构建敏感信息库

本系统研究结合今后在广播电视广告类内容方面进行舆情监控的实际需求，部分参照新颁布的广播电视广告法（第 61 号令），构建了敏感信息库，包含了治愈率、肿瘤、夸大等众多敏感词，通过人工逐一对敏感词进行倾向性分析，赋予每个词一个权重。

(2) 广告违规判别

结合敏感信息库对从终端网页中提取的内容信息进行语义关系分析，当提取出的药品广告内容包含这些敏感信息时，即断定这个广告违规。

5 结语

本文提出了一个基于互联网的数字媒体内容舆情分析模型，设计并实现了基于互联网的数字媒体内容舆情分析系统，并从系统工作流程、体系架构、关键技术应用等方面加以详细讨论。下一步工作是该系统将在相关广播电视监管部门试运行，通过测试分析将对系统存在的问题做进一步的改进和完善。

参考文献

1 刘磊.网络舆情分析系统研究.情报探索,2010,10:106–108.
2 许鑫,章成志.互联网舆情分析系统及应用研究.情报科学,2008,26(8):1194–1204.
3 钱爱兵.基于主题的网络舆情分析模型及其实现.现代图书情报技术,2008,163(4):49–55.
4 戴媛,程学旗.面向网络舆情分析的实用关键技术概述.信息网络安全,2008,6:62–65.
5 张焕明.网络舆情分析系统的研究与设计.网络与通信,2010, 26(6-3):118–121.
6 路斌.互联网舆情热点自动发现与分析技术研究与实现[学位论文].北京:北京大学,2007.
7 来火尧,刘功申.基于主题相关性分析的文本倾向性研究.信息安全与通信保密,2009,3:77–81.