引入时间机制的网络舆情演化分析方法研究

郑步青[1] 邹红霞[2] 胡欣杰[2] 王 桢[1]
(航天工程大学研究生管理大队 北京 101416)[1] (航天工程大学信息装备系 北京 101416)[2]

摘 要 网络舆情的迅速发展使舆情发展成为研究热点,它对舆情的预测预警具有重要意义。从文本聚类入手,针对舆情的演化分析过程,在时间序列上进行K-means聚类研究,得到聚类中心,又依此对聚类中的词频统计进行时序加权处理,使统计所得关键词更具有代表性。通过对时间聚类法和时序加权统计法所得关键词的分析,得到了舆情演化的趋势。研究结果表明该方法降低了聚类的维数,减少了噪声,提高了聚类的准确度,增强了演化分析的可信度。

关键词 网络舆情,时间聚类,加权,关键词,演化分析
中图法分类号 G203 **文献标识码** A

Research on Evolution of Network Public Opinion Introducing Time Mechanism

ZHENG Bu-qing[1] ZOU Hong-xia[2] HU Xin-jie[2] WANG Zhen[1]
(Company of Postgraduate Management, Space Engineering University, Beijing 101416, China)[1]
(Department of Information Equipment, Space Engineering University, Beijing 101416, China)[2]

Abstract The rapid development of network public opinion makes the evolution of public opinion become the research hotspot, which is of great significance for the forecast of public opinion. In this paper, we started from text clustering, for the evolution of public opinion analysis process, making the K-means clustering research in time series, and got clustering center. The time-weighted weighting of word frequency statistics in clustering was made, which makes the statistical keywords more representative. Through the analysis of the keywords obtained by time clustering and time series weighted statistical method, the trend of public opinion evolution was got. The results show that the method reduces the dimension of clustering and the noise, improves the accuracy of clustering, and enhances the reliability of evolution analysis.

Keywords Network public opinion, Time clustering, Weighting, Keywords, Evolutionary analysis

随着互联网的发展和普及,网络为人们提供了丰富的信息资源,网络媒体已经被公认为继报纸、广播、电视之后的"第四媒体"。根据中国网络信息中心第39次中国互联网发展状况报告显示,截至2016年12月,我国网民规模达7.31亿,全年共计新增网民4299万人,手机网民规模达6.95亿。互联网成为反映社会舆情的主要载体,是舆情传播的重要平台。

舆情是指在一定的社会空间内,围绕中介性社会事件的发生、发展和变化,民众对社会管理者产生和持有的社会政治态度[1]。网络舆情存在于海量的网络数据中,它具有海量高维的特点,在时间上具有继承性和延续性,在内容上具有交互性;同时舆情信息涉及的话题具有演化性,也具有一定的可预测性。一般来讲,舆情也是大数据,具有大数据的5"V"特点[2]。对于舆情的研究,难点在于网络舆情数据的大量化、多维化,数据的多元化和非结构化会影响处理的效率和质量。

本文着重对舆情的演化性进行研究,首先对网络舆情数据进行聚类,提出了基于时间序列的聚类研究,建立了时间片内的K-means聚类模型,各个时间片内聚类的结果就是各个时间阶段话题关键词的概括;然后基于舆情中公众的情感倾向,在时间片内对聚类中的TFIDF的词频统计进行了改进,得到时序的关键词统计结果;最后通过对两个部分所得到的聚类中心和关键词进行比较分析,得到舆情的演化过程。

1 K-means时间序列聚类

1.1 K-means聚类

K-means聚类是目前聚类分析中应用得最为广泛的一种方法,其定义是提取到原始数据的集合为$\{x^1,x^2,\cdots,x^i\}$,每个x^i为d维的向量,在给定分类组数$k(k\leqslant n)$值的条件下,将原始数据分成k类,用每个类中所有数据的平均值或加权平均来表示每个类,这个平均值就是聚类中心。K-means算法[3]一开始在数据集合中随机选择k个数据对象作为初始的聚类中心,然后利用欧氏距离方法把所有剩余的数据与初始中心点进行距离比较,计算离初始中心较近的数据均值,得到新的聚类中心,如果聚类中心变化,则继续进行迭代,直到k个聚类中心不再变化为止。

1.2 时间序列

时间序列是指数据按照时间顺序排列的一种数据形式,目的[4]是从大量时间序列数据中发现未知的重要模式和知识,并据此做出具有知识驱动的决策,它广泛存在于各种大型的商业、医学、工程和社会科学等数据库中。时间序列也可称为动态序列[5],它不同于传统的静态数据,是一种复杂的数据

