Opinion Mining for Web Public Sentiment Based on Dynamic Knowledge Base

Wei WEI [1], Gao HUI [2], She LI [3]

[1] *School of Computer Science and Engineering, UESTC, ChengDu, China*

[2] *School of Computer Science and Engineering, UESTC, ChengDu, China*

[3] *School of Computer Science and Engineering, UESTC, ChengDu, China*

Email: lywei1982@163.com, huigao@uestc.edu.cn, sheli@uestc.edu.cn

Abstract: This paper applies opinion mining to the system of web public sentiment, putting forward an opinion collecting system of public sentiment of the Internet based on dynamic knowledge base. The system builds an opinion base through extracting rules automatically and adding to opinion base. For a hot event, we can extract people's opinions through matching with the opinion base. The experiment has shown that the proposed approach in this paper is feasible in early warning system for web public sentiment.

Keywords: web public sentiment; opinion mining; dynamic knowledge base; opinion base

基于动态知识库的 Web 舆情观点挖掘

卫　伟 [1]，高　辉 [2]，余　莉 [3]

[1] 电子科技大学计算机科学与工程学院，成都，中国，610054

[2] 电子科技大学计算机科学与工程学院，成都，中国，610054

[3] 电子科技大学计算机科学与工程学院，成都，中国，610054

Enail: lywei1982@163.com, huigao@uestc.edu.cn, sheli@uestc.edu.cn

摘　要：本文将观点挖掘技术应用于 Web 舆情系统，提出了一种基于动态知识库的网络舆情观点收集系统。该系统通过自动提取规则并将其加入观点库的方式，构建了一种舆情观点规则库。对于一个热点事件，通过匹配观点规则库就可以自动地提取出网民的观点。实验表明，该方法在实际的舆情预警系统中是切实可行的。

关键词：Web 舆情；观点挖掘；动态知识库；观点库

1 引言

网络舆情是通过互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点，主要通过 BBS 论坛、博客、新闻跟贴、转贴等实现并加以强化。由于互联网具有虚拟性、隐蔽性、发散性、渗透性和随意性等特点，越来越多的网民乐意通过这种渠道来表达观点、传播思想。网络舆情是一股强大的舆论力量，会反作用于热点事件并对社会发展和事态进程产生一定的影响。如果引导不善，负面的网络舆情将对社会公共安全形成较大威胁。对相关政府部门来说，如何加强对网络舆论的及时监测、有效引导，以及对网络舆论危机的积极化解，对维护社会稳定、促进国家发展具有重要的现实意义，也是创建和谐社会的应有内涵。

因此，舆情信息的观点收集具有相当重要的意义，在一个热点事件的发展过程中网民观点起着至关重要的作用，甚至可以被认为是网络舆情的核心。如何及时准确地收集网民对热点事件的观点成为一个亟待解决的问题。

目前已经有很多单位针对舆情监控提出了一些不同的解决方案。但这些系统都将重点放在舆情的准确发现和分类归类方面，而对于某个舆情事件中人们所持的观点和看法并没有做深入的研究。

2 研究进展

近年来，文本观点信息抽取和倾向性分析已经成为自然语言处理领域的一个热点问题。就中文领域的研究现状而言，主要采用两种方法：知识工程和自动

资助信息：国家高技术研究发展计划（863 计划）(2007AA01Z440)

978-1-935068-07-5 © 2009 SciRes.

训练[8]。基于知识工程的方法，主要是构建一个领域词典，包含该领域内所有可能的评价对象以及其特征[1]，经过匹配，识别出句子中的观点评价对象。北京大学言语听觉中心的张猛等人通过分别构建照相机、汽车、笔记本和手机四个领域的评价对象列表来识别各领域的观点评价对象。基于自动训练的方法，主要是采用有监督的机器学习算法来识别句子中观点评价对象，这种方法需要人工标注好的训练样本。不同领域所需训练样本也不同。北京大学信息科学技术学院计算语言所的蒙新泛等人根据评价对象抽取和命名实体识别都是针对某类特定的名词进行识别的相似性，采用序列标注模型 CRF(Conditional R andom Fi eld)模型来识别观点评价对象。同时他们初步研究了评价对象抽取的领域独立性试验。还有一种基于模板的评价对象抽取方法，该方法通过人工配置一些出现频率较高的名词性短语的模板来识别观点对象。华中师范大学的何婷婷、北京拓尔思信息技术股份有限公司的乔春庚等人都进行这方面的研究。

从以上分析可以看出，目前针对 Web 舆情监控和观点挖掘的研究还是两个独立的领域。Web 舆情监控侧重于舆情的准确发现和归类，而观点挖掘侧重于产品评论领域。如何将观点挖掘技术应用于舆情监控系统，具有非常重要的研究意义和现实意义。

