Web

曹建平[1]，曾柯[2]，王晖[1]，程佳军[1]，乔凤才[1]

(1. 国防科技大学信息系统与管理学院 ， 2. 西安交通大学电子信息工程学院)

:随着社会化媒体的兴起，近年来情感分析发展迅速.但其在交通领域的应用还远远跟不上现代智能交通发展的旺盛需求.本文提出了交通情感分析（Traffic Sentiment Analysis）的方法和模型，分析对比了该领域内基于规则的和学习的两种方法在Web 数据的优势和劣势.在具体方法应用上，本文利用基于规则的方法来解决交通情感分析问题，从架构设计、相关库构建、处理流程、数据收集等方面具体阐述了基于 Web 的交通情感分析的方法.最后通过“新交规”、“油价”案例研究中验证了方法和应用的可行性.

基于 Web；情感分析；规则库；情感词库

0

发展更安全、科学的现代智能交通系统成为管理者和学术界的共识[1].随着 Web 2.0 的发展，社交网络成为人们表达个人看法、了解他人意见的重要平台[2].丰富的网络信息成为研究者进行数据挖掘的重要资源，人们获取信息和学习知识的方式已经发生了重大变化[3-6].本文提出基于网络数据的交通情感分析（Traffic Sentiment Analysis,TSA）方法，该方法作为跟踪民意、支持决策工具可以作为现代智能交通系统的有机组成部分.

情感分析就是挖掘数据有效信息，计算正面或负面、积极或消极，“同意”或“不同意”的情感值，发现情绪演变规律.交通情感分析有以下几点应用：(1)调查，获得公众有关交通意见.由于网络数据规模庞大，采集、处理和分析的难度很大，如果仅靠人工判读显然是不现实的，而这正是 TSA 的必要之所在.(2)评估，通过对采集的文本数据分析直接感知舆论、掌握公众观点以评估政府的交通政策及服务.(3)预测，TSA 可以帮助用户分析信息，预测交通相关事件发展和趋势.

1 TSA

现有的情感分析方法可分为基于规则的和基于学习的两种方法[7]，两种方法对于 TSA 都适用，但各有优缺点.基于学习的方法的优点是不需要专家的知识，不需要建立相关库，只是在训练分类器，不需要考虑上下文.由于文本长短不一，如果直接训练，文本的特征向量稀疏性不同，比较结果没有意义.如果按文档级别和句子级别进行训练，需要足够大的手工标注的正面和负面的例子，费时费力.由于网络数据的特点是不同用户、不同发布时间的语言风格不同，因此训练集难以覆盖整个数据集足够的特征.基于规则的方法缺点是，情感倾向的结果受到文本背景的影响，且需要建立专家知识库.但在处理中文网络数据时不受文本长短的影响，受文本特征影响不大且易于扩展.本文采用基于规则的方法，通过对中文文本的情感分析来说明 TSA 的关键问题.

基于规则的方法 TSA 针对 Web 数据的情感分析方法的主要有以下三个问题：1）框架设计，2）情感库构建，3）情感计算方法.下面将针对这三个问题进行探讨.

1.1

详见本文第 2 节.

1.2

情感词库 定义 $Seedp_0=\{$快，通畅，方便$\}$，$Seedn_0=\{$慢，拥堵，麻烦$\}$ 为正面情感词种子和负面情感词种子.输入哈工大的哈工大同义词词林扩展版 [8].通过寻找两个种子集的同义词和反义词对种子集分别进行扩展 $Seedp_1$ 和 $Seedn_1$.将这两个集合座位新一次迭代的输入，经过 k 次迭代，最后形成情感词库 $Seedp_k$ 和 $Seedn_k$.但所得的情感词库并没不完美，因为本文再添加由 CNKI（cnki.com）公布的情感词集来完善正面词库和负

面词库.因为有一些的词在交通领域具有特殊的意义，如“超载”和“U 型转弯”.因此手动增加了在交通领域的情感词库.最后构建一个包含 4893 个正面情感词、5416 个负面情感词的情感词词库.

