Using Complex Network Model for Online Comment Target Extraction and Identification in Opinion Mining

Tao Xu[1,a,*]

[1]Hangzhou Dianzi University, Hangzhou 310018, China

[a]txu@xjtu.edu.cn

*Corresponding author

Keywords: Online comments, Complex network, Opinion mining, Shortest path.

Abstract. Identification on the comment target contained in online comments plays a guiding role in opinion mining and sentiment analysis. This paper proposes a Directed-Weighted-Network-based model for modeling online comments, which aggregates important information from numerous comments as a whole object for research. Based on this network model, this paper further studies a candidate comment target set extraction algorithm based on network statistical features and an implement comment target identification algorithm for the comments containing no explicit comment target. A group of empirical experiments on public available English product reviews dataset and manually annotating Chinese news comments dataset are conducted. Experiment results show that the proposed algorithms can achieve satisfactory precision for comment target set extraction and comment target identification. It also shows that the proposed complex network model well captures key features from comment sets.

基于复杂网络模型的隐式评论对象识别方法

徐涛[1,a,*]

[1]杭州电子科技大学，杭州，中国

[a]txu@xjtu.edu.cn

[*]通讯作者

关键词: 在线评论; 复杂网络; 观点挖掘; 隐式评论对象; 最短路径

中文摘要. 评论对象识别对于从网络在线评论中挖掘评论者的观点、情绪具有重要的意义。本文提出了一种在线评论复杂网络模型，该模型能够将评论集中的重要信息聚集成一个整体作为研究对象。基于在线评论复杂网络模型，本文更进一步研究了基于网络统计特征的评论对象候选集抽取算法，以及基于网络最短路径的隐式评论对象识别算法。以公开的英文产品评论数据集及本实验室标注的中文新闻评论数据集为例，对所提出的方法进行了分析验证，实验结果表明本文提出的评论对象候选集抽取算法和隐式评论对象识别算法能够达到较高的抽取精确率和识别准确率，在线评论复杂网络模型很好的捕获了评论集中的关键性特征。

1. 引言

以购物网站商品评论、新闻报道跟帖评论、社交网络转发评论、在线社区讨论为代表的网络在线评论为普通民众提供了直接发表自己意见、观点、感受，以及抒发某种情绪的平台。网络上发表的大量在线评论为及时准确获取公众的观点和情绪特征提供了便利条件，因此如

何对网络在线评论进行有效地利用引起了越来越多研究者的关注。评论对象（comment target）识别旨在通过分析评论中自然语言词汇的词性、搭配关系、语义联系等特征来确定该条在线评论的讨论目标主题，也称之为话题（topic）、焦点（focus）、特征（feature）[1]。传统的评论对象识别（target identification）研究，如基于名词短语结构的相似性测试方法[2]、基于主题词与指示词的共现特征（co-occurrence）的方法[3]、基于互信息的方法[4]等，研究对象均是存在显式评论对象的评论数据，并且大多数研究面向的都是产品评论，而产品评论具有很强的领域性，评论对象一般都是领域内的术语。然而，现实中并非所有的评论都包含显式的评论对象，并且大量的关于突发性新闻事件的网络在线评论一般都不具备领域性，从而限制了以上方法应用的有效性。因此，从在线评论序列的内在整体特征出发，研究不需要借助于领域知识的隐式评论对象识别方法，将具有重要的现实意义和应用价值。

近年来，随着复杂网络模型研究的不断深入，人们以复杂网络模型研究了现实世界中的很多复杂性现象，其中基于复杂网络模型对自然语言领域一些现象的研究也逐渐被报道，如英语语言体系中存在的小世界特性[5]、单篇英语文章中存在的小世界特性[6]等。这些工作从系统科学的角度对自然语言研究做出了有益的尝试。网络在线评论通常主题鲜明、众人发表、观点之间交叉影响，整体上呈现的是众多人的集体行为，因此更值得从系统的角度对其进行深入研究。本文以复杂网络模型为基础，提出了一种在线评论的系统建模方法，并在系统模型的基础上研究了评论文本的评论对象候选集抽取、隐式评论对象识别的自然语言处理问题。