3 观点主题抽取

产品评论领域的观点对象抽取是一种具体的、细颗粒度的过程。对于一个具有情感倾向的词，需要找出其所修饰的评价对象。例如，“蒙迪欧就显得比较沉闷，满眼深色的内饰稳重有余却又显得单调”这句话，其中包含两个情感词“沉闷”和“单调”，这两个情感词修饰的评价对象分别为“蒙迪欧”和“内饰”。

我们认为这种细颗粒度的观点评价对象抽取并不适用于 Web 舆情系统。

本文针对 Web 舆情的特点，提出一种基于动态知识库的 Web 舆情信息观点收集系统，对舆情事件中网民所发表观点进行提取。对于一句话中不同的观点评价对象，若其属于一个同义词集，则将它们作为一个观点对象处理。

系统流程框架如图 1 所示。首先从数据库中获取所有热点事件的文档集 I，针对 I 中一个热点事件文档子集，使用 TF-IDF 算法提取其中的关键词，然后找到该文档子集中包含关键词的句子集合，即关键句集 O；对关键句集 O 进行人工过滤，过滤掉与事件主题毫无关系的句子，再根据观点规则库对关键句集合 O 中剩下的句子进行过滤，然后对剩余句子按照主题词进行聚类，对于前几个大类，通过句法分析自动提取句子中的主谓结构作为候选主题，将候选主题中满足条件的主题加入观点规则库。最后对该热点事件进行情感分析。对文档集 I 中的每个热点事件都采用相同的方法，最终收集到网络中所有最新舆情信息的网民观点。

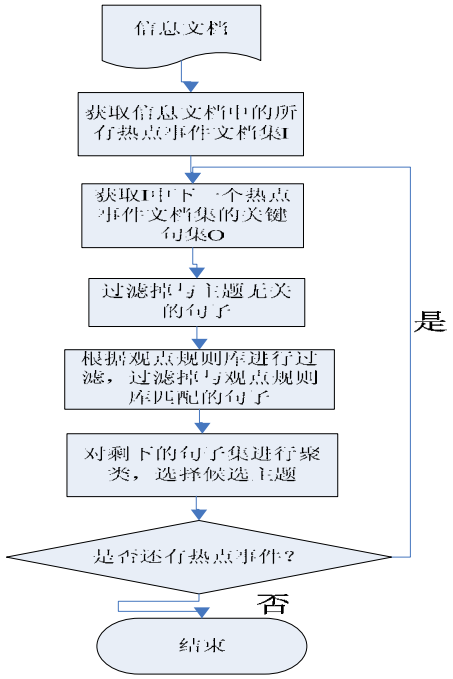

Figure 1. Flow chart:opinion collecting system
图 1. 舆情信息观点收集系统流程图

3.1 主题词选择算法

我们利用 Ku 和 Chen[2]提出的主题词选择算法，即找出那些在不同文档中出现频率较高的词或在一篇文档的不同段落中出现频率较高的词。公式如下：

$$W_{d_i,t} = TF_{d_i,t} \times \log \frac{N}{N_{d_i}} \quad (1)$$

$$W_{p_j,t} = TF_{p_j,t} \times \log \frac{N}{N_{p_j}} \quad (2)$$

$$Disp_{d_t} = \sqrt{\frac{\sum_{i=1}^{m}(W_{d_i,t} - mean)^2}{m}} \times TH \quad (3)$$

[1] 在本文中，我们对评价对象与其特征不做区分，统称为观点主题。

$$Dev_{d_i,t} = \frac{W_{d_i,t} - mean}{Disp_{d_t}} \quad (4)$$

$$Disp_{p_t} = \sqrt{\frac{\sum_{j=1}^{n}(W_{pj,t} - mean)^2}{n}} \quad (5)$$

$$Dev_{pj,t} = \frac{W_{pj,t} - mean}{Disp_{pt}} \quad (6)$$

其中，W 代表词的权重，d 代表文档层次，p 代表段落层次，TF 是词频， N 在公式（1）中为文档数，在公式（2）中为段落数，i, j 分别为文档索引和段落索引，t 为词索引。公式（1）和（2）分别表示词 t 在文档 d_i 和段落 p_j 中的 TF-IDF 值；公式（3）和（5）分别定义了词 t 出现在不同文档和不同段落中频率；公式（4）和（6）分别定义了词 t 出现在每个文档和每个段落中的频率。阈值 TH 用于控制符合条件的主题词的个数，TH 值越大，主题词数量越多。

满足下面两个公式之一的词就是我们所要找的主题词。

$$Disp_{d_t} \le Disp_{pt}, \exists d_i, \forall p_j \in d_i, Dev_{d_i,t} \le Dev_{pj,t} \quad (7)$$

$$Disp_{d_t} > Disp_{pt}, \exists d_i, \forall p_j \in d_i, Dev_{d_i,t} > Dev_{pj,t} \quad (8)$$