郑步青(1993—),男,硕士生,主要研究方向为舆情数据的处理;邹红霞 女,教授,主要研究方向为信息对抗和信息处理。

对象，可以用来描述舆情的演化过程。目前，基于时间序列数据的聚类研究还很少见。在当前时间序列聚类算法的研究中，根据聚类对象可以把时间序列聚类分为全序列聚类、子序列聚类、时间点聚类[6]。

1.3　时间序列聚类模型

K-means 全序列聚类研究的重点是在聚类的基础上添加了时间规则。在时间上进行数据的分类整合，整合得到的数据往往代表着该时间片内的演化主题。

假设舆情的原始数据为$\{x^1,x^2,\cdots,x^i\}$，初始化 k 个随机数据$\{u_1,u_2,\cdots,u_k\}$作为初始中心点，同时选取时间序列$\{T_0,T_1,\cdots,T_n\}$。在一个时间片内根据下列 K-means 聚类的两个迭代公式求出最终所有类的聚类中心 u_k。

首先求出时间片内所有数据和初始化的随机数据的距离，找出距离每个初始数据最近的原始数据 C^i。

$$C^i=\arg\min_j \| x^i-u_j \|^2, T_n<t<T_{n+1} \tag{1}$$

然后计算初始数据和最近原始数据的距离均值，不断迭代两个公式，直到 u_j 不再变化，即得到最终的聚类中心 u_j。

$$u_j=\frac{\sum_{i=1}^{m}\{c^i=j\}x^i}{\sum_{i=1}^{m}\{c^i=j\}}, T_n<t<T_{n+1} \tag{2}$$

2　时间序列词频加权统计

在对数据进行聚类时，需要对数据进行 TFIDF 词频统计处理。TFIDF(Term Frequency-Inverse Document Frequency)是一种统计方法，用以评估某字词对于一个文件集或一个语料库中的其中一份文件的重要程度。若某个词语在一篇文章中出现较多，且在其他文章中出现较少，那么这个词就具有较好的区分能力，该篇文章中的这些词构成的含义就是文章的核心思想，适用于分类与聚类的研究。

获得的舆情数据往往带有公众的情感倾向，从网民的评论、点赞数等可以反映出来，点赞数表明网民的认可程度，评论数表示网民的参与度。文献[7]针对微博中的转发数、表态数和评论数，采用了层次分析法，对传统的 TFIDF 特征权重计算方法做出改进，提出一种微博文本特征权重计算新方法，实验表明改进方法具有较好的聚类效果。

在演化过程中考虑加入对网民的态度分析可以提高聚类的准确度，增加舆情分析的真实性、正确性。本文对 TFIDF 词频统计的计算进行了改进，加入了时间序列，采用了层次分析法，并在每个时间序列内以舆情中带有网民情感倾向的评论数、点赞数等为舆情指标，对指标进行加权来实现舆情的演化分析，最后将统计所得的关键词与时间序列聚类的中心进行对比分析。

假设第 i 个指标的权重为 A_i，且该指标的数据大小为 X_i。首先对各指标进行标度，对加以权重的指标进行量化并建立判断矩阵，如以下矩阵所示，在判断矩阵中标度 a_{ij} 表示第 i 行指标与第 j 行指标的权重大小比较，且 $a_{ij}=\frac{1}{a_{ji}}$，$a_{ii}=1$。

$$\begin{bmatrix} 1 & a_{12} & \cdots & a_{1j} \\ a_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & \cdots & 1 \end{bmatrix}$$

其中矩阵中标度 a_{ij} 的含义如表 1 所列。

表 1　标度的含义

标度	含义
1	表示 i 元素与 j 元素同等重要
3	表示 i 元素比 j 元素稍微重要
5	表示 i 元素比 j 元素较为重要
7	表示 i 元素比 j 元素明显重要
9	表示 i 元素比 j 元素极端重要

对每行元素计算几何平均值得到指标的权重向量$w=(\bar{A}_1,\bar{A}_2,\bar{A}_3,\cdots,\bar{A}_i)^T$；然后通过对指标进行归一化处理后得到$\bar{X}_i$，则加权后关键词出现的次数为：

$$n_{ij}=\sum_{i=1}^{j}[1+\lambda*(\bar{A}_1\bar{X}_1+\bar{A}_2\bar{X}_2+\cdots+\bar{A}_i\bar{X}_i)]*I_i,$$
$$T_n<t<T_{n+1} \tag{3}$$

