基于随机网络的在线评论情绪倾向性分类

杨 锋 [1, 2, 3] 彭勤科 [1, 2, 3] 徐 涛 [1, 2, 3]

摘 要 提出了一种基于随机网络的在线评论情绪倾向性分类模型 SCP-X (Shortest covering path-X). 首先引入了一种增量式创建词语顺序共现随机网络的方法, 并基于此随机网络以及情绪词表, 提出了一种基于评论序列最短覆盖路径 (Shortest covering path, SCP) 的情绪倾向性分类方法. 该方法具有以下两个优点: 1) 能够对相对短小、随意性较强、完整性较差的评论文本展开词语联想, 从而对完整性较差的评论数据进行属性值扩展; 2) 能够对评论文本的冗余属性进行约简, 约简后数据的属性规模为一般 VSM 模型的 10 % 左右. 本文最后设计了一组实验, 对以下算法进行了对比测试: TC, SVM, SCP-TC, SCP-SVM, SCP-HMM, SCP-Bayes. 结果表明本文提出的 SCP-X 方法对在线评论文本的倾向性分类效果更佳.

关键词 在线评论, 随机网络, 最短路径, 属性约简, 情绪倾向性

DOI 10.3724/SP.J.1004.2010.00837

Sentiment Classification for Online Comments Based on Random Network Theory

YANG Feng[1, 2, 3] PENG Qin-Ke[1, 2, 3] XU Tao[1, 2, 3]

Abstract We propose a new method of sentiment classification named SCP-X (shortest covering path-X) for online comment based on the random network theory. A new approach which is proved to be effective by experiments is presented to create the word co-occurrenced network incrementally. With the network, the sequences of online comments, which are shorter, more optional and more fragmentary, are extended by shortest covering path (SCP) proposed in this paper. Using this algorithm, the amount of attributes is reduced to about 10 % compared to VSM. Finally, experiments are designed to compare the results of the algorithms such as TC, SVM, SCP-TC, SCP-SVM, SCP-HMM, and SCP-Bayes. The results indicate that SCP-X is remarkably effective to classify online comments by sentiment orientation.

Key words Online comment, random network, shortest path, attributes reduction, sentiment orientation

近年来, 随着经济全球化和信息技术的飞速发展, 突发公共事件发生的频率、产生的影响和造成的财产损失逐渐增大, 突发公共事件的应急管理已经引起国内外政府和学术界的高度重视. 在线评论起源于某个公共事件或热点话题, 特点是反应快、内容短小精悍、口语化且有独特的非正规词语, 虽然有时在表达意见、想法与感受时比较激烈和主观, 但是能反映评论者对突发公共事件的直接反应和心理感受[1], 所以对其进行研究是突发公共事件信息获取及分析的重要方面. 国际上把 Blog、论坛和网络新闻在线评论等相关的文本称为新型文本[2−3] — 通常都是以短文本的形式出现的, 对新型文本的分析与处理正在成为当前研究的一个热点. 由于在突发公共事件的应急管理中, 情绪是分析民众应急反应的重要观察点, 其不仅能反映民众对风险的认知, 而且会影响民众的行为, 所以利用网络上的在线评论研究网民情绪的途径已经开启了突发公共事件信息分析的一个新方向. 在实际中, 人们的情绪是可以相互影响的, 在危急时刻这种情绪的传播特征更加显著[4]. 情绪影响和传播现象具体反映在网络在线评论中, 表现为一个表达某种情绪的评论会影响后续评论的情绪表达, 多个评论形成聚合效应, 对网民的情绪起到加强或缓解作用. 因此, 对评论文本的情绪倾向性研究作为突发公共事件应急管理研究的基础, 显得至关重要.

对短文本的倾向性研究, 比较成功的主要集中在产品评论[5−6] 和影评[7] 中, 对新闻评论情绪倾向性研究的效果还不够好[8−11]. 心理学研究发现, 词汇和人类情感之间的关系是可度量的, 独立的词汇或短语的语义倾向对于传达人类情感是重要的[12]. 有研究表明, 词汇和短语的语义倾向主要有两个现象: 1) 相同倾向的情感术语经常同时出现[6]; 2) 相反倾向的情感术语一般不同时出现[13]. 由于这两个现象的存在, 可以仅使用少量的种子词汇, 从一个大

收稿日期 2008-11-21 录用日期 2010-01-13
Manuscript received November 21, 2008; accepted January 13, 2010
国家高技术研究发展计划 (863 计划) (2007AA01Z475, 2007AA01Z464), 国家自然科学基金 (60774086), 教育部博士点基金 (20090201110027) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2007AA01Z475, 2007AA01Z464), National Natural Science Foundation of China (60774086), and the Ph. D. Programs Foundation of Ministry of Education of China (20090201110027)
1. 机械制造与系统工程国家重点实验室 (西安交通大学) 西安 710049 2. 智能网络与网络安全教育部重点实验室 西安 710049 3. 西安交通大学电子与信息工程学院自动化系 西安 710049
1. State Key Laboratory for Manufacturing Systems Engineering (Xi′an Jiaotong University), Xi′an 710049 2. Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, Xi′an 710049 3. School of Electronic and Information Engineering, Xi′an Jiaotong University, Xi′an 710049

