文章编号：1671-0444(2018)08-0586-04

基于多源数据的教育网络舆情分析

殷　红[a]，孙　凯[b]，王长波[b]
(华东师范大学 a. 经济与管理学部；b. 计算机科学与软件工程学院，上海 200062)

摘要：针对教育网络舆情表达形式的多元化，基于多源网络数据，构建一个教育网络舆情的分析系统。通过收集和融合处理教育网络数据，对其进行分词、聚类与关联度分析和教育舆情的时变特征分析；随后分析教育舆情的情绪，并对比分析教育事件的发展规律。案例验证本文系统分析教育舆情的有效性。

关键词：网络数据；教育舆情；情绪；可视分析
中图分类号：TP 301.6　　**文献标志码**：A

Analysis of Education Public Opinion Based on Multi-source Data

YIN Hong[a], *SUN Kai*[b], *WANG Changbo*[b]
(a. Faculty of Economics and Management; b. School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China)

Abstract: Facing with the diversity of the expression form of network public opinion in education, this paper constructed an analysis system for educational public opinion based on multi-source network data. Through the collection and data fusion processing to analyze the participle, clustering and correlation, the time-varying characteristics of educational public opinion were analyzed as well. Then, through the analysis of the sentiment of educational public opinion, the development of educational events was compared and analyzed. Some cases show that this system is intuitive and effective on public opinion analyze in education.

Key words: network data; education public opinion; sentiment; visual analysis

随着移动互联网与社交网络技术的发展，教育舆情的表达形式日益多元化，教育事件的热点层出不穷。另外，随着教育综合改革的深化，教育主管部门也希望得到公众对有关教育政策调整以及改革措施的反馈，为进一步的教育决策提供依据。但是，当前关于教育舆情分析的工作较多地聚焦于教育事件的定性分析，如采用问卷等形式，使得数据的来源不够，分析的角度和内容也多局限于统计结果分析等。随着大数据技术的发展，通过网络数据的挖掘和关联对比分析，可以更深入地发现教育舆情的传播规律，挖掘相关教育事件的深层次观点。

目前网络舆情分析主要集中于事件监测分析等，包括网络舆情观点、情绪的挖掘[1]以及网络集群行为的监测分析[2]。目前有关于教育舆情的分析和监测报告[3-4]主要偏重于统计情况分析。

利用大数据分析技术来进行教育舆情分析是近年来的研究热点。兰月新等[5]研究了大数据背景下网络舆情的主体交互机制；张鹏高等[6]进行了基于

收稿日期：2018-04-30
基金项目：上海市科委软科学研究资助项目(1769210440)；国家自然科学基金资助项目
作者简介：殷　红(1976—)，女，湖北随州人，博士，副教授，研究方向为信息管理系统，E-mail：hyin@jjx.ecnu.edu.cn
王长波(联系人)，男，教授，E-mail：cbwang@sei.ecnu.cn

大数据的教育舆情监控与分析；王丹丹等[7]构建了新媒体和大数据背景下的多校区高校网络舆情体系；Sun 等[8]基于网络数据分析了教育舆情中的知识图谱。上述研究分析的直观度和跨媒体的融合度还不够。本文通过对微博、论坛、新闻网站等的融合分析，挖掘其中的规律，并研发相关在线分析和可视化系统。

1　多源教育网络数据

针对教育网络舆情的分析需求，首先需要采集和收集数据并建立教育舆情数据库。这里关注的数据源主要包括：(1)门户网站，包括教育相关新闻数据和用户评论数据；(2)教育新闻网，涵盖全面的教育政策、地方教育动态；(3)论坛、博客、微博、微信公众号等自媒体平台；(4)在线版电视节目和报纸等。

本文收集了近 2 年的教育网络舆情数据，其中包含：约 250 万篇文章、帖子和微博数据，约 2000 万条用户评论和回帖数据。舆情数据来源分布如表 1 所示。

表 1　教育网络舆情的数据来源分布

Table 1　The data sources of educational public opinion

序号	教育舆情数据来源	分布占比/%
1	门户网站	15.12
2	政府教育网	7.23
3	博客	3.11
4	微信公众号	12.76
5	网络论坛	35.23
6	电视报纸	9.73
7	新浪微博	16.82