其中，λ 是权重影响因子，j 表示总文档数，I_i 表示词语在第 i 篇文档中出现的次数。则词频的计算公式如式(4)所示，$\sum_k n_{kj}$ 表示文档中所有字词出现的次数之和。

$$tf_{ij}=\frac{n_{ij}}{\sum_k n_{kj}}, T_n<t<T_{n+1} \tag{4}$$

3　网络舆情实验结果分析

本文以微博为实验数据抽取平台，以 2016 年里约奥运会女排夺冠事件作为核心，对微博博文进行部分数据的采集。事件大概是 2016 年 8 月 21 日的里约热内卢奥运会，中国女排在决赛中以 3:1 击败塞尔维亚，时隔 12 年再次夺冠，这一振奋人心的消息在网络上引起了网民广泛的关注和讨论。本文采集的数据时间从决赛前一天到 12 月份，包括一部分较为热门的原创博文，以及这些原创博文下的转发及评论。据统计，采集的博文共有 882 条，对其进行数据的清洗，去除垃圾数据和重复数据，得到有效博文共 843 条。博文中包含的有效信息有博主 ID、博主头像、博文、发布时间、转发数、评论数、点赞数等。

首先在时间上对发布的博文数进行 MATLAB 曲线拟合，如图 1 所示。

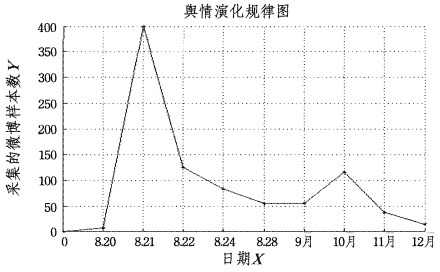

图 1　微博样本随时间变化图

通过所拟合的曲线图，发现通过微博爬取的女排夺冠事件符合一般舆情的演化规律，即舆情的产生、爆发、逐渐减弱，最后到舆情消亡，可以对其进行舆情分析。女排夺冠时间是 8 月 21 日，在决赛开始之前女排就拥有一定的关注度，夺冠后话题热度迅速增长，而后持续几个月话题还未消亡，说明了话题的爆发速度快、传播速度快、话题影响力较大等特点，反映了该舆情的突发性和持续性。

3.1 词频统计对比分析

将博文的发布时间作为时间序列排序的依据，博文中的点赞数、评论数和转发数为舆情指标，对其进行指标权重分配，得到时间序列词频加权统计结果。后期除去介词、语气词等一些无用词语后，得到每段时间片内的重要关键词。将其与无加权词频统计结果进行对比，结果如图 2－图 5 所示。图 2－图 5 是截取的几个时间片内的词频统计对比图，其中纵坐标是关键词，横坐标是关键词出现的次数。从词频统计上看女排决赛很受关注，不少网民收看直播为女排加油。在里约奥运会上，女排夺冠后，网民第一反应是感动、致敬。而后主要关注于女排的拼搏顽强的精神和女排精神所带来的时代烙印。网民讨论的话题也从事件本身转移到事件内在和给社会带来的促进作用上。