2. 在线评论复杂网络模型

孤立地分析一条评论，单条评论中通常包含评论对象（target）、针对评论对象的陈述（state）、意见持有者（holder）、以及评论者的情绪倾向（sentiment polarity）四部分内容[7]。其中评论对象、陈述通常在评论中显式可见，而后两者则蕴含在前两者中，通常并不显式可见。因此，单条评论文本可以按图1a所示方式抽象成由评论对象、围绕评论对象的陈述词构成的一条有向词链。而若将针对某一产品（事件）的所有评论抽象成有向词链后，利用节点间的共用关系则可以聚集在一起，最后从形式上呈现为不同评论对象、陈述词的网状交织，如图1b所示。

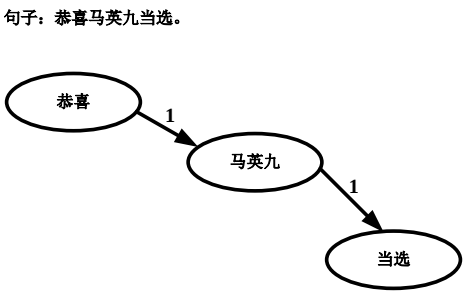

图1a 基于单条评论的有向词链

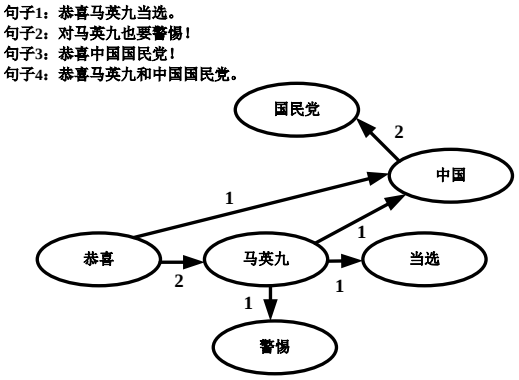

图1b 新闻在线评论集的网络模型示例

2.1 模型定义

为了将针对某一产品（事件）的所有在线评论聚集在一起进行系统研究，本文提出了一种基于词汇共现关系的有向网络模型，该有向网络模型在对单条评论抽象成有向词链的基础上，将所有评论汇织成网络，因此网络模型中包含了新闻在线评论的整体信息。为了方便叙述，首先给出如下了定义：

$\sum$：汉语词汇集，本文用到的词汇集为去除停用词、无意义实词后的汉语词汇集；

w：词，显然$w \in \sum$；

S：句子，句子由多个词按一定的顺序组成，即$S = w_1 \to w_2 \to \cdots$；

R：评论，评论由多个句子按一定的顺序组成，即 $R = S_1 \to S_2 \to \cdots$；

$G = (W, E, N_W, N_E)$：在线评论网络模型；

$W = \{w_1, w_2, \cdots, w_N\}$——在线评论网络模型的节点集合；

$E = \{< w_i, w_j >\}$——在线评论网络模型的边集合，其中 $< w_i, w_j >$ 表示从节点 w_i 指向节点 w_j 的有向边；

$N_W = \{n_1, n_2, \cdots, n_N\}$——在线评论网络模型中节点的权重集合；

$N_E = \{n_{i,j}\}$——在线评论网络模型中边的权重集合，其中 $n_{i,j}$ 表示有向边 $< w_i, w_j >$ 的权重。

2.2 在线评论网络模型生成算法

基于2.1节定义，本文提出的在线评论网络模型 G 的建立方法可描述如下：

Algorithm：NetworkCreate(R)