假设情感词的情感倾向是由语素决定的，通过语素计算情感倾向.假设如果一个词的词素更频繁地出现在正面词词典中，它往往是一个正面情感的词，反之亦然.为了测量语素的正面和负面的倾向程度，分别赋予正面和负面的比重如下：

$$WeightP_{c_i}=\frac{fp_{c_i}\Big/\sum_{i=1}^{n}fp_{c_i}}{fp_{c_i}\Big/\sum_{i=1}^{n}fp_{c_i}+fn_{c_i}\Big/\sum_{i=1}^{n}fn_{c_i}} \tag{1}$$

$$WeightN_{c_i}=\frac{fn_{c_i}\Big/\sum_{i=1}^{n}fn_{c_i}}{fn_{c_i}\Big/\sum_{i=1}^{n}fn_{c_i}+fp_{c_i}\Big/\sum_{i=1}^{n}fp_{c_i}} \tag{2}$$

$$S_{c_i}=WeightP_{c_i}-WeightN_{c_i} \tag{3}$$

在式（3）中，情感词的极性取决于 S_{c_i} 的符号，强度取决于 S_{c_i} 绝对值.情感极性的计算步骤如下. ①扫描的正面和负面词库;②如果单词出现在积极的字词库，$S_w=1$③如果单词出现在否定词词典，$S_w=-1$④否则，计算的情绪极性由语素式（4），

$$S_w=\frac{1}{p}\sum_{j=1}^{p}S_{c_i} \tag{4}$$

其中 S_w 代表词 w 的情感倾向，是由 $c_1,c_2,\cdots c_p$ 组成，如果 $S_w>0$，情感词极性为正，否则为负.如果该值接近于零，则表明情感接近中性.

修饰词库利用专家知识构建否定副词、程度副词词典，给不同副词赋予不同的情感值修改等级[9]（表 1）.

表 1　程度副词分类

程度副词	修改等级	简介
极其|extreame/最|most	grade=2.5	增强语气
非常、很|very	grade=2.1	
较|more	grade=1.8	一定程度上增强语气
稍|-ish	grade=0.8	较小程度上增强语气
欠|insufficiently	grade=0.5	削弱预期
超|over	grade=-1	改变极性

语义规则库 语义规则就是情感词（S）和修饰词（否定副词（N）和程度副词（D））的布局使用 SND 模式来表示. 这 3 个因素中，最重要的是 S.因此首先从句子中找到 S.再从 S 的周围找到相应的 N 和 D，从而建立 SND 模型.

每个情感词都有其独特的修饰词.情感词和修饰词之间的关系主要取决于其在句子中的位置和类别.通过观察从网络上选择的 10000 句子的情感词及其修饰词的位置和类别，发现情感词的修饰语应该在同一个句子中，他们之间的距离一般小于 5 个汉字.本文总结了情感词和修饰词主要的位置顺序，如表 2 所示.需要特别指出的是，在 N + S 规则中，中文中存在多重多否认，偶数个否定词等于没有否定词而奇数个否定词相当于只有一个否定词.在 N + D + S 规则中，N 是程度副词 D 的否定词，N + D 的情感词（S）修饰.因此其特点是类似于 D + S.但是在规则中 D + N + S，否定词（N）修饰情感词（S），程度副词（D）为修饰词 N + S，

S 通常是一个动词或名词.

表 2 规则库简介

规则	例子	特征	计算公式
S+D	好(good)|极了(great) 方便(convenient)|得多(more)	S 通常为形容词	$p = p_s * p_d$
D+S	永远(forever)|支持(support) 绝对(absolute)|的|安全(safe)	副词 +形容词 副词 +动词 形容词 + 名词 形容词 +"的" +名词	
N+S	不(not)|支持(beautiful) 不(not)|安全(safe)		$p = -p_s$
N+D+S	不是(not)|很(very)|安全(safe)	类似 D+S	$p = -(1/3) * p_d * p_s$
D+N+S	很(very)|不|安全(safe)	S 通常为动词或名词	$p = -p_d * p_s$

设 p 是 SND 模式的情感极性值， p_s 表示情感词 S 的值， p_d 表示程度副词 D 的值，情感值的计算公式就如表 2 所示.例如，单词“安全”的得分为 2，程度副词“很（非常）”的值为 2.1.因此，根据规则 D + S 短语“很安全”的情感值为 4.2. 同样地根据规则 N + S，N + D + S，D + N + S，“不安全”、“不很安全”、和“很不安全”的情感值分别为-2，-1.4 和-4.2.