由于涉及到多种媒体数据，且数据的语言风格、格式、表达方式各异，因此在分析之前需要对数据进行预处理，这里需要进行跨媒体的关联分析以及数据间关联和索引。数据处理的基本思路如图 1 所示。这里从不同的媒体数据源采集和收集数据，然后存储到教育网络舆情数据库中，进一步可进行数据清洗和关联分析，以及数据融合和信息索引，相关结果也可以更新到数据库中。

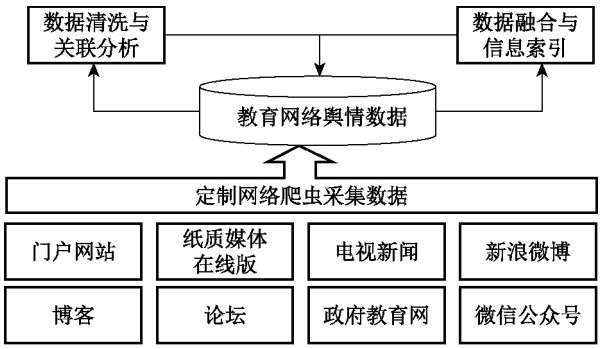


图 1　舆情数据的预处理方案

Fig. 1　The pre-processing method for the public opinion data

2　网络教育事件的挖掘分析

2.1　网络事件的挖掘分析

针对采集到的数据，利用主题模型来分析文本中出现的词汇，并依据词汇的关联信息进行聚类，从而推断出文本中隐含的主题。每个网络事件或者话题一般可以由关键词集合及其权重进行表示，例如与“校园暴力”相关的话题，可以包含“学生”“欺凌”“校园”等关键词，通过分析每篇文档在这类话题上关键词的分布，自动检测出与该类话题最为相关的舆情新闻和公众评论。

将所有数据按照与教育相关或不相关进行分类，保留教育舆情相关的信息，去除数据源中与教育相关度较低的数据。相比传统的关键词匹配方法，本文方法可以自动将数据划分到不同的教育主题，提升了数据处理的效率。

2.2　多源数据的融合聚类

对于同一个教育事件，不同的媒体可能有不同的描述和评论，因此多源数据的融合处理是个难点。将相同或相似地名、人名、机构名等要素的信息聚成一个簇，不同网络媒体数据的新闻自动聚成若干个簇，每个簇内包含一个热点事件。

具体而言，将从文本中挖掘出的关键词及核心信息看作一个个实体，进行实体抽取和归一化。这里的关键问题是有些教育事件的相关文档数量较少，较难采用基于模式的方法来进行关系抽取[9]。因此，提出一种基于实体关系对的方法来识别候选实体。实体关系对是指具有上下文环境的多个实体组合，可以将拥有相似的上下文环境的实体对聚合成一个原始的关系，如式(1)所示。

$$R = \{(e_i, e_j, C_{ij})\} \tag{1}$$

式中：e_i 和 e_j 为通过关键词抽取出来的实体对；C_{ij} 为 e_i 和 e_j 的上下文环境。实体对中若某一个实体被识别，则其他实体仍旧作为候选实体。

2.3　网络舆情的情绪分析

教育舆情的另外一个重要方面是教育事件中的大众情绪。根据教育舆情特点对中文情绪词典进行调整，具体情绪分析思路如图 2 所示。

这里需要解决两个问题：一是在情绪词典里面可能没有一些网络新名词，需要对其进行人工标注，同时增加与教育相关的名词情绪值；二是有些词无法判断是正面还是负面的情绪，可以利用有监督的学习

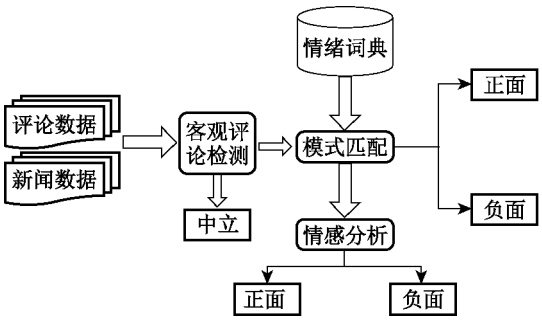


图2 教育舆情的情绪分析

Fig. 2 The sentiment analysis of educational public opinion

方法,即把一部分评论样本数据进行训练得到分类器,然后用这个分类器去区别正面和负面的情绪。

通过情绪分析可得到不同分词的情绪,对于整个新闻或者评论段落,将综合不同分词的情绪进行统计。如某句话中每个分词都有一个情绪值 m_i,正面情绪取值为(0, 1),负面情绪取值为(−1, 0),中立情绪取值为0,则这句话的整体情绪为 $\sum m_i (0<i<N)$,其中 N 为分词的个数。图2中情感分析即找到评论里面的主观句子,根据主观句子里的情绪属性,再计算属性对应的情感分。

