doi: 103969/j. issn. 0490-6756. 2016. 11. 009

基于复杂网络重叠社团发现的微博话题检测

尹 兰[1,2], 程 飞[1], 任亚峰[1], 姬东鸿[1]
(1. 武汉大学计算机学院, 武汉 430072; 2. 贵州师范大学大数据与计算机科学学院, 贵阳 550001)

摘 要: 社交媒体话题检测一直是个热点问题,由于社交数据杂乱异构,且具有时效性,语义模糊性等特点,话题检测也是个难点问题. 研究利用复杂网络对社交文本数据进行建模,并结合一种基于极大团凝聚层次聚类的重叠社团发现方法实现了社交话题的检测. 文本数据建模中,通过自定义突发系数量化话题词,即把话题词看作具有时域分布偏好的关键词,并通过自定义相关系数连接话题词,构建话题网络. 为使自定义系数更适用于动态数据环境,实验结合真实数据进行了适应性测试优化系数. 文章把采用 EAGLE 重叠社团发现方法在公开数据集上评测,根据 Q 函数值显示结果明显优于当前一些重叠社团发现策略,研究对采样的 60 万条青少年社交数据进行了话题分析并可视化了分析结果.

关键词: 复杂网络; 重叠社团发现; 话题检测; 青少年
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2016)06-1233-08

Topic detection based on overlapping community in complex network

YIN Lan[1,2], *CHENG Fei*[1], *REN Ya-Feng*[1], *JI Dong-Hong*[1]
(1. School of Computer Science, Wuhan University, Wuhan 430072, China;
2. School of Big Data and Computer Science, Guizhou Normal University, Guizhou 550001, China)

Abstract: Topic detection in social media is a hot yet challenging issue in social computing given most data there are heterogeneous, time-evolving and linguistically ambiguous. In this paper, the authors explore the idea of achieving this goal through complex network modeling which has demonstrated excellent interpretability of the real world. Specifically, a complex network was constructed based on preprocessed topic words where two parameters, namely the emergency and correlation coefficients, were also introduced to allow us to filter social data through the network as well as determine their possible correlations. This approach was then applied to analyze 600,000 messages by teenager users in Weibo.com to identify overlapping communities with the help of the well-established algorithm EAGLE. It was demonstrated that, compared to other popular approaches such as CONGO and Peacock a much better Q-value results has been obtained by the method proposed here.

Keywords: Complex network; Overlapping community discovery; Topic detection; Teenagers

1 引 言

社交话题的检测一直是个热点问题,由于社交网络数据杂乱异构,数据常常具有时效性,突发性和模糊性等特点,加之中文语义切分歧义性等复杂特点,中文社交话题的检测一直是个难点问题[1].

青少年作为社交网络的原住民,其社交生活极

收稿日期: 2015-11-12
基金项目: 国家自然科学基金(61133012,61373108); 贵州省科技厅联合基金(黔科合 J 字 LKS201237)
作者简介: 尹兰(1979-), 女, 副教授, 博士生, 研究方向为自然语言处理, 复杂网络. E-mail:yl@gznu.edu.cn

具影响力,因此也受到各方面的关注.社交网络对青少年信息传播,心理健康,人格调节等具有重要的影响[2].因此,对其社交群体生活进行关注具有深刻的社会现实意义.

话题模型是自然语言处理领域一个关键的应用模型,经典模型有PLSA(Probabilistic Latent Semantic Analysis)[3,4]和LDA(Latent Dirichlet Allocation)[5].PLSA基于多项式分布和条件分布混合建模词和文档的共现概率,其主要思想是建立在传统的潜在语义分析基础上利用EM算法进行参数学习.LDA模型是一个概率生成的贝叶斯模型,其基本思想是假设一篇文档的具体内容信息能够由一些潜在话题的多项式分布来表征,而话题又能够由一系列词的多项式分布来表征,文档由一系列语义上相关的词语和在某些话题中出现的概率来表征.PLSA和LDA都是建立在概率独立性假设的基础上,即植根于传统的词袋(Bag of Words)理论.因此,不可避免忽略了文本语义的复杂关系.LDA模型在话题发现领域被广为应用及改进[6,7].同时,受Google的PageRank启发下的TextRank[8,9]及各种图方法[10]、实体链接[11]方法被提出并应用到各类文本计算任务.在具体媒体环境中,结合真实应用场景的话题模型也得到相关研究[12].