名词库 交通领域的特殊名词可能在文本分割时被忽略从而影响情感分析的质量.因此有必要构建的交通领域专用名词术语库.本文通过相关网站收集数据得到了包含汽车品牌以及交通类专业术语在内的 1732 个名词.

2 TSA

3.1 TSA

TSA 架构的核心处理过程，其主要组成部分如下(图 1 左)：①Web 数据采集，②预处理，③提取意见持有者和对象，④提取情感特征，⑤情感计算与分类，⑥评估或应用，⑦评估的反馈，从反馈中可以改进情感词库、规则库、以及 TSA 对象库.由于相关工作已经在之前介绍过了，在此本文将主要聚焦于应用的处理过程，简要阐述每个组件的基本内容（2-6），然后再讨论的整个处理过程和数据集合.

数据采集 针对这个问题，本文收集了如新浪微博，腾讯微博，天涯汽车论坛等网站的数据.本文所收集的数据基本来自网络用户，以此确保结论可以代表公众的意见，至少代表部分可检测到的公众意见[10].

预处理 中文文本处理的基本内容包括，①文本分割，②词性标注，③必要的替换.在论坛中，常常有各种各样的表达式表示相同的含义.例如，一些用户通常用“ d ”表示“顶”（支持）来表达他/她的赞同某个意见.采用了中科院 2011 年推出的 ICTCLAS 3 版进行分词[11].

分词优化 为避免不必要的干扰，提高分词的精度，预处理是根据材料和算法需求进行的.但是在某些情况下，这些处理会降低算法的精度 [12].

主客体提取 对于 TSA 而言，应当根据不同的数据集和数据来源设计相应的模型对文本背景进行挖掘，背景挖掘应当追求尽可能细腻完善的结果，以为以下步骤提供必要的背景知识.通过文本分析提取主体和客体对象.同时，也可通过文本分析提取相关的主客体.

特征提取 本文采用由 CL Zhang 等人在[7]中提出的 3 步战略进行特征抽取.

性能评估 在方法应用之前应当构建科学的标准数据集[13，14]来对相关方法进行测试.一方面，我们可以测试算法的效率和精度，另一方面，我们可以发现的相关库的不足之处，通过适当的词更新 TSA 数据库.