3 教育网络舆情的分析

基于上文的数据处理和分析结果,可以分析教育舆情事件的时序分布和地域分布,从而得到整个教育网络舆情的时变趋势。2017年中国教育舆情事件在每个月份上的分布占比如表2所示,其中9月份的教育舆情事件是最多的。2017年中国教育网络舆情在地域上分布如表3所示,由此可知东部地区对于教育的关注度最高。

表2 2017年中国教育网络舆情事件的时间分布

Table 2 The time distribution of Chinese educational public opinion in 2017

时间	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
教育舆论事件数	12	6	16	20	27	22	20	23	33	25	20	12
分布占比/%	5.1	2.5	6.8	8.5	11.4	9.3	8.5	9.7	14.0	10.6	8.5	5.1

表3 2017年中国教育网络舆情事件的地域分布

Table 3 The location distribution of Chinese educational public opinion in 2017

地域	东部	中部	西部
教育舆情事件数	107	76	53
分布占比/%	45.2	32.3	22.5

图3是对某一教育事件的大众情绪时变图,可以看到开始时负面情绪比较多,后来慢慢地趋于正面情绪。

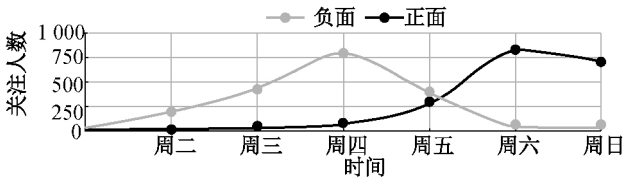


图3 教育事件的情绪时变图

Fig. 3 The sentiment change map of educational public opinion

不同媒体对于同一教育舆情事件的观点不一定相同,由此可以进一步分析它们的观点差别。例如,采用雷达图的形式对比分析不同媒体对“新高考改革”事件的报道如图4所示。雷达图的6个轴分别表示媒体报道的六维属性,即文档数量、中立观点数、反向评论数、正向评论数、回复数量、覆盖媒体数量。由图4可知,针对“新高考改革”事件,新浪教育的负向评论较多,而凯迪论坛的中立观点偏多。

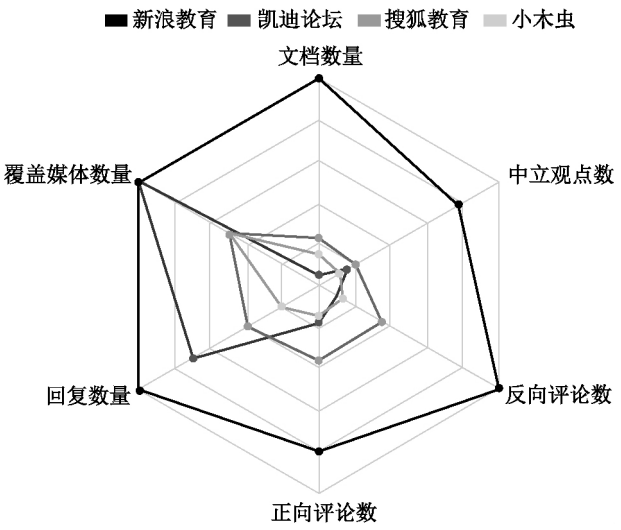
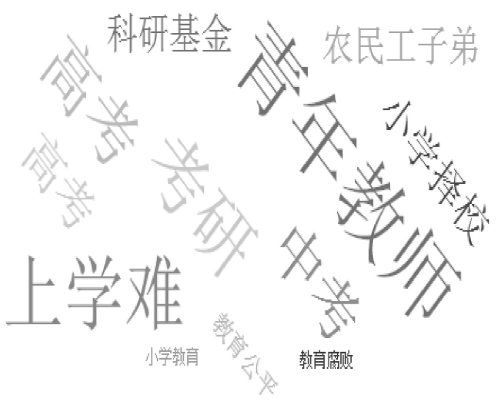