总的说来,现行话题模型的问题在于复杂场景下,很难揭示文本数据复杂多变的语义关系,因此利用复杂网络对文本语义进行数据建模,并揭示其中潜在的丰富关系是一项重要的探索任务.研究利用复杂网络模型对社交数据进行话题建模,把话题定义为具有时间偏好的主题词网络集合,结合复杂网络中重叠层次社团发现算法进行社交话题检测.研究采用了EAGLE(agglomerativEhierarchicALclusterinG based on maximaLcliquE)算法,一种基于极大团凝聚层次聚类的重叠社团发现算法,实现了基于复杂网络重叠社团的话题检测,算法在公开网络数据集(http://www-personal.umich.edu/~mejn/netdata/)上选取开放数据集进行了相关评测,结果均优于流行的一些重叠社团发现策略.

研究以青少年社交数据为案例,从开放微博获取60万条真实青少年社交文本,对其进行数据清洗,话题词抽取,并结合提出策略实现了基于重叠社团发现的话题检测.

2 复杂网络及重叠社团发现

复杂网络,简言之就是呈现复杂性的网络,其复杂性往往体现在网络具有小世界,长尾分布,无标度,社区结构等属性,复杂网络因其对现实世界具有很好的解释性,得到了各个领域的广泛应用和研究.本节对复杂网络、社团发现及研究所采用的重叠层次社团发现算法进行简述.

2.1 复杂网络及社团发现相关

复杂网络一直是科学研究的重要议题[13-16],近年来,随着社交网络的兴起,网络分析在数据科学领域得到广泛应用,典型研究如斯坦福大学Jure Leskovec的网络分析项目(http://snap.stanford.edu/).与此同时,国内相关学者把复杂网络看作一门新型的交叉学科展开研究[17,18].通常把复杂网络映射到社交网络上进行网络研究[19],然而,充分利用复杂网络进行中文文本语义关系研究进而有效检测相关话题仍然是一项具有重要理论意义及实践价值的研究.

复杂网络作为复杂关系的度量模型具有很好的解释性.而其中社团结构是一个重要的属性,其表现为社团内部节点联系紧密,而社团间的节点联系相对稀疏.有效的检测社团结构是认识复杂网络的关键.社团发现具体就是找到网络划分或网络中高密度边的节点集.经典算法有GN算法[20],算法通过删除边界数进行网络划分,并引入了模块度函数度量衡量网络划分质量[21],高模块度意味着明显的社团结构.除此之外,常见的社团发现方法有建立在谱图划分理论上的矩阵分析[22,23],根据特定图矩阵的特征向量导出对象特征,并利用导出特征来推断对象之间的结构关系.此外,从信息论角度求解拓扑结构的有损压缩[24,25]或根据复杂网络中关联对象进行标签传播[26]来实现社团发现都有相关研究.总体说来,复杂网络社团发现研究主要考虑自底向上的凝聚(Agglomerative)和自顶向下的分裂(Divisive)两种策略.但常规的社团发现方法不能很好的解释复杂网络中重叠的层次的社团结构.常见的Newman算法虽可以通过将其进一步应用到首次划分得到的子社团中,从而得到层次性结构,但不能够揭示社团之间的重叠关系.而传统的一些k-cliques算法虽然可以揭示社团之间的重叠关系,但不能够得到社团之间的层次性关系.

在真实的社交话题网络中,话题常常在时间序列上呈现复杂的重叠性,即话题结点可能分布于一个或多个话题社团,因此,本文把社交话题检测看作一个重叠层次社团发现的过程,采取一种基于极大团凝聚层次聚类的方法进行话题检测.参照EA-

GLE[27]（Agglomerative Ehierarchic ALclusterinG based on maximaLcliquE)算法，本文在构造的话题网络基础上揭示了社交话题的重叠层次关系.

2.2　本文采用的重叠层次社团发现方法

几个关键概念定义如下.

定义 1　最大团. 给定网络 $G=(V,E)$，设 V' 是该图的顶点集 V 的一个子集，若 V' 中的任何两个节点在图 G 中均相邻，则称 V' 为该图的一个团. 假如图 G 不包含适合 $|V''|>|V'|$ 的独立集 V'，则称 V' 为该图的极大团，其中 $|V'|$ 为图 G 的团数.