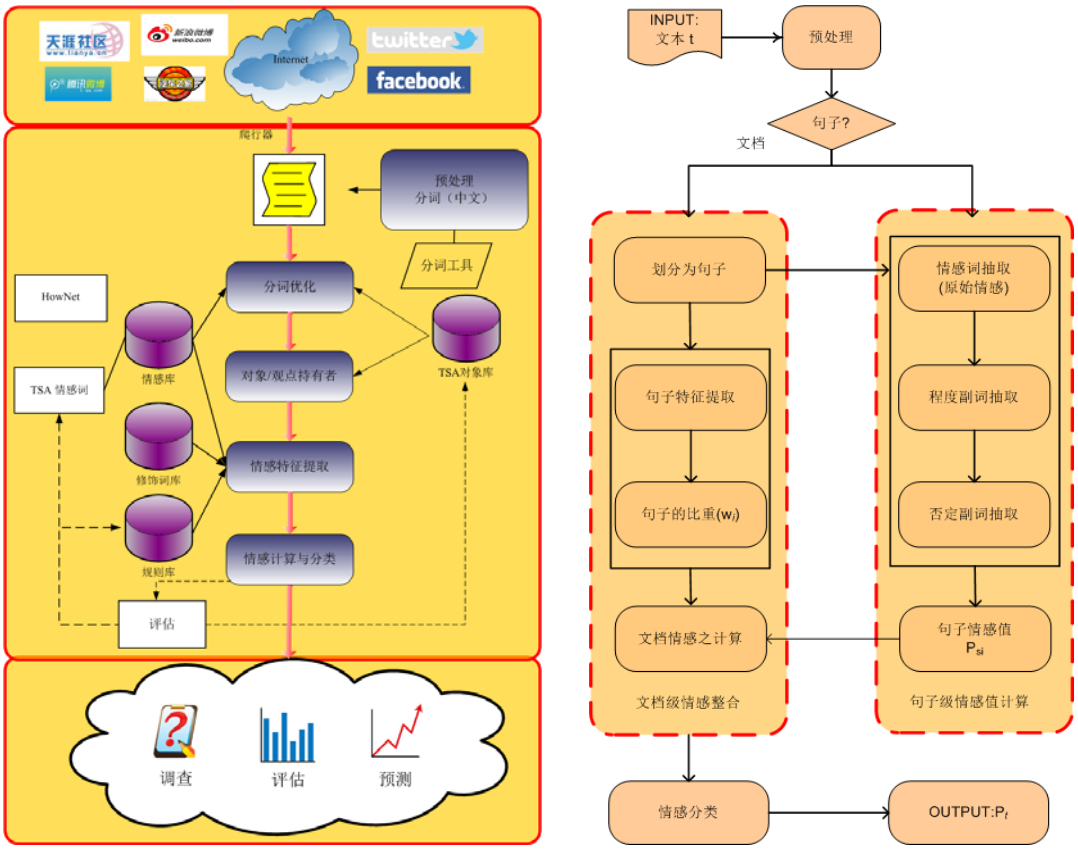

图 1　基于规则的 TSA 框架（左）、TSA 过程（右）

3.2 TSA

本文将单词或短语作为另一种形式的句子考虑.处理的文本包括两个主要的部分，句子和文档的情感极性值的计算.图 1 右半部分显示了本文方法的全过程，包括两个主要步骤：句子级的情感分析和文档级的情感叠加.本文首先将文档分解成句子，然后确定每个句子情感极性.设句子的极性为 p_s.计算 p_s 的核心步骤即提取句子的 SND 模式然后根据文本的 SND 模式计算情感极性值.关于 SND 模式的抽取本文在第 2 节第二部分的已经介绍.然后根据规则库中定义的规则计算句子 s_i 的极性值.

4

本节以研究“新交规”和“油价”个案例，并将本文算法同 Ku 算法[15]进行对比实验.

4.1

本文数据收集最大的中文在线社区天涯论坛（tianya.cn），该论坛有关于交通主题的讨论.其基本信息如表 3：

表 3 数据集基本信息

	话题	来源	记录	用户	事件
案例 1	新交规	*tianya.cn*	10286	5329	12/10/9 - 13/1/26
案例 2	油价	*tianya.cn*	12794	8834	04/8/22 - 13/2/18

4.2

标准集　选择了三个人的对文本进行情感标记，将所有的标注者都相同的标签的数据组成的一个实验

数据集.实验数据集中，主题 1 包含 547 条正面情感和 5937 条负面情感的数据，主题 2 包含 2516 正面的消息和 7418 负面消息的数据.

评价方法　本文用混淆矩阵评估算法优劣，选择了常用的精度、召回率和强度三项指标评估这两种算法的性能.由于精度、召回率已比较常见，在此简要介绍一下强度算法公式如（5）：

$$D-value=\frac{\sum_{i=1}^{N}(a_i-b_i)}{N} \tag{5}$$

式 5 中 $D-value$ 表示的算法结果和专家的标准之间的差异.差异越小越接近专家标准，N 文本的总数，a_i 表示计算的第 i 个文字的情感强度，由专家给出；b_i 表示的第 i 个文本表示的感情强度.式（8）可以进一步理解为式（9）：

$$D-value=\left|mean_f-mean_a\right| \tag{6}$$

这里 $mean_f$ 指算法的平均情感值，$mean_a$ 表示由专家给出的平均强度.

结果表 4 给出了两种算法的效果对比.