S1：对每一条评论 R 分句，得到一组有序的句子 $S_1 \to S_2 \to \cdots$；

S2：对每一个句子 S 分词，并去除停用词和无意义的实词，得到一组有序的词 $w_1 \to w_2 \to \cdots$；

S3：对每一个句子 S，采用2位滑动窗从句子中抽取出词汇对 $< w_i, w_j >$。若 $w_i \notin W$，则向 W 中添加一个新节点 w_i，并为节点 w_i 的权重 n_i 设初始值为1；否则 n_i 加1。对 w_j 的操作与 w_i 类似。若 $< w_i, w_j > \notin E$，则向 E 中添加一条新的有向边 $< w_i, w_j >$，并将权重 $n_{i,j}$ 设初始值为1；否则 $n_{i,j}$ 加1；

S4：整个评论集的完全网络模型建好以后，将节点权重 $n_i < n_{thr}$（n_{thr} 为一常数）的节点 w_i 及相邻的边删除，并修改相应的节点权重与边权重，从而降低网络模型的复杂性，及去除掉一些很少出现的词汇所产生的噪声；

S5：对每一条边 $< w_i, w_j >$ 计算*Jaccard*系数 J_{w_i, w_j}，如果 $J_{w_i, w_j} < J_{thr}$（J_{thr} 为一常数），则将边 $< w_i, w_j >$ 删除，并修改相应的节点权重与边权重，进一步降低网络模型的复杂；

图1b是基于上述算法生成的在线评论网络模型示意图。

3．隐式评论对象识别

统计显示，关于某一产品（事件）的网络在线评论通常都是围绕少量的热点评论对象展开，这部分热点评论对象与产品/事件本身相关，并且被不同的评论者反复提及。由于热点评论对象反映的众多评论者的集体行为，因此若考虑全体评论数据的整体信息，在评论序列整体模型的基础上识别评论对象，将可能获得优于针对单条评论的各种评论对象识别方法。

3.1 基于网络统计特征的评论对象候选集抽取算法

我们第2节提出的在线评论复杂网络模型实现了对评论序列整体信息的建模，本节将在网络模型的基础上研究基于网络统计特性的评论对象候选集抽取方法。为了描述本文提出的评论对象候选集抽取方法，下面是一组需要用到的定义：

定义1 节点间的扩展最短路径距离

$$d'(i,j) = \begin{cases} d_{\min}(i,j) & \text{如果节点} w_i, w_j \text{之间连通} \\ sum & \text{如果节点} w_i, w_j \text{之间不连通，} sum \text{为网络的总节点数} \end{cases}$$

式中：$d_{\min}(i,j)$ 表示两连通节点 w_i、w_j 之间的最短路径距离。

定义2节点对网络平均路径长度的贡献程度

$$LC_i = L_i^{'} - L_i \quad (1)$$

式中：$L_i^{'}$表示将节点w_i及w_i与其他节点的所有连线删除后，网络的平均路径长度，由于删除节点及连线后，连通网络可能变为非连通网络，所以对非连通网络计算$L_i^{'}$时应该采用节点间的扩展最短路径距离；L_i表示将节点w_i删除，但仍保留w_i与其它节点的所有连线时，网络的平均路径长度。

在已有的关于评论对象的抽取与识别研究中，都是将评论对象默认为名词或是名词性短语[2-4]。我们延续从名词中抽取评论对象思想，提出的评论对象候选集抽取方法可描述如下：

Algorithm：TargetExtract(*G*, *m*)

S1：对网络中所有名词节点的*LC*值、节点的权重n_{w_i}值及节点的出入度之和d_{w_i}值排序，得到三个列表L_c、L_n、L_d；

S2：计算所有名词节点在3个列表中的序号之和，并将其排序，获得列表L_{sum}；

S3：列表L_{sum}中前*m*个名词节点即为抽取出的评论对象候选集；

3.2 基于网络最短路径的隐式评论对象识别算法

由于在线评论具有高度口语化的特点，通常很多评论中并不显式地包含评论对象，但是这些评论仍然是围绕某一评论对象展开，因此对这些评论隐含的评论对象识别就需要用到推理的方法。在3.1节方法抽取出评论对象候选集的基础上，我们提出了基于在线评论复杂网络模型最短路径的隐式评论对象识别方法。