型的语料库中, 通过迭代的方法扩充情感倾向词集[13−14]; WordNet 中词汇之间的距离可以揭示情感倾向的关系[15]; 同义词和反义词集也可以用来预测情感倾向[16]. 然而, 当考虑到 “不”、“很”、“有点” 这样本身没有情感倾向性, 但可以传递相反极性的情感倾向性或者改变情绪倾向强度的词语, 就需要同时关注陈述中多个元素之间的相互作用. 关于这方面的研究主要集中到两种方法上: 一种方法采用了手工构造语言模式或者赋予陈述极性的限制[17]或者专注于上下文各元素之间的相互作用[18], 这些方法中用到的语言模式和限制都需要大量人工信息的干预; 另一种方法采用了机器学习的方法, 采用多种方法提取情绪特征去训练情感统计模型[19].

有研究表明, 英文[20] 和中文[21] 的普通文本建立的词语共现网络都满足小世界特性, 并在此网络的基础上进行了文本分割和主题抽取方面的研究[21]. 文献 [22] 提出了一种基于词语依赖图最短路径的关系抽取方法, 并将此关系用于构建分类器的核函数. 但是, 将随机网络理论应用到短文本的情绪倾向性分类的研究中, 目前未见相关研究报道, 本文将致力于这方面的研究.

本文主要研究新闻在线评论短文本的属性约简和扩展方法, 以及情绪倾向性分类方法. 首先基于随机网络理论, 采用词语顺序共现图的方式对整个评论序列建立随机网络模型, 并分析了该网络的小世界特性. 然后利用网络对评论数据进行属性的约简和扩展, 使得评论数据具有足够且有效的语言特征, 能够更加精准地表示评论者的意见和语义倾向. 并在此基础上, 对在线评论序列进行情绪倾向性分类研究, 提出了一种将网络最短覆盖路径 (Shortest covering path, SCP) 算法和机器学习算法相结合的评论情绪倾向性分类模型 SCP-X (Shortest covering path-X). 该模型考虑了评论数据相对短小、随意性比较强、完整性比较差的特点, 使用 SCP 评论序列的属性进行约简和扩展, 然后用机器学习算法进行情绪倾向性分类. 最后, 在实验中将属性处理前后数据的测试结果进行了比较与分析.

1 基于词语顺序共现的在线评论随机网络模型

为了便于对短文本进行研究, 挖掘短文本词语之间的内在规律, 本文提出了一种新的建立词语顺序共现随机网络模型的方法.

1.1 建模方法

文献 [21] 中的方法是根据词语之间的普通共现关系建立随机网络模型, 这种普通的共现关系只能度量词语对在同一句子中共同出现的频数, 忽略了词语共现的顺序和距离; 而词语共现的顺序一般来说体现着语义方面的信息, 比如同位、修饰等, 词语的共现距离与词语的语义关系也有很大的关联. 本文是根据词语的紧密顺序共现关系建立模型, 即共现区域的窗口长度 WL 较小 (一般取 2), 而且考虑到词汇共现时的次序关系.

为了描述在线评论随机网络模型的建立方法, 下面是一组需要用到的数学定义:

$\sum$: 汉语词汇集, 本文使用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 $S = w_1 \rightarrow w_2 \rightarrow \cdots$;

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

$G = (W, E, N_W, N_E, D_E)$: 在线评论序列网络模型;

N: G 的节点个数;

M: G 的边的个数;

$W = \{w_i | i \in [1, N]\}$: G 的节点集合;

$E = \{(w_i, w_j) | w_i, w_j \in W$, 且 w_i 和 w_i 之间存在顺序共线关系 $\}$: G 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;

$N_W = \{n_{w_i} | w_i \in W\}$: G 中节点的权重;

$N_E = \left\{n_{w_i,w_j} | (w_i, w_j) \in E\right\}$: G 中边的权重, n_{w_i,w_j} 表示节点 w_i 与 w_j 之间边的权重;

$D_E = \left\{d_{w_i,w_j} | (w_i, w_j) \in E\right\}$: G 中边的长度, d_{w_i,w_j} 表示节点 w_i 与 w_j 的距离, 文中为了方便将相似权图转换为相异权图, 近似地取 d_{w_i,w_j} 为 n_{w_i,w_j} 的倒数.

下面给出在线评论序列网络模型 G 的建立方法:

1) 对每一条评论 R 进行分句, 得到一组有序的句子 $S_1 \rightarrow S_2 \rightarrow \cdots$.

2) 对每一个句子 S 进行分词, 并去除停用词和无意义的实词, 得到一组有序的词 $w_1 \rightarrow w_2 \rightarrow \cdots$.