3.2 构建观点规则库

我们通过将观点规则作为知识库中的概念的方式来构建观点规则库。

我们将舆情事件分为“政治主题类”、“经济主题类”、“安全主题类”等十八个大类，每个大类下都有若干个相关的热点事件。

通过对舆情事件的分析，可以发现对于同一类别的舆情事件下的不同的热点事件，网民很可能会发表相似的观点。例如，对于“安全主题类”舆情事件下的两个热点事件：“拐卖儿童”事件和“孙伟铭案”事件，网民都发表了“加强法制建设”和“加强公民道德教育”两种类型的观点。若我们提取过“拐卖儿童”事件的观点主题，那么，对于“孙伟铭案”事件，通过匹配观点规则库，就可以直接提取出观点主题而不需要人工进行干预。

观点规则库是动态增长的。开始时观点规则库为空，根据图 1 所示的流程图，处理完一个热点事件，观点规则库就增加一次，随着规则库的日趋完善，人工干预的因素就会越来越小。

观点规则库的结构是一种树状结构，第一级节点代表不同类型的抽象观点，第二级节点表示该类型抽象观点下的若干条规则，每一条规则代表该类型抽象观点下的一种观点主题。对于一条规则，基于同义词典抽取相似的观点主题，是一种粗粒度的观点主题抽取。

例如，对于“三鹿奶粉”热点事件，网民会发表“加强法制建设”类型的抽象观点。在观点规则库中，我们添加了“加强法制建设”抽象观点，该抽象观点包含一条规则，该规则由两个同义词组组成，分别是“加强、强化、…”和“法律、法制、法规、…”。根据观点规则库，就可以提取“三鹿奶粉”事件中网民所发表的“加强法制建设”类的抽象观点。

4 情感分析

情感分析的任务是对网民所发表的抽象观点的倾向性进行分析，以便得到网民对该观点的态度是肯定、否定还是保持中立。

通过匹配观点规则库，我们得到每类抽象观点的观点句集，其中每个句子都属于某个用户。

算法流程如下：

步骤 1：对一个抽象观点的观点句进行分词、词性标注。

步骤 2：根据情感词库对每个观点句进行标记，然后根据标记计算出每个观点句的情感倾向值；

步骤 3：根据该句子所属用户标记，将属于同一用户的观点句的情感值相加，得到不同用户对该观点类型的情感值。

步骤 4：根据不同用户对该观点类型的情感值进行投票。投票规则如下：

$$vote = \begin{cases} \text{正面评价，} p_{ij} > 1.5 \\ \text{中立态度，} |p_{ij}| \le 1.5 \\ \text{负面（消极）评价，} p_{ij} < -1.5 \end{cases}$$

其中，p_{ij} 为用户 i 对抽象观点 j 的情感值。

在步骤 2 计算一个观点句的情感倾向值时，我们采用基于情感词库的方法。情感词库存储结构如表 1 所示。

首先分析句子中的词语成分，然后对其中的情感词进行标记，根据情感词的权值计算出该句子所体现出的针对某类抽象观点的情感倾向。计算规则如下：

978-1-935068-07-5 © 2009 SciRes.

首先将句子中相邻的相同词性的情感词采用以下规则进行合并：

1） 对相邻的形容词、动词、名词和副词词性的词语的权值采用加法规则；

2） 对相邻的转义词语的权值采用乘法规则；

然后对下面几种情况采用乘法规则进行合并：

a） 副词和其修饰的形容词采用乘法规则；

b） 副词和其修饰的动词采用乘法规则；

c） 形容词和其修饰的名词采用乘法规则；

Table 1. The structure of the base of the sentiment words
表 1. 情感词库存储结构

类型	标签	权值	词语举例
褒义形容词	Aj	1	好，美丽，…
贬义形容词	Dj	-1	差，丑陋，…
强烈副词	Ad	2	很，非常…
普通副词	Dd	0.5	略加，稍微…
褒义动词	Av	1	推动, 促进，…
贬义动词	Dv	-1	导致，破坏，…
褒义名词	An	1	快乐，优点，…
贬义名词	Dn	-1	人渣，缺点，…
转义词语	Mr	-1	没有，不，…

最后将剩下的权值应用加法规则得到一个关键句的情感倾向。同时我们考虑了标点对句子情感倾向的影响，尤其是问号对句子情感的影响。我们是基于这样的考虑，对于舆情事件，网民在很大程度上会以反问的形式发表自己的观点，这就需要反转句子的极性，即句子的情感倾向。

5 实验结果与分析

在实验中，我们对“猪流感”事件中网民所发表的观点进行了提取。该事件有 380 篇文档。通过提取关键句并进行聚类，我们发现网民主要发表了四种类型的抽象观点，分别为：“引起民众恐慌”、“预防体系健全”、“疫情传染性强”和“减缓经济发展”。