说明：最大团可解释为网络中连接最为紧密的一组节点，不能被任何一个更大的团包含，通常表现为顶点集 V 中取 K 个顶点，其两两间有边连接.

定义 2　模块度. 给定网络 $G=(V,E)$ 有 $|E|$ 条边，则社团 C 可利用模块度 Q 来度量社团结构紧密程度，Q 值越大，则表明网络具有很强的社区结构，其计算方法如式(1)所示.

$$Q(C)=\sum_{i=1}^{t}(e_{ii}-a_i^2)=Tr(e)-\|e^2\| \quad (1)$$

把整个图 C 分割成 t 个社团，e 表示对称矩阵，e_{ij} 指的是社团 i 与社团 j 边的条数除以整个图所有边的条数得到的值. 其中，$a_i=\sum_{j-1}^{t}e_{ij}$（t 为矩阵的阶数）. $Tr(e)=\sum_i e_{ii}$ 表示矩阵的迹.

说明：本文利用模块度作为衡量社团紧密程度的量化指标.

定义 3　社团相似度. 用度量，C_1，C_2 两个社团的相似度公式如式(2)所示.

$$M=\frac{1}{2m}\sum_{v\in C_2,w\in C_2,v\neq w}\left[A_{vw}-\frac{k_v k_w}{2m}\right] \quad (2)$$

其中，A_{vw} 即网络邻接矩阵，表示点 v 和点 w 之间是否有边，没有边标记为 0，有边则标记为 1；m 表示网络中的边数 $m=\frac{1}{2}\sum_{vw}A_{vw}$；$k_v$ 表示点 v 的度数；k_w 表示点 w 的度数.

说明：社团相似度反映了两个社团之间的链接紧密程度，本例中相似度高的社团往往可以合并成一个更大的社团.

定义 4　扩展模块度. 扩展模块度是用来评估重叠社团分解的效果的量化指标，度量公式如式(3)所示.

$$EQ=\frac{1}{2m}\sum_i\sum_{v\in C_i,w\in C_i}\frac{1}{O_v O_w}\left[A_{vw}-\frac{k_v k_w}{2m}\right] \quad (3)$$

其中，m，A_{vw}，k_v，k_w 同上；C_i 表示第 i 个社团；O_v 表示点 v 所属的社团个数；O_w 表示点 w 所属的社团个数；A_{vw} 表示点 v 和点 w 之间是否有边，表示点 v 的度数，表示点 w 的度数.

算法步骤如下.

步骤 1　利用 Bron-Kerbosch 算法[28]计算出复杂网络中大于或等于指定阈值的最大团；

步骤 2 合并最相似社团并计算合并前后扩展模块度 EQ，输出使得 EQ 值最大的社团发现结果；

步骤 3 重复步骤 2 直至合并为一个社团.

其中 Bron-Kerbosch 算法是无向图中计算图最大连通分量的一种方法，其通过定义 R，P，X 三个集合，R，X 为空集，P 为所有定点集，算法依次从 P 中递归回溯搜索，最后返回最大团 R，其伪代码如下.

BronKerbosch(R, P, X)：
if P and X are both empty：
report R as a maximal clique
for each vertex v in P：
BronKerbosch1($R\cup\{v\}$, $P\cap N(v)$, $X\cap N(v)$)
$P:=P\backslash\{v\}$
$X:=X\cup\{v\}$

对从集合 P 中获得的每个节点 $\{v\}$，有如下的处理步骤.

步骤 1 将顶点 $\{v\}$ 加到集合 R 中，集合 P、集合 X 分别与顶点 $\{v\}$ 的邻接顶点集合 $N(v)$ 取交集，之后递归新的集合 R、P、X.

步骤 2 从集合 P 中去掉节点 $\{v\}$，并将节点 $\{v\}$ 加进集合 X 中.

若集合 P、X 都为空，则集合 R 即为最大团. 算法是每次从集合 P 中取节点 v 后，然后在 $P\cap N\{v\}$ 的集合中取，总是获得邻接的节点，保证所得到的集合 R 中任意两个顶点之间都相邻.

算法把社团看作较剩余网络更具关联性的点集，即社团有更高的链接密度. 算法结合 EQ 度量值，呈现复杂网络的层次社团关系.