3) 对每一个句子 S, 采用 WL (一般取 2) 位滑动窗从句子中抽取出词汇对 $\langle w_i, w_j \rangle$. 若 $w_i \notin W$, 则向 W 中添加一个新节点 w_i, 并为 w_i 的权重 n_{w_i} 设初始值为 1; 否则 n_{w_i} 加 1. 对 w_j 的操作与 w_i 类似. 若 $(w_i, w_j) \notin E$, 则向 E 中添加一条新边 (w_i, w_j), 并为 (w_i, w_j) 的权重 n_{w_i,w_j} 设初始值为 1; 否则 n_{w_i,w_j} 加 1.

4) 所有评论处理完成后, 网络模型 G 就建立完成了, 可以将节点权重 $n_{w_i} < n_{\text{thr}}$ (n_{thr} 是一参数) 的节点 w_i 及相邻的边删除, 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.

5) 对每一条边 (w_i, w_j) 计算 Jaccard 系数

J_{w_i,w_j}, 如果 $J_{w_i,w_j} < J_{\rm thr}$ ($J_{\rm thr}$ 是一参数), 则将边 (w_i, w_j) 删除, 从而进一步降低网络模型的复杂性. Jaccard 系数的计算公式为

$$J_{w_i,w_j} = \frac{n_{w_i,w_j}}{n_{w_i} + n_{w_j} - n_{w_i,w_j}}$$

6) 保留网络的最大弱联通子图, 删除其余节点和边 (为了满足节点的可达性, 方便计算最短路径).

7) 为了使用 FLOYD 算法计算最短路径, 需要将相似权网络转化为相异权网络, 即计算 $d_{w_i,w_j} = 1/n_{w_i,w_j}$.

由建网过程可以看出, 本文所提方法可以增量式地添加评论数据. 随着评论数据的增加, 可以逐步建立更加符合该评论序列统计特性的网络模型. 图 1 是本文实验中用到的《马英九当选台湾地区领导人》在线评论序列构建的网络模型. 其中调整模型参数 $n_{\rm thr}$ 和 $J_{\rm thr}$, 使得节点个数在 30 左右, 边数在 50 左右.

从图 1 中, 可以看出评论序列中大量出现的语句模式在网络中联系将比较紧密. 如: 热烈 → 祝贺 → 马英九 (→*同志/先生*)→ 当选; 实现 (→*祖国/中国*)→ 和平统一 → 大业; 等等. 对于网络新闻评论数据这种语法性较差、表述较随意的语言, 可以用图 1 所示的共现图来推理、完善数据, 分析数据的某些特性. 下面正是基于本节提出的词语共现图, 进行在线评论情绪倾向性分类算法研究.

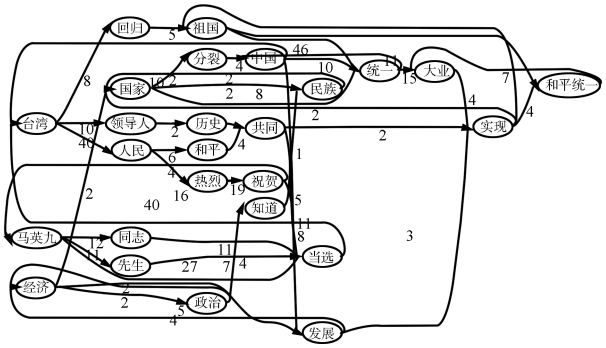

图 1　在线评论序列网络模型示意图

Fig. 1　An example of word co-occurrenced network

1.2　小世界网络特性分析

复杂网络中的两个主要模型是 Watts 与 Strogatz 提出的小世界网络模型[23] 和 Barabasi 与 Albert 提出的无标度网络模型[24]. 相对于规则网络和经典随机网络, 小世界网络和无标度网络具有以下两个显著特性: 小的平均路径长度、大的聚集系数. 文章 [20−21] 分别验证了英文和中文普通文本的小世界特性. 为了分析本文方法所构建网络模型的小世界特性, 需要先给出两个网络特性参数的定义:

定义 1 (平均路径长度 L). 网络 G 的平均路径长度 (也称特征路径长度), 描述了网络中节点间的分离程度, 如式 (1):

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N} d_{\min}(w_i, w_j) \tag{1}$$

其中, N 表示节点的数目; $d_{\min}(w_i, w_j)$ 表示两连通节点 w_i 和 w_j 之间的最短路径距离.

定义 2 (聚集系数 C). 聚集系数衡量了网络 G 中近邻节点间的联系紧密程度, 如式 (2):

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum\limits_{j=1}^{N}\xi_{i,j}\left[\sum\limits_{k\in\Gamma_i;k>j}\xi_{j,k}\right]}{C^2_{|\Gamma_i|}} \tag{2}$$

其中, N 表示节点的数目; $\xi_{i,j}$ 表示节点 w_i、w_j 之间是否存在边, 若存在边则 $\xi_{i,j} = 1$, 否则 $\xi_{i,j} = 0$; $\Gamma_i = \{j|\xi_{i,j} = 1\}$ 表示节点 w_i 的邻接节点集合; $|\Gamma_i|$ 表示集合 Γ_i 的元素个数.

