基于藏文网络的舆情传播模型①

邓竞伟, 邓凯英, 李永生, 李应兴

[1](西北民族大学 数学与计算机科学学院, 兰州 730124)
[2](西北民族大学 经济学院, 兰州 730124)

摘 要: 通过研究网络舆情传播的发展现状、趋势及特点, 介绍了藏文网络舆情传播的特点和作用. 根据复杂网络理论的研究方法对藏文网络舆情传播规律进行实证分析, 设计了一个藏文网络舆情模型结构. 研究表明, 通过聚类可以提高藏文网络舆情的准确性, Web 挖掘能够有效地从藏文网络上获取并分析相关舆情信息.

关键词: 网络舆情; 信息分析; 文本挖掘

Opinion Spreading Models on Tibetan Networks

DENG Jing-Wei, DENG Kai-Ying, LI Yong-Sheng, LI Ying-Xing

[1](School of Mathematics and Computer Science , Northwest University for Nationalities, Lanzhou 730124, China)
[2](School of Economics, Northwest University for Nationalities, Lanzhou 730124, China)

Abstract: This paper discusses the research of the complex network opinion spread situation, trends and characteristics. The features and functions of tibetan networks opinion spread are introduced. Based on the complex network theory research methods, we design tibetan networks public opinion model. Research shows that Web mining can effectively obtain the relevant opinion information and enhance the accuracy about tibetan networks public opinion.

Key words: internet public opinion; information analysis; text mining

随着网络信息技术的飞速发展, 网络舆情受到许多学者的广泛关注. 网络舆情传播是重要的研究方向之一[1]. 能够造成网络舆情的因素主要包括帖子、Blog、图片、短信、评论等等. 网络舆情是社会舆情的一种表现形式, 是公众在网上公开表达对某种社会现象或社会问题的具有一定影响力和倾向性的共同意见[2,3]. 与全国用户增长速度相比, 藏族用户的增长速度尤为突出. 因此, 藏文网络舆情是当前必须关注的信息传播与舆论涌现现象. 藏文网络舆情的数量也在不断增加, 网络信息传播的途径也日益呈现多样化. 根据藏文网络的这些特点, 在技术上不容易实现, 如果想在短时间内迅速获取大量信息相对不容易. 另外, 目前藏文网页数量巨大, 处理速度相对很慢, 以往的网络舆情页面的统计是基于手工统计, 效率非常低, 很难对网络舆情变化做出迅速响应.

由于藏文网络传播媒介的特殊性, 网络舆情表现出一些不同的特点[4]: 它具有广泛性和匿名性, 自由度和不可控制性, 互动性和即时性, 且网络传播影响范围和程度较大. 网络舆情[5]由网民, 公共事务, 情绪, 意愿, 态度和意见, 时空因素, 强度等构成要素组成. 信息传播中受传者有着从众心理, 会受到周围人的影响程度与其中接受信息的人的比例有关[6,7].

1 Web信息挖掘

互联网已经成为网络舆情监控的重点, 怎样获取更准确的网络舆情信息是目前的首要任务. 它对网上大量文本进行表示、特征提取、分类、聚类、内容总结、关联分析、语义分析以及利用网络文本进行趋势预测等[8,9]. 主题网络爬虫通过链接内容推测和链接聚类推测[10,11]. Web 信息挖掘可以有效地从互联网上获取并分析相关舆情, 达到预警和辅助决策的目的, 为网络舆情预警提供了极大的帮助[12].

① 基金项目:教育部人文社科青年基金(12YJCZH027);规划基金(11YJAZH053)
收稿时间:2012-08-25;收到修改稿时间:2012-09-22

1) 度分布: 网络中节点的度指的是与该节点连接的边数, 网络中节点的度所占的比例, 即随机选取一个网络节点, 节点度的概率分布函数用 $P(k)$ 来表示, 它指的是节点有 k 条边连接的概率, $P(k)$ 的期望称作网络的平均度. 如公式(1)所示.

$$p(k)=\sum_{k'}p(k') \tag{1}$$

2) 平均路径长度: 在网络拓扑结构中, 节点到所有其它节点间距离的平均最短距离, 节点间的距离指的是从一个节点到另一个节点所要经历边的最小数目, 其中所有节点对之间的最大距离称为网络的直径. 如公式(2)所示.

$$L=\frac{\sum_{u\neq v} d(u,v)}{N(N-1)} \tag{2}$$

其中, 公式(2)的 v 表示为网络节点集合, N 表示为节点的总数, $d(uv)$ 表示为节点 u 和节点 v 之间的最短距离.