在“引起民众恐慌”类型下，我们构造了一条规则，该规则包含两个同义词组：“H1N1、流感、甲流、…”和“可怕、吓人、恐慌、…”。通过匹配，我们得到了 497 个含有该类型观点的句子，分属于 255 个文档。一个文档对应于一个用户，可知有 255 个用户发表了该类型的观点。其它观点类型如表 2：

通过对各抽象观点类型的观点句集进行情感分析，我们得网民对各类型观点的态度，如表 3 所示：

Table 2. The statistics of opinion
表 2. 观点类型统计

观点类型	句子数	用户数
引起民众恐慌	497	246
预防体系健全	559	236
疫情传染性强	153	130
减缓经济发展	94	67

Table 3. The statistics of sentiment analysis
表 3. 情感分析统计

观点类型	支持数	中立数	反对数
引起民众恐慌	80	27	139
预防体系健全	161	34	41
疫情传染性强	68	20	42
减缓经济发展	34	5	28

6 结论与展望

本文在较高的抽象层次上提出了一种 Web 舆情观点收集系统，并对不同的观点进行了情感倾向分析。该方法在实际的 Web 舆情预警系统中是切实可行的。下一步的研究将侧重提高自动提取观点规则的准确率以及如何减少描述同一个观点的观点句所聚的类数，以减少人工干预。

致 谢

本论文在国家高技术研究发展计划（863 计划）项目《WEB 舆情的社会网络关系挖掘》（合同号为 2007AA01Z440）中完成。

References (参考文献)

[1] Yu H, Hatzivassiloglou, Towards answering opinion questions: separating facts from opinions and identitying the polarity of opinion sentences, In M. Collins and M. Steedman(eds): Proc. Of the EMNLP-03: The 8th Conference on Empirical Methods in Natural Language Processing, 129~136, Sapporo, Japan, July 11-12, 2003

[2] Ku, L.-W., and Chen, H.-H. (2007). Mining Opinions from the Web: Beyond Relevance Retrieval. Journal of American Society for Information Science and Technology, Special Issue on Mining Web Resources for En-hancing Information Retrieval, 58(12), 1838-1850.

[3] Kim, Youngho, Jung, Yuchul and Myaeng, S.-H. (2007).

Identifying Opinion Holders in Opinion Text from Online Newspapers. Proceedings of the 2007 IEEE International Conference on Granular Computing.

[4] Dong Z, Dong Q, HowNet, http://www.keenage.com/

[5] Hatzivassiloglou V, McKeown K, Predicting the semantic orientation of adjectives, In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics(ACL-97), 174~181, Madrid, Spain, July 7-12, 1997.

[6] Zhu Yanfeng, Min Jin, Zhou Yaqian, etc. Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing, 2006,20(1),14-20.
朱嫣凤,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算[J].中文信息学报,2006,20(1),14-20.

[7] Lou Decheng, Yao Tianfang. Semantic polarity analysis and opinion mining on Chinese review sentences. Computer Applications, 2006,26(11):2622-2625.
娄德成,姚天昉.汉语句子语义极性分析和观点抽取方法的研究[J].计算机应用,2006,26(11):2622-2625.

[8] HeHui, LiSi, etc. PRIS Sentiment Analysis System Technical Report. The Collection of Papers published on the First Conference of Chinese Opinion Analysis Evaluation (COAE 2008).
何慧,李思,等.PRIS 中文情感倾向性分析技术报告[C]. 第一届中文倾向性分析评测会议论文集(COAE 2008).

[9] Qiao Chungeng, Sun Lihua, etc. Pattern Based Chinese Semantic Orientation Analysis. The Collection of Papers published on the First Conference of Chinese Opinion Analysis Evaluation (COAE 2008).
乔春庚,孙丽华,等.基于模式的中文倾向性分析研究[C].第一届中文倾向性分析评测会议论文集(COAE 2008).

[10] Meng Xinfan, Wang Houfeng. Detecting Target based on CRF and Evaluation Affects of Domain Specific Data on Target Detection. The Collection of Papers published on the First Conference of Chinese Opinion Analysis Evaluation (COAE 2008).
蒙新泛,王厚峰,基于 CRF 的对象抽取及对对象抽取的领域特定性研究[C].第一届中文倾向性分析评测会议论文集(COAE 2008).

[11] He Tingting, Wen Bin, etc. Research on Sentiment Terms' Polarities Identification and Opinion Extraction. The Collection of Papers published on the First Conference of Chinese Opinion Analysis Evaluation (COAE 2008).
何婷婷,闻彬,等.词语情感倾向性识别及观点抽取研究[C].第一届中文倾向性分析评测会议论文集(COAE 2008).

978-1-935068-07-5 © 2009 SciRes.