记 $\bar{k} = M/N$, 表示网络节点的平均连结数; $L_r = \ln(N)/\ln(\bar{k})$ 表示随机网络的特征路径长度; $C_r = \bar{k}/N$ 表示随机网路的聚集系数; $\mu = (C/L)/(C_r/L_r)$ 表示小世界网络的衡量指标. 如果同时满足: $C << C_r$、$L \approx L_r$、$\mu << 1$, 则该网络是小世界网络[20−21].

表 1 和表 2 分别是利用本文网络和文献 [20] 网络对于两组数据的测试结果. 结果表明对于短文本序列, 本文构建网络的小世界特性更加显著. 词语顺序共现关系更加符合语法习惯, 经常使用的短语通常都是顺序共现的, 因此, 词语顺序共现网络在结构上比词语简单共现关系更加规则, 正如图 1 中所示, 评论中经常出现的短语模式, 在词语顺序共现网络上就会处于比较重要的位置, 具体表现就是 (词语共现边权重比较大, 距离比较近). 在一个新闻主题的系列评论文本中, 常用的短语模式少而集中, 因此, 通过词语共现网络就可以得到这些处于关键位置的短语模式, 与这些短语模式关系比较紧密的词语往往体现了群体对新闻的观点. 因此本文后续将应用该词语顺序共现网络进行评论情绪的倾向性分类研究.

2　在线评论情绪倾向性分类

根据短文本的特性 — 词语数量相对较少、随意性比较强、完整性比较差, 一般的基于情感词汇计数的 TC (Term count) 算法或者基于向量空间模型的支持向量机 (Support vector machine, SVM) 算法在进行短文本分析时显得力不从心, 因为属性值过少使得 TC 算法的效果变差, 而在向量空间模型中如

表 1 词语序列共线网络小世界特性验证结果 (本文方法)

Table 1 The empirical results about small-world property of order word co-occurrence network using the proposed method

	J_{thr}	n_{thr}	$\overline{k}$	N	M	L	L_r	C	C_r	μ
奥运火炬传递相关新闻评论	0.01	10	5.52	897	4 191	3.35	3.92	0.22	0.0062	40.02
5·12 地震相关新闻评论	0.01	10	5.68	914	4 280	3.36	3.92	0.23	0.0062	43.27

表 2 词语普通共现网络小世界特性验证结果 (文献 [20] 方法)

Table 2 The empirical results about small-world property of word co-occurrence network using the method in [20]

	J_{thr}	n_{thr}	$\overline{k}$	N	M	L	L_r	C	C_r	μ
奥运火炬传递相关新闻评论	–	–	6.43	674	4 460	3.92	3.14	0.33	0.019	14.25
5·12 地震相关新闻评论	–	–	7.26	664	4 820	3.59	2.93	0.29	0.022	10.73

果用情绪词作为属性, 则向量维数会很多, 而且会非常稀疏. 本文提出的网络模型体现出了在线评论文本中的词语顺序共现关系的统计特性, 因此, 对于一条特定的评论数据, 可以根据网络的统计特性来推理、扩展或者约简它的属性. 基于这种思想, 本节提出了一种属性约减和扩展算法 — 最短覆盖路径法 (Shortest covering path, SCP). SCP 算法结合 TC 或者机器学习算法就可以提高普通的情绪倾向性分类算法对评论情绪倾向性分类的准确率.

2.1 情绪词表

本文的情绪特征词以知网 (HowNet)[25] 的《情感分析用词语集》中的 "负面评价词语 (中文)" 和 "负面情感词语 (中文)" 作为消极情绪种子词 (4 370 个), 将 "正面评价词语 (中文)" 和 "正面情感词语 (中文)" 作为积极情绪种子词 (4 567 个), 根据《同义词林》[26], 在第五级上扩充, 最后去掉交叉词汇, 共得到 8 354 个积极词汇和 6 358 个消极词汇, 以此作为算法的情绪特征词集: PSD (Positive sentiment dictionary, 正面情感词集)、NSD (Negative sentiment dictionary 负面情感词集). 并用同样的办法, 扩展得到否定词集: PD (Privative dictionary).

2.2 基于评论最短覆盖路径的情绪倾向性分类方法

由于短文本本身的特性: 词语数量相对较少、随意性比较强、完整性比较差, 一般的基于倾向性词汇计数的 TC 算法[6] 或者基于向量空间模型的 SVM 算法[27] 在进行短文本分析时显得力不从心, 因为属性值过少使得 TC 算法的效果变差, 而在向量空间模型中如果用情绪词作为属性 (8 353 积极词汇 + 6 358 消极词汇 = 14 711 个), 则向量维数会很多而且会非常稀疏. 从对图 1、表 1 和表 2 的分析中可以看出, 该网络模型体现出了在线评论文本中的词语顺序共现关系的统计特性, 因此, 对于一条特定的评论数据, 可以根据网络的统计特性来推理、扩展或者约简它的属性. 用网络中节点代表的词语作为属性, 属性规模可以约简为 N 个, 文中一般调整参数使得 N 介于 600 至 1 500, 那么属性规模将会约简为 5 % 到 11 %. 基于这种思想, 本文提出了一种无监督的情绪倾向性的分类方法 — 最短覆盖路径法.