3　社交话题复杂网络建模

对杂乱异构的社交数据清洗工作繁重且关键，研究通过主题词过滤筛选出大量关键结点词汇进行相关性度量构建复杂网络. 社交话题常常具有很强的时间敏感性，其相关传播模型已得到广泛关注[29, 30]，但就话题的突发性度量指标很难实现且没有统一的标准. 本文提出采用话题突发系数偏好设置进行词汇结点度量.

3.1　数据预处理及网络初始化

考虑社交数据的具体应用场景，文章对社交文本进行了数据清洗，本文利用 Python 语言进行文

本处理,采用结巴分词(https://github.com/fxsjy/jieba)对文本进行了分词和词性标注(注:案例中网络结点词为名词及关键动词).

数据经过预处理后网络初始化的两个关键点是主题词选择及相关性度量.由于社交数据常常具有突发性特点,而就"突发"的度量本身是个难点,也没有参照标准,时间参数作为一个关键指标已得到社交数据计算的应用[31],社交话题通常被看作一个时间序列函数,如 Kleinberg 的突发模型[32],即突发主要考虑两个指标:消息队列的时间间隔参考和消息呈现的内容.图 1 粗略勾勒了常规社交话题突发的一个过程,但由于社交数据及中文信息语义演化的复杂性,话题的演化很难模拟.

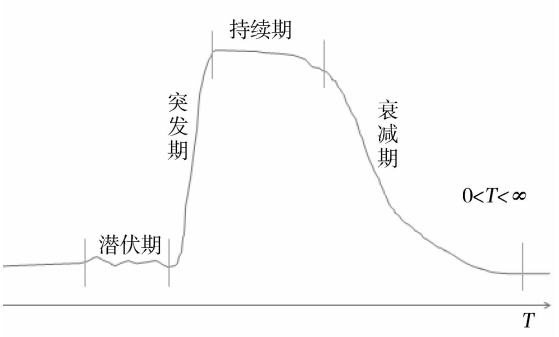

图 1 一般话题突发分布

Fig. 1 Normal topic life span

本文简化了度量,把话题词看作一个具有时间系数的网络结点词,在充分考虑时间间隔基础上,利用话题词的时域分布设计了话题词 i 的突发系数,并自定义函数如式(4)所示.

$$G_i = \frac{\max_j(F_{ij}/F_j)}{(\sum_j^k F_{ij}/F_j - \max_j F_{ij}/F_j))/(T-1)} \quad (4)$$

其中,T 为话题测量的时间范围,可根据粒度定义成天或小时,本案例设定为天;F_{ij} 为第 j 个时间窗口中第 i 个词的微博量;F_j 指的是在第 j 个时间窗口中所有的微博量;$\max_j(\frac{F_{ij}}{F_j})$ 为频率最高的一天的量.设置目的是为了避免不同时间窗口微博条数不同导致话题词出现分析误差.

结合主题词选取,我们自定义词向量和相关系数,进行了话题网络构建.词向量定义如式(5)所示.

$$w_i \rightarrow (f_{f1}, \cdots, f_{if} \cdots f_{ik}) \quad (5)$$

式(5)中,w_i 为选取的第 i 个词;k 为所有词的个数;f_{ij} 表示微博语料中第 i 个和第 j 个词的共现数.

根据选取的结点词,之间的相关度计算式如式(6)所示计算得出.

$$R(w_i, w_j) = \frac{\sum_{t=1}^{k} f_{it} \times f_{jt}}{\sqrt{\sum_{t=1}^{k} f_{it}^2 \sum_{t=1}^{k} f_{jt}^2}} \quad (6)$$

式(6)中,f_{it} 表示第 i 个词和第 t 个词共现的微博条数;f_{jt} 表示第 j 个词和第 t 个词共现的微博条数;k 表示整个语料中的词汇总量.

为考虑自定义突发系数及相关系数的适应性,结合实验真实语料,我们进行了系数适应值测试,最终选取突发系数为 3,相关系数为 0.4 作为阈值.

4 实验及相关结果

相关实验主要分成如下三个部分.

(1) 在公开数据集上进行重叠社团发现算法评测;

(2) 结合真实数据对自定义系数进行适应值测试,以便选取适配的网络相关系数构建话题网络;

(3) 结合构建的青少年话题网络实现话题检测.