图4 不同媒体的对比分析

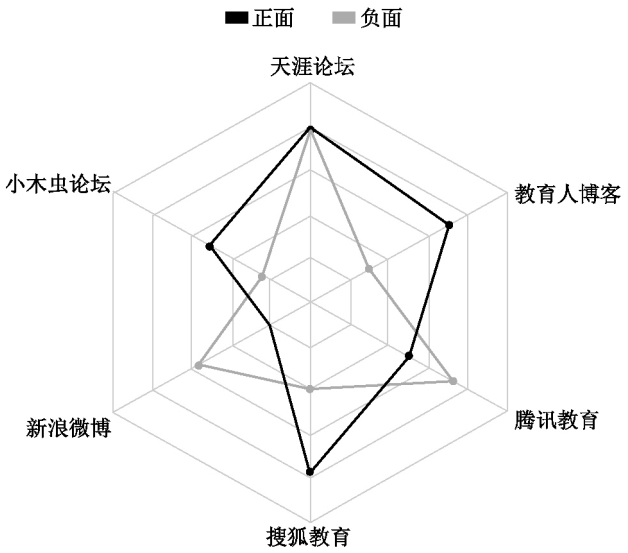
Fig. 4 The comparison analysis of different media

某时间的热门教育舆情事件的分布如图5所示。由图5(a)可以看到,“高考”“上学难”“青年教师”等还是关注的焦点,这里字体越大表示关注的人数越多。由图5(b)可以看出,腾讯教育、天涯论坛、

新浪教育的观点是负面的，而小木虫论坛、搜狐教育、教育人博客的观点是正面的。



(a)关注焦点



(b)不同媒体对比

图 5　教育舆情分析图

Fig. 5　The analysis of educational public opinion

基于上面分析，本文进一步采用 Java 编程语言研发了一个教育网络舆情的在线分析系统，该系统包括舆情热点、舆情时间、舆情对比等功能，并且具备多视图切换和交互功能，可以点击不同视图进行缩放、平移和高亮显示。

4　结　语

教育网络舆情受到了越来越多的关注，本文通过收集和分析包括教育网站、社交媒体等不同数据来源的教育舆情数据，利用网络事件挖掘以及舆情情绪分析方法，研究面向教育网络数据的舆情分析，可以有效地得到教育网络舆情的时变趋势，以及对某些教育事件的观点对比分析，从分析结果可以看出所提出的方法是可行和实用的。

然而，本文所构建的教育网络舆情分析系统在深度上还有可改进的空间，包括进一步挖掘不同类型人群对于舆情的观点，并揭示舆情传播的规律。目前的舆情分析系统还不是实时的，下一步可以研发实时采集和分析处理的系统。另外，如果能对网络教育事件的发展进行预测，对于舆情的管理及决策将有更大的作用。

参　考　文　献

[1] WANG C B, LIU Y H, XIAO Z, et al. Analyzing internet topics by visualizing microblog retweeting[J]. Journal of Visual Language and Computing, 2015, 28: 122-133.

[2] PRIYANKAR S, SHIVANI S, NAGPAL D. Sentiment analysis and opinion mining [J]. International Journal of Computer Applications, 2018, 180 (20) :14-16.

[3] 何晓丰，朱益明，王长波，等. 2015 年中国教育网络舆情分析报告[M]. 上海：华东师范大学出版社，2016.

[4] 王保华. 中国高等教育舆情报告(2017)[M]. 北京：高等教育出版社，2017.

[5] 兰月新，王芳，张秋波，等. 大数据背景下网络舆情主体交互机理与对策研究[J]. 图书与情报，2016(3)：28-37.

[6] 张鹏高，毕曦. 基于大数据的教育网络舆情监控与分析[J]. 中国教育信息化，2015(15)：7-9.

[7] 王丹丹，彭利美，余锦燕. 新媒体和大数据时代背景下多校区高校网络舆情体系的构建[J]. 法制与社会，2018(10)：167-168.

[8] SUN K, LIU Y H, GUO Z C, et al. EduVis: visualization for education knowledge graph based on web data[C]// Proceeding of the 9th International Symposium on Visual Information Communication and Interaction. 2016: 138-139.

[9] SHEN W, WANG J, LUO P, et al. REACTOR: a framework for semantic relation extraction and tagging over enterprise data [C]//In Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM, 2011: 121-122.

(责任编辑：徐惠华)