表 4　两种算法对比

实际	预测				总量
	本文算法		Ku 算法		
	积极	消极	积极	消极	
积极	463	84	351	196	547
消极	1054	4883	2018	3919	5937
总量	1517	4967	2369	4115	6484
积极	467	49	346	170	516
消极	1339	6079	1927	5491	7418
总量	1806	6128	2273	5661	7934

从表 4 中可以看到，本文的算法整体精度分别为 82.45％和 82.51％，较 Ku 算法为 65.85％和 73.57％，分别提高了 16.6％和 8.94％.这表明，不管是正面的还是负面的案例，其精度都有所增加.这是因为本文的规则是更适合与交通主题相关的数据集.我们积极情感的召回率是 30.52％和 25.86％，准确度为 84.64％和 90.50％;负面情绪召回是 98.31％和 99.20％，82.27％和 81.95％的精度.我们已经在一定程度上提高了 Ku 算法，古的积极情感召回率是 14.82％和 15.22％，精密率 64.17％和 67.05％;负面情感召回率 95.24％和 97.00％，精确率分别是 66.01％和 74.02％.相比之下，负文本召回率和准确率增加更明显，这是因为负面数据中大多数有复杂的句子的结构，这一点在我们的算法中被审慎考虑过了.

表 5　两种算法的强度比较

案例	专家 $mean_a$	本文算法		Ku 算法	
		$mean_f$	$D-value$	$mean_f$	$D-value$
1	0.13	0.128	0.002	0.244	0.114
2	0.18	0.17	0.01	0.381	0.201

表 5 显示了专家评审的情绪强度和两种算法计算得分相同的文字.比较的结果，我们发现：①我们的算法在情感的强度方面比 Ku 的算法具有更高的精度.在两种情况下，平均情感强度 Ku 算法最高.这是因为情感计算时，Ku 算法没有考虑对一般意义上的情感进行修改.②本文的算法强度接近专家评审的强度.这表明，我们所提出的积极的情感计算模型的算法更符合人类认识模式.

5

专门针对交通领域进行情感分析，这将是一个前景广阔的很好的应用.本文针对 Web 上的非结构化文本提出了交通情感分析（TSA），据我们所知，这是一个新的视角，尤其在智能交通系统的研究方面是第一次提出.我们的主要工作如下：⑴构建 TSA 应用处理框架的; ⑵建立 TSA 的相关库; ⑶提出了处理 TSA 的基于规则的算法; ⑷在情感值计算中考虑修饰关系，句型和情感词的位置.下一步我们将进一步完善算法的性能，充实相关情感词库并尝试用不同的方法去实现 TSA 功能.

[1]N. Zhang, F.-Y. Wang, F. Zhu, D. Zhao, and S. Tang, "DynaCAS: Computational experiments and decision support for ITS," Intelligent Systems, IEEE, vol. 23, pp. 19-23, 2008.

[2]X. Li, D. Zeng, W. Mao, and F.-y. Wang, "Online communities: a social computing perspective," in Intelligence and Security Informatics, ed: Springer, 2008, pp. 355-365.

[3]F. Y. Wang, "Social computing: Concepts, contents, and methods," International Journal of Intelligent Control and Systems, vol. 9, pp. 91-96, 2004.

[4]F.-Y. Wang, R. Lu, and D. Zeng, "Artificial Intelligence in China," Intelligent Systems, IEEE, vol. 23, pp. 24-25, 2008.

[5]S.-M. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," in Proceedings of the Workshop on Sentiment and Subjectivity in Text, 2006, pp. 1-8.

[6]B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the Web," in Proceedings of the 14th international conference on World Wide Web, 2005, pp. 342-351.

[7]C. L. Zhang, D. Zeng, J. X. Li, F. Y. Wang, and W. L. Zuo, "Sentiment Analysis of Chinese Documents: From Sentence to Document Level," Journal of the American Society for Information Science and Technology, vol. 60, pp. 2474-2487, Dec 2009.

[8]W. Che, Z. Li, and T. Liu, "Ltp: A chinese language technology platform," in Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, 2010, pp. 13-16.

[9]Y. Guo and Y. Zhou, "Chinese text orientation analysis based on phrase," in Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on, 2009, pp. 1-6.

[10]Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 355-362.

[11]ICTCLAS. (2011). http://ictclas.org/index.html.

[12]K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519-528.

[13]C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 625-631.

[14]N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima, "Collecting evaluative expressions for opinion extraction," in Natural Language Processing - Ijcnlp 2004. vol. 3248, K. Y. Su, J. Tsujii, J. H. Lee, and O. Y. Kwong, Eds., ed Berlin: Springer-Verlag Berlin, 2005, pp. 596-605.

[15]L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion Extraction, Summarization and Tracking in News and Blog Corpora," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 100-107.