4.1 社团发现算法测试及结果

实验对比了本文采用的 EAGLE 算法同当前一些社团发现算法 CONGO[33] 和 Peacock[34] 实施重叠社团发现实验,并在开放的 Karate 和 Dolphin 数据集上进行了对比.实验结果通过 Q 函数值进行度量,其中 Q 函数值越大,表示社团发现的结果越好,实验效果明显,验证了选取的重叠社团发现算法的可靠性.实验结果如表 1 所示.

表 1 重叠社团度量

Tab. 1 Overlapping communities detection strategies

数据集	CONGO-2	CONGO-3	BGLL+Peacock	Infomap+Peacock	EAGLE
Karate	0.2449	0.2575	0.2635	0.2203	0.3106
Dolphin	0.131	0.3173	0.2930	0.2614	0.4075

数据集说明：Karate 数据集是一个空手道俱乐部成员关系网络. Wayne Zachary 从 1970 年～1972 年间，根据对美国一所大学里空手道俱乐部成员之间的社会关系的长期观察而得. Karate 复杂网络总共包括 34 个节点以及 78 条边. Dolphin 数据集是一个海豚关系网络. Lusseau 等人对生活在新西兰 Doubtful Sound 峡湾一个宽吻海豚的群体长达 7 年的观察而得. 包括 62 个节点和 159 条边.

算法说明：CONGO-2，CONGO-3 表示 CONGO 算法中分列中介度参数设置为 2 和 3，BGLL+Peacock，Infomap+Peacock 表示 Peacock 算法混合 BGLL 算法和 Infomap 算法从而实现重叠社团发现.

4.2　突发系数及相关系数适应性测试实验及结果

为了使本文定义的突发系数及相关性系数在真实语料中更具适应性，实验结合真实语料，随机选取了 100 个关键词作为测试样本，使用 0，1 标记进行话题词判定，并进行人工标注，结合提出的系数度量方法，进行计算. 根据计算结果算出准确率，召回率和 F-Score 三项检测指标.

对突发系数适应值测试，本例中设定了 1～5 个阈值查看话题词挖掘的指标变化，具体结果如图 2 所示，同理，设定相关系数在 0～1 之间进行取值，根据随机选取的关键词计算出的准确率，召回率和 F-score 进行了相关系数的适应值测试，实验结果如图 3 所示.

结合实验结果，相对取整，最终选取突发系数值为 3，相关系数为 0.4 进行语料的话题复杂网络构建.

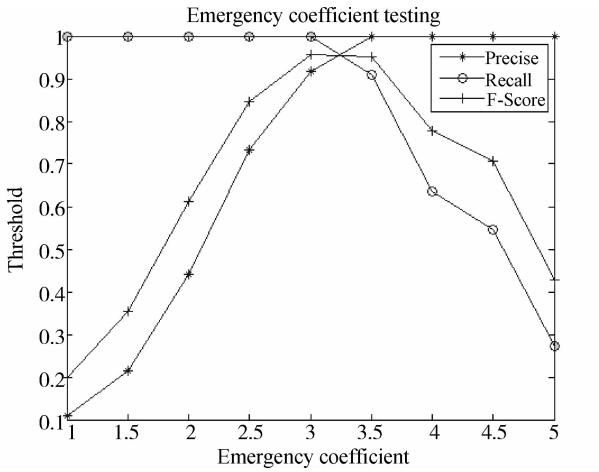

图 2　突发系数适应值测试

Fig. 2　Emergency coefficients testing

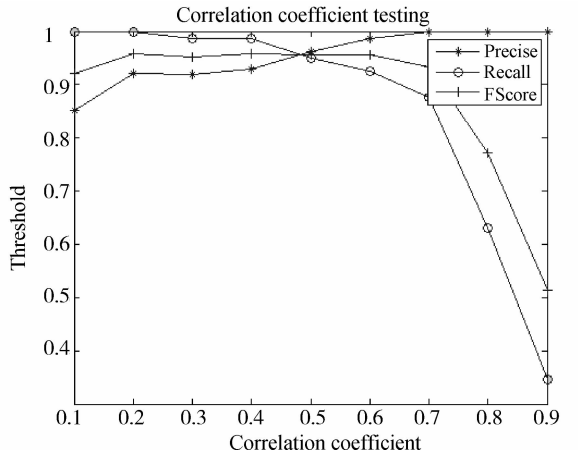

图 3　相关系数适应值测试

Fig. 3　Correlation coefficients testing

4.3　青少年社交话题检测实验及结果

实验案例研究中，我们获取了新浪微薄 2014 年 6 月，7 月两个月抽样青少年用户约 60 万条微博记录，并进行了数据清洗（其中包括一些社交网络页面中无用数据的过滤），进行了分词，词性标注，词抽取，突发系数设置，结合提出的方法进行了社交话题网络构建，利用 EAGLE 重叠社团发现策略实现了复杂网络重叠社团的话题检测.

其中，算法扩展模块度 EQ 值的变化如图 4 所示. 在话题社团数（见图 4 中横坐标所示）为 10，EQ 值最大（见图 4 中纵坐标所示，具体值为0.504136160152).

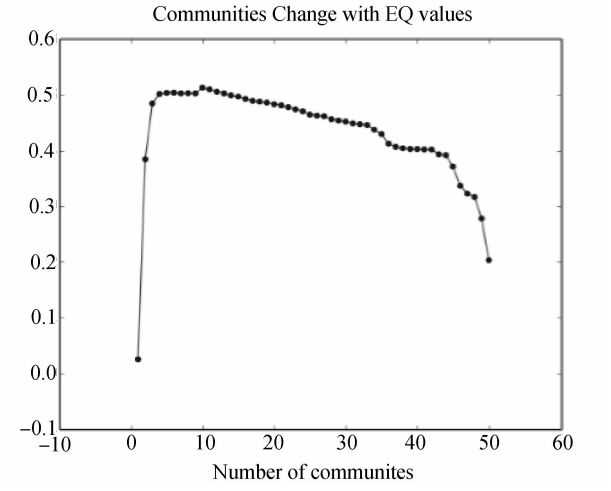

图 4　社团及 EQ 值变化

Fig. 4　Communities Change with EQ value

因此，利用 EQ 值进行数据切分，系统选择在最高 EQ 值下的话题输出的 10 个小社团对应的话题结果如表 2 所示.

表 2 话题检测结果

Tab. 2 Topic results

序号	检测结果
1	做人、别太、行星
2	亲亲、可怜、花心
3	儿子、看看、结果
4	得意、起来、还有
5	TFBOYS、sEYuVh、专属、加入、口袋、四叶草、城堡、守护、屁屁、没有、装进、魔法
6	cn、http、zYFwdTq、不想、中考、人生、关注、加油、单词、参加、取消、小升初、弟弟、录制、斩学、时候、期末考试、没有、生活、百词、背单词、认证、词汇
7	TFBOYS、产品、体验、俱乐部、偷笑、分享、协助、品牌、四叶草、家教、希望、成立、支持、改善、易烊千玺、有点、步步高、王俊凯、王源、称号、粉丝、组合、肯定、谢谢、进行、酷哦、首席
8	世界杯、中考、准备、加油、可怜、大学、学霸、孩子、安心、小凯、少女、开始、弟弟、感觉、时候、有没有、期末考、期末考试、王俊凯、老师、考场、考试、能力、英语、还有、送祝福、预习
9	世界、呵呵、拜拜、期末考、王俊凯、老王、考完
10	不用、奥特曼、开心、成立、成绩、组合、考上、高考、鼓掌

从整体结果看,采样数据的青少年话题主要集中在考试,娱乐明星两大块. 需要说明的是案例中青少年话题由于其特殊性,一些结果不能直观被理解,同时,因为实验利用 Python 脚本实现,调用结巴工具中词性标注,其主要考虑词典词及利用传统的 Viterbi 算法进行词序列译码,所以针对中文短文本而言,预处理也存在一定的误差.

但有四类明显的话题社团得到发现,其中两个是围绕两个娱乐明星群体的话题(话题 5 和话题 7),一个关于英语学习的话题(话题 6),一个考试话题(话题 8). 研究利用了 Gephi 软件(http://gephi.github.io/)对话题结果进行了相关可视化如图 5 所示.