定义 3 (最短覆盖路径 S'). 网络 G 中, 顺序覆盖评论语句 S 中所有词语的路径中最短的路径. 如式 (3):

$$S' = \sum_{i=1}^{n-1} P_{\min}(w_i, w_{i+1}) \tag{3}$$

其中, $P_{\min}(w_i, w_{i+1})$ 是从 w_i 到 w_j 的最短路径. 记 $S' = w'_1 \to w'_2 \to \cdots \to w'_{n'-1} \to w'_{n'}$.

定义 4 (最短覆盖路径长度 $d_{\min}(S)$). 网络 G 中, 顺序覆盖评论语句 S 中所有词语的路径中最短的路径的长度, 即 S' 的长度. 计算公式如下:

$$d_{\min}(S) = d_{\min}(S')$$

$$d_{\min}(S) = \sum_{i=1}^{n-1} d_{\min}(w_i, w_{i+1})$$

$$d_{\min}(S') = \sum_{i=1}^{n'-1} d_{w'_i, w'_{i+1}}$$

其中, $d_{\min}(w_i, w_{i+1})$ 是路径 $P_{\min}(w_i, w_{i+1})$ 的长度.

根据上述定义, 下面将最短覆盖路径算法和 TC 算法结合起来, 给出 (SCP-TC) 情绪倾向性分类算法的步骤:

1) 对每一个句子 S 分词, 保留网络 G 中存在的词, 得到一个有序的词序列, 记为:

$$S = w_1 \to w_2 \to \cdots \to w_{n-1} \to w_n$$

2) 求解 S 的最短覆盖路径: $S' = w'_1 \to w'_2 \to \cdots \to w'_{n'-1} \to w'_{n'}$;

3) 结合情绪词表 PSD, NSD, PD, 对 S' 进行以下处理 (该步骤其实是 TC 算法):

a) 定义变量 C_{PSD}、C_{NSD}, 分别记录情绪词的个数, 初始值是 0;

b) 对于 S' 中的每个词 w'_i: 如果 $w'_i \in$ PSD, 则 $C_{\text{PSD}} = C_{\text{PSD}} + 1$; 如果 $w'_i \in$ NSD, 则 $C_{\text{NSD}} = C_{\text{NSD}} + 1$; 如果 $w'_i \in$ PD, 则下一个情绪词将取相反的极性.

c) 计算情绪倾向性因子 (Sentiment orientation factor, SOF):

$$SOF\,(S') = \begin{cases} \frac{(1-\lambda)C_{\text{PSD}} - \lambda C_{\text{NSD}}}{(1-\lambda)C_{\text{PSD}} + \lambda C_{\text{NSD}}}, & (1-\lambda)C_{\text{PSD}} + \lambda C_{\text{NSD}} \neq 0 \\ 0, & (1-\lambda)C_{\text{PSD}} + \lambda C_{\text{NSD}} = 0 \end{cases}$$

其中, $0 < \lambda < 1$ 表示负面倾向性权重 (取 0.6). 显然有: $SOF\,(S') \in [-1, +1]$. 句子 S 的情绪倾向性为式 (4):

$$SOF\,(S) = \begin{cases} 1, & SOF(S') > \delta \\ 0, & -\delta \leq SOF(S') \leq \delta \\ -1, & SOF(S') < -\delta \end{cases} \tag{4}$$

其中, $0 < \delta << 1$ 表示倾向性阈值.

评论文本比较短小, 且有很多数据格式不标准、语义信息不全、修辞手法特殊, 简单的分类算法不能取得好的结果. 本文提出的词语顺序共现网络具有小世界特性, 从某种意义上说是一个语义网络, 网络的边表示的是词对之间的有向共现关系, 代表了特定的修辞格式. 基于 SCP 的属性值扩展算法修复了词语之间某些特定的修辞格式, 尽可能使得一个评论句子的局部语义特性更加明显. 当然这种属性值扩展的方法可能导致样本差异减小或者分类器泛化性能的下降, 但是由于直接应用评论文本分类算法准确率并不高, 因此引入 SCP 后带来的负面影响不会很大. 后续实验结果也表明本文的分析是正确的.

2.3 基于 SCP-HMM 的情绪倾向性分类方法

HMM 在自然语言处理和模式识别等相关领域的研究中都有着出色的表现, 而且对于不定长样本序列可以进行训练, 鉴于评论序列的特点, 使用 HMM 对评论情绪倾向性进行学习, 以提高 SCP-TC 的预测准确率. 将此算法称为 SCP-HMM (Shortest covering path – hidden Markov model).

SCP-HMM 的训练方法如下:

1) 使用 SCP 算法对样本进行第一次情绪倾向性分类, 计算评论序列的情绪倾向性因子 (SOF).

2) 对于所有的样本: 如果 $SOF\,(S') > SOF_{\text{thr}}$ 则将 S 加入正例样本集 PS; 如果 $SOF\,(S') < -SOF_{\text{thr}}$ 则将 S 加入反例样本集 NS. 其中 SOF_{thr} 是表示样本的情绪倾向性显著性阈值的参数, 且 $SOF_{\text{thr}} \in (0, 1)$.

3) 用 PS 和 NS 作为训练集训练 HMM 分类模型. SCP-HMM 的预测方法如下:

a) 对于一条评论句子 S, 用 SCP 算法计算最短覆盖路径 S', 计算情绪倾向性因子 $SOF\,(S')$.

b) 如果 $|SOF\,(S')| < \delta$ $(0 < \delta << 1)$, 表示 S 情绪倾向性比较难以判断, 用 HMM 分类模型预测 S' 的情绪倾向性; 否则用式 (4) 预测情绪倾向性.

2.4 评论的情绪倾向性

前面所叙述的研究都是句子级别的, 本节将介绍基于句子级别的情绪倾向性来计算评论 $R = (S|S \in R)$ 的情绪倾向性. 基于评论中句子的情绪倾向性因子与情绪倾向性预测结果, 本文采用下面的公式计算评论的情绪倾向性因子为式 (5):

$$SOF\,(R) = \frac{1}{N_R} \sum_{i=1}^{N_R} |SOF\,(S_i)|\, SO\,(S_i) \tag{5}$$

同样, $SOF\,(R) \in [-1, +1]$. 评论 R 的情绪倾向性计算式为式 (6).

$$SO\,(R) = \begin{cases} 1, & SOF(R) > \delta \\ 0, & -\delta \leq SOF(R) \leq \delta \\ -1, & SOF(R) < -\delta \end{cases} \tag{6}$$

3 实验及结果分析

本节利用实验检验了本文所提出算法的有效性.

3.1 数据集与实验安排

本文选取 2008 年影响较大的新闻事件: 雪灾、奥运火炬传递、5·12 地震事件、拉萨 3·14 事件等 4 个事件的相关评论作为数据集来测试本文的算法. 每个事件, 选取其中的 6 000 条评论进行研究, 对评论进行分句, 删除不含情绪词的句子, 手工标定句子的情绪倾向性, 以此作为测试数据集. 然后对本文提出的 SCP-HMM 算法进行测试, 并与传统的情绪倾向性分类算法 (基于 TC、基于 SVM 的算法) 进行比较, 同时也测试了 SCP-Bayes, SCP-SVM 的分类效果. 实验采用十倍交叉验证, 取分类算法性能指标的平均值作为测试结果.

3.2 实验结果与结果分析

实验结果如表 3 所示, 由于篇幅的限制, 文中只给出了火炬传递和地震两个数据集的测试结果, 其他几个数据集也得到了相似的结果. 对网络参数的解释和建网参数的确定方法见表 3 的注释.

3.2.1 算法执行速度

TC 和 Bayes 算法具有最快的速度; 而 SVM 和 HMM 算法的建模过程比较慢, 且所有的 SCP-X 算法都需要建立网络和处理属性的时间; 在预测速度上, SVM 相对较慢, 而 HMM 和 TC 则比较快. SCP-X 算法比 X 算法执行时间少的原因主要是文中应用了 SCP 算法对数据属性进行了约减, 数据维数得到了极大的减少. 其中 SCP-HMM 比 HMM 执行时间长是因为 HMM 的参数选择不同对算法复杂度影响比较大.

3.2.2 算法效果比较

SCP-X 算法均比单纯的 X 算法的效果好, 准确率平均提高了约 8 个百分点, 这也证明了文中的共现网络的建立方法、基于共现网络的 SCP 样本扩展方法和 SCP-X 分类算法的思想是可行的和有效的. 主要是因为使用 SCP 算法对属性缺失严重的数据进行了属性值扩展, 使得这些评论文本的情绪倾向性变得比较显著, 分类准确率随之提高. 实验结果表明: 对于网络新闻评论, SCP-X 情绪倾向性分类算法是可行的和有效的. SVM 效果好于 TC, SCP-SVM 的效果好于 SCP-TC, 而 SCP-HMM 的效果与 SCP-SVM 相当.

3.2.3 算法消耗资源

SVM 算法是基于 VSM 模型的, 对于在线评论这种短文本数据来说, 样本向量会非常稀疏, 计算时会耗费大量的内存空间; 而 HMM 模型可以对变长的原始样本进行学习, 不用将数据表示成 VSM 模型, 对评论序列这种比较短而属性非常多的数据比较合适, 计算时不会因为样本的稀疏性问题而耗费额外的内存空间.

3.2.4 存在的问题

分类结果比较偏置, 比如: 大部分算法积极倾向性样本的召回率都比消极倾向性样本的召回率高很多, 相应的大部分算法的积极倾向性样本的精度都比消极倾向性样本的精度低很多. 这是由于分类结果中有大量的消极倾向性样本被错误的判断为积极倾向性样本. 导致这种结果有多种原因: 样本不对称、算法参数问题、算法设计问题、自然语言特有的性质. 当新闻评论中频繁使用句式: 反问、反语等时, 评论者大多具有消极情绪倾向. 因此需要重点研究如何采取措施有效的提高消极样本的分类效果.

由于小世界网络有着短路径和高聚集度的特性, 并且结合该网络的语义特性, 一个评论句子的网络最短覆盖路径 (SCP), 将具有修辞关系且具有高聚集度的词语对用最短路径联系起来,将网络的统计特

表 3 在线评论情绪倾向性分类实验结果

Table 3 The classification results about sentiment orientation of internet comments

	算法	建模时间 (s)	预测时间 (s)	准确率 (%)	积极情绪召回率 (%)	积极情绪准确率 (%)	消极情绪召回率 (%)	消极情绪准确率 (%)
	TC	0	0.01	57.20	63.80	56.36	50.60	58.29
奥运	SVM	1 292.17	341.94	79.05	61.20	93.42	95.92	72.33
火炬	HMM	6 272.63	10.77	79.21	77.99	81.75	81.72	77.96
传递	SCP-TC	31.33	0.01	85.95	86.44	86.81	85.38	85.24
相关	SCP-SVM	294.39	142.72	91.79	98.19	87.99	84.32	97.55
评论	SCP-HMM	7 605.45	15.20	86.10	93.14	80.23	77.87	92.12
	SCP-Bayes	4.93	0.79	87.00	88.93	87.21	84.72	86.81
	TC	0	0.01	63.53	65.51	63.02	61.57	64.09
5.12	SVM	1 262.23	364.13	69.14	83.75	69.20	49.28	69.02
地震	HMM	6 234.11	11.57	64.08	73.75	66.37	51.45	60.16
相关	SCP-TC	68.34	0.01	80.34	77.65	84.56	84.61	77.82
新闻	SCP-SVM	292.56	91.42	87.40	94.88	84.66	77.64	92.08
评论	SCP-HMM	7 825.30	12.22	82.52	94.83	76.09	70.20	93.14
	SCP-Bayes	2.28	0.43	84.07	89.53	83.46	76.88	85.01

注: 指定参数: $WL = 2$, $SOF_{\text{thr}} = 0.5$, $\delta = 0.05$, $\lambda = 0.6$

调节网络参数 J_{thr} 和 n_{thr}, 使得: $N \in (600, 1\,500)$, $\bar{k} \approx 6$

性应用到对语言信息缺失严重的句子进行属性值得到扩展, 使得扩展后的句子能够更加准确地表示评论者的意见, 比如只有一个词的评论句子 “祝贺!”, 扩展后为 “热烈祝贺马英九!”, 增加了带有积极倾向的程度副词 “热烈” 和评论对象 “马英九”, 句子的情绪倾向性就更加明显了. 本文第 2.2 节分析指出 SCP 扩展会存在负面效应, 由于评论文本的短文本特点, 并且应用简单分类算法处理这种属性稀疏的短文本文分类问题效果不佳, 因此 SCP 算法的负面效应在实验结果中并没有体现出来. 总体来说, SCP 属性处理算法对评论文本的情绪倾向性分类是有很大帮助的.

4 结论

实验结果表明, 本文提出的基于词语顺序共现随机网络模型的在线评论情绪倾向性分类算法 SCP-X, 对短文本的属性处理和情绪倾向性分类具有比较好的效果. SCP 算法可以扩展属性值, 在一定程度上解决了短文本数据的属性完整性较差、随意性较强的问题; 可以约简属性, 极大地减少向量空间模型中向量的维数 (减少至约 10 %), 从而降低向量的稀疏程度. 由于共现网络的建立是一个增量式的过程, 因此本算法可以改造成增量式学习算法, 能够随着训练数据的增加逐渐提高预测准确率. 由于 SCP 仅对样本进行属性的处理, 因此算法中的 X 可以替换为除 HMM 外的其他机器学习算法, 如: Bayes, SVM 等. 另外, SCP-X 算法中网络 G 的建立、SCP 算法和机器学习算法 X 都可以进一步优化, 以改善算法预测效果, 因此后续研究可以围绕这些方面展开. 本文的研究作为新闻评论短文本情绪分析研究的基础, 将为进一步研究开发突发公共事件应急管理决策分析系统提供支持.

References

1 Balog K, Mishne G, de Rijke M. Why are they excited? identifying and explaining spikes in blog mood levels. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations. Trento, Italy: Association for Computational Linguistics, 2006. 207−210

2 EACL. The workshop on new text Wikis and blogs and other dynamic text sources [Online], available: http://www.sics.se/jussi/newtext/, March 22, 2009

3 World Wide Web. The 3rd annual workshop on the weblogging ecosystem: aggregation, analysis and dynamics [Online], available: http://www2006.org/workshops/#W16, May 10, 2008

4 Wei Jiu-Chang, Zhao Ding-Tao. Research on the crisis information communication model and its impact factors. *Information Science*, 2006, **24**(12): 1782−1785
(魏玖长, 赵定涛. 危机信息的传播模式与影响因素研究. 情报科学, 2006, **24**(12): 1782−1785)

5 Turney P D, Littman M L. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems*, 2003, **21**(4): 315−346

6 Peter D. Turney thumbs or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2002. 417−424

7 Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifter. *Computational Intelligence*, 2006, **22**(2): 110−125

8 Vermeij M J M. The orientation of user options through adverbs, verbs and nouns. In: Proceedings of the 3rd Twente Student Conference on IT. Enschede, The Netherlands: University of Twente, 2005. 1−8

9 Mishne G, Glance N. Leave a reply: an analysis of weblog comments. In: Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem. Edinburgh, UK: IEEE, 2006. 1−7

10 Chen G W, Chiu M M. Online discussion processes: effects of earlier messages′ evaluations, knowledge content, social cues and personal information on later messages. In: Proceedings of the 6th International Conference on Advanced Learning Technologies. Kerkrade, The Netherlands: IEEE, 2006. 756−760

11 Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives. In: Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics. Madrid, Spain: Association for Computational Linguistics, 1997. 174−181

12 Yao Tian-Fang, Cheng Xi-Wen, Xu Fei-Yu, Uszkoreit H, Wang Rui. A survey of opinion mining for texts. *Journal of Chinese Information Processing*, 2008, **22**(3): 71−80
(姚天昉, 程希文, 徐飞玉, Uszkoreit H, 王睿. 文本意见挖掘综述. 中文信息学报, 2008, **22**(3): 71−80)

13 Gamon M, Aue A. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In: Proceedings of the Association for Computational Linguistics Workshop on Feature Engineering for Machine Learning in NLP. Michigan, USA: Association for Computational Linguistics, 2005. 57−64

14 Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan: Association for Computational Linguistics, 2003. 129−136

15 Kamps J, Marx M. Words with attitude. In: Proceedings of the 1st International Conference on Global WordNet. Mysore, India: Indian Institute of Technology, 2002. 332−341

16 Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2004. 271−278

17 Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne, Florida: IEEE, 2003. 427−434

18 Popescu A M, Etzioni O. Extracting product features and opinions from reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada: Association for Computational Linguistics, 2005. 339−346

19 Kudo T, Matsumoto Y. A boosting algorithm for classification of semi-structured text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004. 301−308

20 Ferrer-i-Cancho R, Sole R V. The small world of human language. *Proceedings of the Royal Society, Series B: Biological Sciences*, 2001, **268**(1482): 2261−2265

21 Shi Jing, Hu Ming, Dai Guo-Zhong. Topic analysis of Chinese text based on small world model. *Journal of Chinese Information Processing*, 2007, **21**(3): 69−75
(石晶, 胡明, 戴国忠. 基于小世界模型的中文文本主题分析. 中文信息学报, 2007, **21**(3): 69−75)

22 Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada: Association for Computational Linguistics, 2005. 724−731

23 Watts D J, Strogatz S H. Collective dynamics of "small world" networks. *Nature*, 1998, **393**(6684): 440−442

24 Barabasi A L, Albert R. "Emergence of scaling in random networks", *Science*, 1999, **286**(5439): 509−512

25 HowNet. HowNet knowledge database [Online], available: http://www.keenage.com/html/index.html, April 15, 2009

26 HIT IRLAB. HIT center for information retri [Online], available: http://ir.hit.edu.cn/, May 4, 2009

27 Joachims T. Making large-scale support vector machine learning practical. *Advances in Kernel Methods: Support Vector Learning*. Cambridge: MIT Press, 1999. 169−184

杨 锋 西安交通大学自动化系硕士研究生. 主要研究方向为复杂系统研究和突发性公共事件管理.
E-mail: yangfeng1983@gmail.com
(**YANG Feng** Master student in the Department of Automation, Xi′an Jiaotong University. His research interest covers complex systems and emergency response.)

彭勤科 西安交通大学自动化系教授. 主要研究方向为并行计算、符号序列分析、网络安全和生物信息处理. 本文通信作者. E-mail: qkpeng@xjtu.edu.cn
(**PENG Qin-Ke** Professor in the Department of Automation, Xi′an Jiaotong University. His research interest covers parallel computing, symbolic sequence analysis, network security, and biological information processing. Corresponding author of this paper.)

徐 涛 西安交通大学自动化系博士研究生. 主要研究方向为复杂系统研究和突发性公共事件管理.
E-mail: txu.xjtu@stu.xjtu.edu.cn
(**XU Tao** Ph. D. candidate in the Department of Automation, Xi′an Jiaotong University. His research interest covers complex systems and emergency response.)