3) 介数: 节点的介数定量地描述某个节点在网络中的重要性或影响力. 在网络拓扑结构中, 每两个节点之间的最短路径组成一个最短路径的集合. 其中, 经过某一个节点 Vm 的最短路径的条数, 如公式(3)所示.

$$\beta=\sum_{j,k\in v}\frac{N_{jk}(i)}{N_{jk}} \tag{3}$$

其中, 公式(3)中的 N_{jk} 表示为网络节点 j 和网络节点 k 之间的最短路径的个数, $N_{jk}(i)$ 表示为网络节点 j 和网络节点 k 之间的最短路径中经过网络节点 i 的个数. 把网络节点 i 对所有节点对的贡献累加起来再除以节点对总数, 就可以得到网络节点 i 的介数.

通过复杂网络拓扑结构分析可以描述整个网络结构及其构成要素, 根据复杂网络拓扑结构的度量指标以网络的全局属性描述, 通过了解整个网络可以知道网络舆情的情况.

2　网络舆情信息文本挖掘的模型

根据网络舆情传播的复杂网络特征, 运用复杂网络的分析处理方法对网络舆情热点进行挖掘. 首先要对藏文网络中的信息进行采集, 然后对 Web 网页的抓取和对数据的存储. 通过用户在网络上对网页的访问频率的多少进行分析, 在一定程度上发现用户感兴趣的问题, 从而确定目前的热点网络舆情话题. 最后, 将藏文网络舆情热点话题根据关键词搜索, 并且根据每个热点话题进行聚类, 让管理者通过阅读这些话题可以了解正在发生或者已经发生的重要事件, 并提供自动追踪事件的能力, 帮助管理者快速的了解事件的全貌.

研究藏文网络的关键之处是如何结合藏文字、词、句各类形式特征来确定藏文分词[13], 它是研究藏文信息不可缺少的基础性工作. 对于藏文, 只有音节字、句和段可以通过分隔符来划界, 而词是没有分隔符的. 藏文作为拼音文字和二维的书写规则等特点, 使得分词又有别于汉语言分词. 分析藏文网络舆情先把收集的网络信息转换成文本形式, 再对文本进行处理, 其中包括对藏语言文本预处理、文本分词、特征提取、分类、聚类以及关联分析等操作.

网络舆情传播作为互联网的一个典型代表, 符合复杂网络中的一些典型拓扑特性. Web 信息挖掘是从网络信息的内容中发现有效的知识, 主要包括: 基于内容的网页分类、网络聚类、网页与网页之间内容的关联规则、主题相似度的计算和发现以及对网页进行特征提取等.

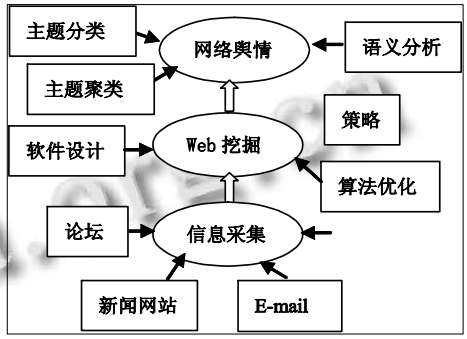

图 1　Web 信息挖掘的网络舆情模型

本文中运用的基于 Web 信息挖掘的藏文网络舆情挖掘分析模型如图 1 所示. 构造藏文网络舆情的传播模型步骤如下:

1) 假设网络中有 N 个节点, 则每个节点代表一个用户, 每个节点在[0, 1]之间随机赋予一个随机数作为初始状态, 即节点的初始意见值 $O(i)$, 初始时间步为 $t=0$.

2) 设定一个参数值 α. 遍历网络中的所有节点 i, 每个节点随机选择与之相关联的一个节点 j, 若 $|O_i-O_j|<\alpha$, 则两个节点在此时间步不需要移动, 即

用户意见需要更新.

3) 令 $t=t+1$，节点在每个时间步都与其所有相邻节点进行交互，若 $|O_i-O_j|>\alpha$，用户意见不需要更新.

4) 假设阈值为 10^{-3}，如果所有用户的意见变化之和大于阈值 10^{-3}，则转向步骤（1），否则结束.

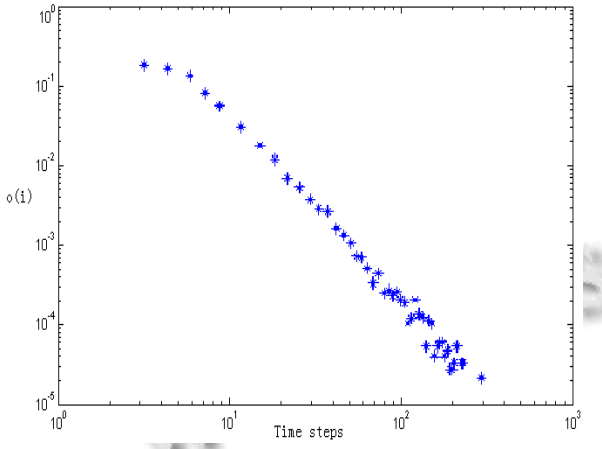

图 2 网络舆情传播模型

在查询的过程中，先将查询条件向量化，主要是根据布尔模型，最后可以计算出每个文本域查询的相似度，从而可以根据相似度的大小，将查询的结果进行排序，排序后的结果与设定的阈值进行比较，将所有计算结果控制在[0, 1]区间中.

计算任意网页文本之间的相似度运用了余弦相似度 Cosine Similarity PageRank (CSP) 分析算法来挖掘藏文网络舆情信息[14, 15].

$$sim(x,c)=\frac{\sum_{j=1}^{N} w_{jx}\times w_{jc}}{\sqrt{(\sum_{j=1}^{N} w^2{}_{jx})\times(\sum_{j=1}^{N} w_{jc}{}^2)}} \quad (4)$$

其中，$sim(x,c)$ 为网络新进来的网页文本 x 对于某一个事件聚类的相似度，N 表示为网页文本集中词的总个数，w_{jx} 表示集中词 j 在簇 c 的权重，θ 为两个向量之间的夹角. 相似度是指两个网页文本 x 和 c 之间的内容相关的程度，用它们之间的相似度 $sim(x,c)$ 来度量两个网页文本相关的程度.

因为藏文网络舆情信息数据量大，为了提高效率，采用网页清理技术，对收集到的网络信息进行分类、聚类和挖掘，按照 Web 网络主题组织信息，生成经过处理的一些有针对性的网络舆情信息.

3 结语

根据藏文网络舆情分析的特点，并通过复杂网络，Web 信息挖掘的知识能够有效地从网络上获取并分析相关的藏文网络舆情信息，对藏文网络舆情传播模型进行了理论分析和数值模拟，得出网络中的各种观点的人数比例与最初人群所占比例有关，最终达到预警的目的和引导的作用，为藏文网络舆情传播和预警提供了很大的帮助，方便相关部门采取有效措施进行预防控制，以此维护社会的安全和稳定.

参考文献

1 Stauffer D, Oliveira PMC. Simulation consensus model of never changed opinions in Sznajd Ceon- sensus model using multi-spin coding. Eprint, 2002, arxiv: cond-mat/0208296.

2 徐晓日.网络舆情事件的应急处理研究.华北电力大学学报(社会科学版),2007(1):89–92.

3 黄敏,胡学钢.基于复杂网络方法的舆情热点挖掘.计算机仿真,2007(1):11–12.

4 刘毅.略论网络舆情的概念、特点、表达与传播.理论界,2011,28(9):114–118.

5 刘毅.网络舆情研究概论.天津:天津人民出版社,2007.

6 牛康.社会传播学.福州:福建人民出版社,2001.344.

7 Jon.Authoritative sources in a hyperlinked envirn-onment. Journal of the ACM. 1999(5).

8 梅中玲.基于 Web 信息挖掘的网络舆情分析技术.中国人民公安大学学报(自然科学版),2007:4.

9 宋聚平,王永成.对网页 PageRank 算法的改进.上海交通大学学报,2003,37(3):397–400.

10 Junghoo C, Garcia-Molina H, Page L. Efficient crawling through URL ordering. Computer Networks and ISDN Systems,1998,30(1–7):161–172.

11 Chakrabart S, van den Beng M, Dom B. Focused crawling: a new approach to topic specific Web resource discovery. Computer Networks, 1999,31(11– 16):1623–1640.

12 刘剑宇.Web 挖掘技术在网络舆情预警中的研究与应用.四川警察学院学报,2009,21(3).

13 陈玉忠,李保利,俞士汶.藏文自动分词系统的设计与实现.中文信息学报,2003,17(3):15–19.

14 孟春艳.用于文本分类和文本聚类的特征抽取方法的研究.微计算机信息,2009,9.

15 丁杰,徐俊刚.IPSMS:一个网络舆情监控系统的设计与实现.计算机应用与软件,2010,27(4):188–190.