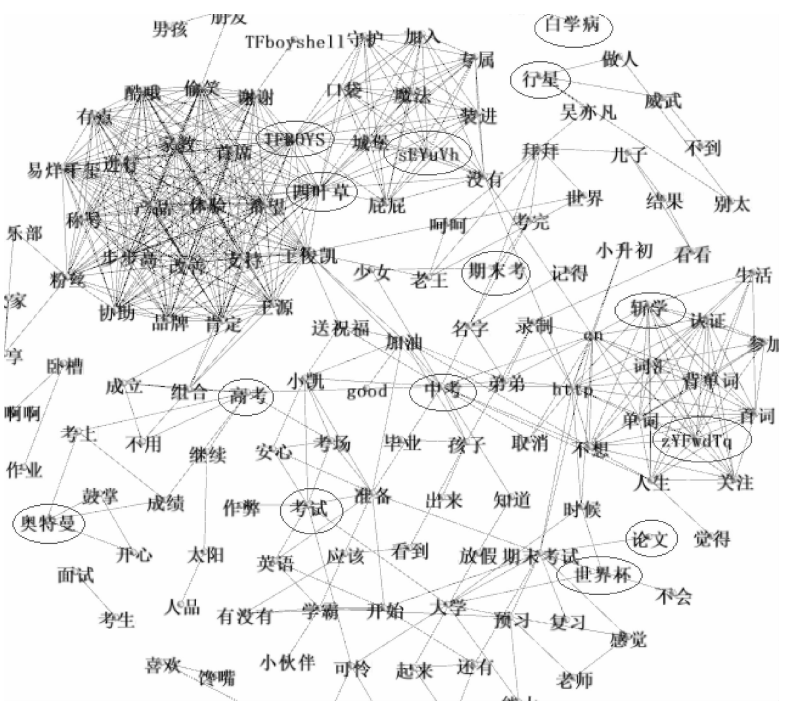

图 5 青少年微博话题复杂网络

Fig. 5 Teenagers' topic complex network

通过案例分析,可以得知"考试学习类"和"明星类"话题仍是青少年的主要话题,特别是考试,普遍存在青少年学习生活中,"论文","大学"关键词也提前呈现于我国青少年社交话题之列. 青少年关心的事件"世界杯"也迸发在话题之列,要说明的一点是在话题结果中,出现的一些 http,cn,zYF-

wdTq 等噪音数据，主要源于一些商业营销行为在公开微博上的商业推送所致. 而青少年话题网络中，高模块度的社团明显呈现于青少年偶像大 V，商业品牌营销构成的社交话题社团. 由此可见，微博作为一个开放环境娱乐导向一直是青少年社交的话题聚合的重要因素，而商业化营销的涌入也破坏了话题网络的网络结构.

青少年作为社交网络的原住民是网络新词的重要来源，在话题检测中，获取了诸如“奥特曼”，“白学病”等网络词汇. 因此，跟进复杂网络相关研究并有效的结合其在社交文本数据挖掘中的应用，展开相关社会科学研究具有重要的意义和研究前景.

5　结　论

考虑社交话题的复杂性和层次性呈现，本文利用复杂网络对社交话题建模研究进行了一次有益的实践，研究自定义了社交话题网络的构建方案，提出了话题词的突发性度量方法和相关性计算方案，为了使得自定义的突发系数和相关系数更适用于动态数据环境，本文结合真实语料进行了适应性测试优化了系数值. 最后结合复杂网络中有效的重叠社团发现策略进行了社交文本话题发现及可视化结果呈现.

利用复杂网络对社交媒体进行研究常常集中在信息传播方面，而利用复杂网络建模语言网络为网络科学打开了一个新视野，同时也为语言定量研究提供了潜在的方法论. 研究把社交语言看作一个多层重叠的复杂网络系统，对中文文本语义进行了网络建模，并结合相关算法进行了话题检测的初步实践和探索.

需要说明的是开放数据环境中社交网络的用户画像本身是一个难点，有效进行数据清洗，从社交网络获取可信的高质量实验数据仍然是一个关键问题；并且，考虑单机处理能力，实验只在开放大数据环境下进行了小样本数据实验，因此，更细时间粒度下的实效话题检测，网络算法性能提升等方面都需要进一步探索和实验，跟进复杂网络相关理论研究，并有效结合进行中文语义表示及文本挖掘有重要的研究意义和研究空间.

此外，当今青少年人群是伴随移动互联网络成长起来的一代，其真实生活常常直接映射到社交生活，其社交语言也是网络新词的重要来源，以该数据题材为案例研究为整合计算科学及语言学、社会学等学科融合提供了一个很好的研究场景.

参考文献：

[1] 吴少华，崔鑫，胡勇. 基于 SNA 的网络舆情演变分析方法[J]. 四川大学学报：工程科学版，2015，47(1)：138.

[2] 荆怀福，吴加强，苏园园，等. 青少年使用社交网络情况的调查与分析[J]. 北京青年工作研究，2013(11)：20.

[3] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning，2001，42(1-2)：177.

[4] Hofmann T. Probabilistic latent semantic analysis [J]. Proc of Uncertainty in Artificial Intelligence Uai'，2013，25(4)：289.

[5] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993.

[6] Perotte A，Bartlett N，Elhadad N，*et al*. Hierarchically supervised latent dirichlet allocation [J]. Advances in Neural Information Processing Systems，2011，24：2609.

[7] Liu Z，Zhang Y，Chang E Y，*et al*. PLDA+：Parallel latent dirichlet allocation with data placement and pipeline processing[J]. Acm Transactions on Intelligent Systems & Technology，2011，2(3)：389.

[8] Mihanlcea R，Tarau P. TextRank：bringing order into texts[J]. Unt Scholarly Works，2004，7：404.

[9] 李鹏，王斌，石志伟，等. Tag-TextRank：一种基于 Tag 的网页关键词抽取方法[J]. 计算机研究与发展，2012，49(11)：2344.

[10] Ioannis K，Mirella L. Unsupervised concept-to-text generation with hypergraphs[C]// Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Montreal，Quebec，Canada：Association for Computational Linguistics，2012.

[11] Han X，Sun L. An entity-topic model for entity linking[C]. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island，Korea：Association for Computational Linguistics，2012.

[12] Tsai F S. A tag-topic model for blog mining[J]. Expert systems with applications，2011，38(5)：5330.

[13] Milgram S. The small world problem[J]. Psychology Today，1967，2(1)：185.

[14] Jeffery T, Stanley M. An experimental study of the small world problem [J]. Sociometry, 1969, 32 (4): 425.

[15] Watts D J, Strogatz S H. Collective dynamics of ′small-world′ networks [J]. Nature, 1998, 393 (6684): 440.

[16] Barabasi A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286 (5439): 509.

[17] 方锦清, 汪小帆, 郑志刚, 等. 一门崭新的交叉科学: 网络科学(上)[J]. 物理学进展, 2007, 27 (3): 239.

[18] 解伯, 汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学, 2005, 2 (3): 1.

[19] 张艺. 基于复杂网络理论的社交网络研究[D]. 浙江: 浙江大学, 2012.

[20] Girnan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2002, 99(12): 7821.

[21] Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics (PRE), 2004, 69(2): 026113.

[22] Donetti L, Munoz M. Detecting network communities: a new systematic and efficient algorithm[J]. Journal of Statistical Mechanics Theory & Experiment (JSTAT), 2004, 2004(10): 10012.

[23] Capocci A, Servedio V D P, Caldaralli G, *et al*. Detecting communities in large networks[J]. Physica A Statistical Mechanics & Its Applications, 2005, 352(2-4): 669.

[24] Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks[J]. Proceedings of the National Academy of Sciences of the United States of America , 2007, 104(18): 7327.

[25] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics (PRE), 2009, 80(2): 2142.

[26] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics (PRE), 2007, 76(3): 1.

[27] Shen H, Cheng X, Cai K, *et al*. Detect overlapping and hierarchical community structure in networks [J]. Physica A Statistical Mechanics & Its Applications, 2009, 388(8): 1706.

[28] Bron C, Kerbosch J. Finding all cliques of an undirected graph (Algorithm 457) [J]. Communications of the Acm, 1983, 83(16): 575.

[29] 史亚光, 袁毅. 基于社交网络的信息传播模式探微[J]. 图书馆论坛, 2009, 29(6): 220.

[30] 张赛, 徐恪, 李海涛. 微博类社交网络中信息传播的测量与分析[J]. 西安交通大学学报, 2013, 47 (2): 124.

[31] Deng Z, Yan M, Sang J, *et al*. Twitter is faster: Personalized time-aware video recommendation from Twitter to YouTube [J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2015, 11(2): 31.

[32] Kleinberg J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373.

[33] Gregory S. A fast algorithm to find overlapping communities in networks[M]. Machine learning and knowledge discovery in databases. Lecture Notes In Artificial Intelligence(LNAI). Berlin: Springer Berlin Heidelberg, 2008: 408.

[34] Gregory S. Finding overlapping communities using disjoint community detection algorithms [M]. Complex networks. Lecture Notes In Artificial Intelligence(LNAI). Berlin: Springer Berlin Heidelberg, 2009: 47.