定义3 覆盖路径：对于任意评论语句$S = w_1 \to w_2 \to \cdots \to w_n$，如果在评论序列网络模型中存在一条路径顺序经过$w_1$、$w_2 \cdots w_n$节点（允许不连续经过上述节点的情况），则称该路径为覆盖路径，并记作$P(w_1, w_2 \cdots w_n)$。

定义4 覆盖路径距离：覆盖路径中经过的所有边的权重的倒数之和则称为覆盖路径距离，并记作$PL(w_1, w_2 \cdots w_n)$。

定义5 最优覆盖路径：给定节点集$w_1 \to w_2 \to \cdots \to w$的所有覆盖路径中，覆盖路径距离最小的路径就成为最优覆盖路径。

采用3.1节的方法抽取出评论对象候选集后，当某条评论语句$S = w_1 \to w_2 \to \cdots \to w_n$中不显式地包含评论对象时，可以按如下的网络最短路径法识别隐式的评论对象：

Algorithm：ImplementTargetIdentify (*G*，*S*)

S1：任意选取评论对象候选集中的某一评论对象，将其依次插入到w_1之前、w_1与w_2之间、w_2与w_3之间、……、w_{n-1}与w_n之间、w_n之后，分别计算出最优覆盖路径的距离，并将最小的最优覆盖路径的距离记录下来；

S2：若评论对象候选集中仍然有剩余对象没有进行S1步的计算，否则转入S3步；

S3：对所有评论对象对应的最小最优覆盖路径的距离进行比较，取最小值作为推理得出的隐式评论对象；

4．实验及结果分析

4.1 数据集及评价指标

为了验证本文算法的有效性，实验所用数据集来自University of Illinois at Chicago的Liu Bing等人所标注的英文产品评论数据（以下简称UIC数据集）[3, 9]，以及本实验室标注的突发性公共事件中文新闻评论数据。上述两个数据集在标注过程中，都对每条句子中涉及到的评论对象做了详细标注，数据集的其它详细信息见表1。

表1 实验数据集描述

数据集	评论数量	评论对象数量
Nokia 6610（UIC数据集）	546	67
Canon G3 （UIC数据集）	597	79
Jukebox Zen Xtra 40GB （UIC数据集）	1716	57
马英九当选台湾领导人 （本实验室标注）	4430	42
杭州飙车撞人事件（本实验室标注）	8618	47
江苏丰县校车事故事件（本实验室标注）	6603	55

注：统计过程中将英文数据集的每条评论句子视作一条评论

实验的评价指标设定为如下的评论对象抽取精确率和隐式评论对象识别准确率：

定义6 评论对象抽取精确率：

$$Precision = \frac{|A \cup B|}{|A|} \tag{2}$$

式中：A表示利用本文算法抽取出的评论对象候选集；B表示手工标定的评论对象集。

定义7 隐式评论对象识别准确率

$$Accuracy = \frac{count_{true}}{count_{total}} \tag{3}$$

式中：$count_{true}$表示正确推理出评论对象的评论数，$count_{total}$表示需要被推理的评论总数。

4.2 实验结果分析和讨论

为了测试隐式评论对象的识别准确率，本文实验中采取了两种策略来构造包含隐式评论对象的测试数据：

策略1：将包含显式评论对象数据中的评论对象手工去除从而获得隐式包含评论对象的测试数据；

策略2：人工挑选出本身没有显式评论对象的评论作为测试数据。

实验结果如表2所示

表2 评论对象候选集抽取精确率与隐式评论对象识别准确率

数据集	*Precision*									*Accuracy*	
	m = 5	10	15	20	25	30	35	40	45	策略1	策略2
Nokia 6610	1.00	0.90	0.87	0.90	0.88	0.80	0.74	0.65	0.56	0.672	0.408
Canon G3	1.00	0.90	0.93	0.95	0.92	0.83	0.74	0.65	0.58	0.681	0.453
Jukebox Zen Xtra 40GB	0.80	0.80	0.80	0.75	0.68	0.60	0.57	0.53	0.47	0.604	0.407
马英九当选台湾领导人	1.00	0.90	0.87	0.85	0.80	0.73	0.69	0.63	0.58	0.756	0.494
杭州飙车撞人事件	0.80	0.70	0.60	0.55	0.48	0.47	0.43	0.40	0.36	0.722	0.472
江苏丰县校车事故事件	0.80	0.70	0.60	0.50	0.44	0.40	0.37	0.33	0.31	0.689	0.444