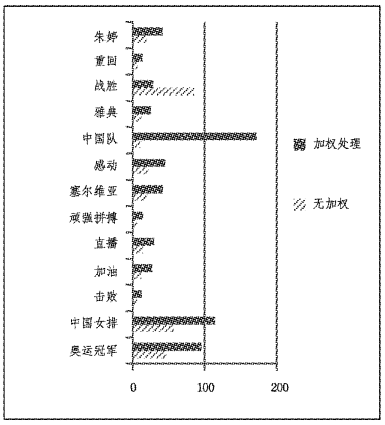

图 2 8 月 21 日词频统计对比

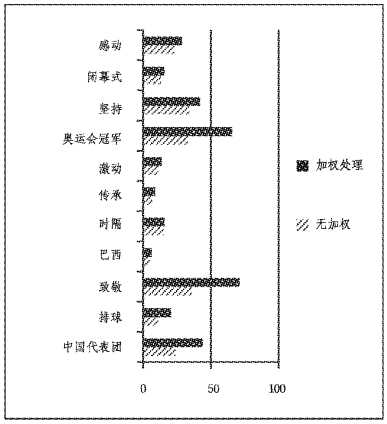

图 3 8 月 22 日词频统计对比

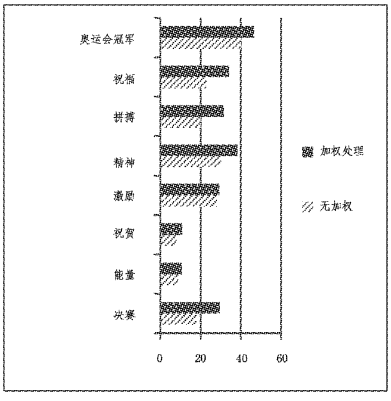

图 4 8 月 28 日词频统计对比

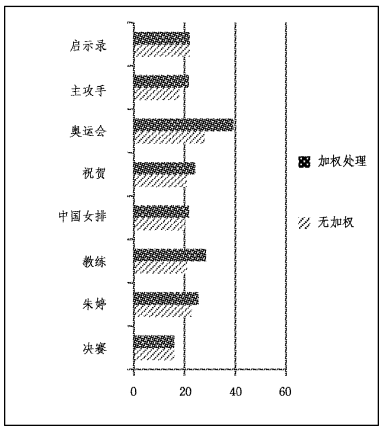

图 5 9 月词频统计对比

通过分析表中数据可以发现，加权 TFDIF 词频处理与无加权的舆情在内容上的演化情况大致一样，始终围绕着女排、奥运会冠军、精神等关键词。但与无加权 TFDIF 词频统计的结果相比较，加权 TFDIF 词频统计出现词数变化的情况，其中分析可得词数变化较大的是舆情内容演化中较为关键的词语。这些关键词带有网民的情感倾向，反映了舆情的演化，使得演化过程更具完整性。

3.2 文本聚类统计分析

对采集的博文信息进行 K-means 聚类，根据事件的走向进行舆情的聚焦和分析。文本的聚类和数据的聚类方法不一样，需要先将文本进行向量化，用向量来代表文本，并通过计算向量之间的欧氏距离来进行文本聚类。表 2 是几个时间片内文本内容聚类的结果。

表 2 K-means 时间聚类的结果

时间序列	聚类簇 k	聚类结果中心概括
8 月 21 日	10	中国女排 顽强拼搏 奥运冠军 击败 塞尔维亚 加油 感动 直播 战胜 雅典
8 月 22 日	12	中国代表团 致敬 排球 巴西 时隔 奥运会冠军 传承 激动 奥运会 坚持 感动 闭幕式
8 月 28 日	8	决赛 能量 祝贺 精神 激励 决赛 拼博 祝福
9 月	8	决赛 朱婷 教练 中国女排 祝贺 奥运会 主攻手
10 月	4	奥运会 启示录 精神 女排

从表 2 可知，在时间序列下聚类的结果较为明显，聚类簇较为确定，每个时间片内聚类中心大都聚焦在奥运会冠军、精神等关键词语上，聚类中心的变化也和词频统计所得到的关键词大致一样，都是从事件本身到事件的内在影响，符合舆情演化的规律。

3.3 结果分析

(1)从话题产生时间和发展上看，事件往往在发生后的 2～4天内关注度及网民讨论程度最高[8]；同时舆情演化发展中会受到多种因素的影响，影响着舆情发展的强度和方向，故在事例中出现了 10 月份关注强度增加的情况，但在整体上仍旧满足舆情演化的一般规律。

(2)从聚类的结果看，在聚类中加入时间序列一定程度上降低了聚类的维数，增加了聚类的准确度。同时在词频统计中加入时间加权，虽然前后关键词变化不大，但是突出了每个时间片内关键词的重点，在一定程度上反映了网民的意愿，提高了演化分析的正确度。

(3)通过改进的词频统计结果和时间聚类中心的对比可以发现演化过程基本一致，因此综合两者之间的关键词变化情况，可以加强对舆情的内容演化分析。同时注意到改进的词频统计所得到的关键词比聚类所得到的要多，分析所得到的内容更加完整，更具有代表性，也更符合舆情的实际演化情况。在进行分类聚类时，可以考虑用改进的 TFIDF 所得到的词频统计结果指导聚类的初始中心和类别数。

结束语　本文通过对加权词频统计和时序聚类结果的分析初步完成了对女排夺冠话题内容演化的模拟分析，有效地实现了时间序列的聚类，能够在舆情分析预测上提供支撑依据。同时以时间为维度，考虑公众的情感倾向，以里约奥运会女排夺冠为例，得到舆情从话题本身往话题内在影响方面转移的演化趋势，提高了话题内容演化分析的准确性，从而能够在舆情发展上提供一定的指导。但本文需要在以下两个方面进一步深化研究：一是数据爬取的全面性和完整性；二是数据预处理中对噪声和无用数据的过滤。

参 考 文 献

[1] 刘毅. 网络舆情研究概论[M]. 天津人民出版社，2007.

[2] 侯万友. 群体性突发事件微博舆情演化分析[D]. 哈尔滨：哈尔滨工业大学，2013.

[3] 王雪梅，李晓峰，高巍巍. 一种改进的 K-Means 聚类算法的研究[J]. 计算机与数字工程，2013，41(11)：1717-1719.

[4] 刘慧婷，倪志伟. 基于 EMD 与 K-means 算法的时间序列聚类[J]. 模式识别与人工智能，2009，22(5)：803-808.

[5] 李深洛. 基于特征的时间序列聚类[D]. 桂林：广西师范大学，2014.

[6] 韩娜. 聚类算法在时间序列中的研究与应用[D]. 广州：广东工业大学，2011.

[7] 黄晓军，王博，包秀国. 基于层次分析法的微博文本特征权重计算方法[J]. 通信学报，2016，37(12)：50-55.

[8] 雷春，付业勤. 旅游网络舆情事件的时空分布与演化规律分析——以海南旅游热点事件为例[J]. 韶关学院学报，2014，35(1)：114-119.

(上接第 410 页)

[17] HAN J W，KOPERSKI K. Discovery of spatial association rules in geographic information databases[C]//Proceedings of the 4th International Symposium on Advances in Spatial Databases，Maine. USA，1995：47-66.

[18] HOUTSMA M，SWAMI A. Set-oriented mining for association rules in relational databases[C]//Proceedings of the Eleventh International Conference on Data Engineering. 1995：25-33.

[19] CHEN C，YAN X F，ZHU F D，et al. Graph OLAP：a multi-dimensional framework for graph data analysis[J]. Knowledge and Information Systems，2009，21(1)：41-63.

[20] 何清，李宁，罗文娟，等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能，2014，27(4)：327-337.

[21] 沈斌，姚敏. 一种新的动态关联规则及其挖掘算法[J]. 控制与决策，2009，24(9)：1310-1315.

[22] 李玲娟，张敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展，2011，21(2)：43-46.

[23] 杨勇，王伟. 一种基于 MapReduce 的并行 FP-growth 算法[J]. 重庆邮电大学学报(自然科学版)，2013，25(5)：651-659.

[24] BOUKOUVALA F，DUBEY A，et al. Computational Approaches for Studying the Granular Dynamics of Continuous Blending Processes，2-Population Balance and Data-Based Methods[J]. Macromolecular Materials and Engineering，2013，297(1)：9-19.

[25] LIN T Y. Granular computing[M]//Rough Sets，Fuzzy Sets，Data Mining，and Granular Computing. Springer Berlin Heidelberg，2003：16-24.

[26] ZADEH L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets System，1997，90(2)：111-127.

[27] PAWLAK Z. Rough sets[J]. International Journal of Computer and Science，1982，11：341-356.

[28] 张铃，张钹. 模糊商空间理论(模糊粒度计算方法)[J]. 软件学报，2003，14(4)：770-776.

[29] 徐计，王国胤，于洪. 基于粒计算的大数据处理[J]. 计算机学报，2014，37(11)：1-22.

[30] 王国胤，张清华. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报，2008，31(9)：1588-1598.

[31] YAO Y. Perspectives of granular computing[C]//2005 IEEE International Conference on Granular Computing. 2005：85-90.

[32] YAO Y Y. Granular computing：basic issues and possible solutions[C]//Proceedings of the Fifth Joint Conference on Information Sciences. 2000：186-189.

[32] 张清华，王国胤，胡军. 多粒度知识获取与不确定性度量[M]. 北京：科学出版社，2013.

[33] 苗夺谦，王国胤，刘清. 粒计算：过去、现在与展望[M]. 北京：科学出版社，2007.

[34] 王国胤，李德毅，姚一豫，等. 云模型与粒计算[M]. 北京：科学出版社，2012.

[35] 张钹，张铃. 粒计算未来发展方向探讨[J]. 重庆邮电大学学报(自然科学版)，2010(5)：538-540.

[36] 钟珞，吴珺. 粒度计算在数据仓库挖掘中的应用[J]. 华中师范大学学报(自然科学版)，2009，43(3)：392-395.