从表2可以看出，本文提出的方法对名词节点综合考虑了其在评论序列网络模型中的关键性、在整个评论序列中的出现频率、在整个评论序列中与其它陈述词的搭配活跃度，因此将

其看作有重要意义的名词节点，并作为评论对象候选词具有一定的合理性。当预设的观点对象集规模较小时（≤10），方法抽取出的观点对象基本上都为真实的观点对象，当预设的观点对象集规模超过25以后，所有方法的抽取结果准确率将急剧下降。另外， 本文提出的方法在对隐式评论对象的识别过程中，利用复杂网络模型构建了全体评论的全局信息特征模型，并且综合考虑了待识别评论中词汇之间的搭配关系、语义联系，因此对隐式评论对象的推理识别具有一定的合理性。实验结果显示，中文评论数据集因为评论对象少，所以相对于因为数据取得更准确的识别效果，通过策略1构建的隐式评论对象由于手工去除，语义特征更明显，所以相对于本身不含评论对象的数据取得了更好的识别效果。

5．结论

网络在线评论作为一种新型文本，无论是在商业决策中的消费者态度评估，还是突发性公共事件管理中的舆情评估与控制，都具有重要的应用价值。目前已报导的在线评论研究主要集中产品评论挖掘，并且涉及到的评论对象研究主要集中于显式评论对象。本文提出了一种网络在线评论复杂网络建模方法，该模型将评论中的全局信息聚集成一个网络系统模型，并在系统模型的基础上，提出了基于网络统计特征的评论对象候选集抽取算法，以及基于网络最短路径的隐式评论对象识别算法。实验结果表明，利用本文提出的评论对象候选集抽取方法能获得令人满意的抽取效果，并且隐式评论对象的识别也能获得不错的效果。上述工作为情感分析和观点挖掘研究提供了坚实的基础。

致谢

本文为国家自然科学基金青年基金项目《基于认知特征的新闻事件在线评论观点自动摘要方法与社会情绪测量模型》(61402142)的阶段性成果之一。

References

[1] T. F. Yao, X. W. Cheng, F. Y. Xu. A Survey of Opinion Mining for Texts [J]. Journal of Chinese Information Processing, 2008, 22(3): 71-80.

[2] J. Yi, T. Nasukawa, R. Bunescu, etc. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques [A]. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003) [C]. Melbourne, Florida: 2003, 427-434.

[3] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews [A]. In: Proceedings of Nineteenth National Conference on Artificial Intellgience (AAAI-2004) [C]. San Jose, USA: 2004.

[4] A.M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews [A]. In: Proceedings of HLT-EMNL-05 [C]. Vancouver, Canada: 2005, 339-346.

[5] R. Ferrer-i-Cancho and R. V. Sole. The small world of human language [A]. In: Proceedings of the Royal Society of London. Series B, Biological Sciences [C] . 2001. 268 (1482) : 2261-2265.

[6] Y. Matsuo, Y. Ohsawa and M. Ishizuka. A document as a small world [A]. In : Proceedings the 5th World Multi-Conference on Systemics, Cybenetics and Infomatics [C]. 2001, 8: 410-414.

[7] S. M. Kim, E. Hovy. Determining the Sentiment of Opinions[C]. In: Proceedings of the Conference on Computational Linguistics (COLING-2004). Geneva, Switzerland: 2004, 1367-1373.

[8] H. Kautz, B. Selman, M.Shah. The hidden Web[J]. AI Magagine, 1997, 18(2): 27-36.

[9] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets