doi:10.3969/j.issn.1000-436x.2014.z2.030

面向热门话题的微博观点挖掘方法

张光磊[1]，徐雅斌[1,2]
(1. 北京信息科技大学 计算机学院，北京 100101；2. 北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 100101)

摘 要：提出了一种微博热门话题的观点挖掘方法。首先通过句法依存关系模板和支持向量机(SVM)共同识别热门话题中的观点句，然后进一步通过词法关系和句法依存关系抽取观点词对，从而简明、清晰展现热门话题中的观点。最后通过实验证明了该方法的有效性。
关键词：微博；热门话题；观点句；观点挖掘；句法依存分析
中图分类号：TP393 **文献标识码**：A **文章编号**：1000−436X(2014)Z2-0220-08

Opinion mining method on hot topic of micro−blog

ZHANG Guang-lei[1], XU Ya-bin[1,2]
(1.School of Computer, Beijing Information Science &Technology University, Beijing 100101, China;
2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing 100101, China)

Abstract: A method of opinion mining was argued in connection with hot topic of micro-blog. Firstly, opinion sentences are identified by syntactic dependencies templates and support vector machine (SVM). Secondly, the opinion words can be get in opinion sentences by syntactic dependencies and lexical relations, thus they can be displayed clearly. Finally, the experiment proves the effectiveness of this method.
Key words: micro-blog; hot topic; opinion sentences; opinion mining; syntactic dependencies

1 引言

微博热门话题通常包含当今政治、经济、体育等领域的各种热点事件。微博用户可以针对某一热门话题发表微博进行讨论，同时热门话题也会自动收录相关微博。对某一热门话题发布和转发的微博数量巨大，而且在这些微博中，往往包含很多不同的观点和看法。因此，如何有效挖掘其中的观点和看法成为文本分析领域又一新的研究方向。

早期的观点挖掘工作主要集中在商品的评论中，抽取用户对商品属性的评论。对于消费者来说，有助于进行商品间的比较，并据此来辅助做出购买决策。对商家来说，有助于改进商品在某一方面的缺陷或不足，提高产品的品质。

目前，研究人员普遍将研究重心放在了普通文本的情感倾向分析上，通过对篇章或句子层面的情感极性进行分析，最终将分析结果展示给特定的用户。例如：要了解观众对于某一部电影的喜好程度，可以通过判断影评的情感极性进行分析来获得，最终将正向评论和负向评论的数量所占的百分比展示给公众。这种方法虽然能提供一定的有用信息，但不能向用户展示更加具体的观点。例如，在“#菲军舰恶意撞击#”这个热门话题中，虽然大部分用户评论的情感极性为负向，但是缺少具体的观点和看法，难以了解民众表达的强烈意愿和诉求。

2 相关工作

目前，在观点句识别方面，方法大致可以分为 3 类：

收稿日期：2014-10-28
基金项目：国家自然科学基金资助项目(61370139)；网络文化与数字传播北京市重点实验室基金资助项目（ICDD201309）；北京市属高等学校创新团队建设与教师职业发展计划基金资助项目(IDHT20130519)
Foundation Items: The National Natural Science Foundation of China (61370139); The Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDD201309); Beijing Collages Innovation Team Building and Teacher Occupation Development Program (IDHT20130519)

基于词典的方法、基于统计的方法和基于图的方法。

基于词典的方法：Janyce Wiebe M[1]通过预先建立好的词典，统计文本中的情感词数量，由此判断是否为观点句。

基于统计的方法：叶强等[2]提出连续双词词类组合模式(2−POS)，采用 CHI 统计方法计算主客观程度，并通过设定阈值来判断是否为观点句。蒙新泛[3]通过提取观点句或者非观点句中的特征，并利用这些特征对分类器进行训练，通过分类模型对观点句进行识别。

基于图的方法：Bo Pang 等[4]使用最小分割法把文本分为观点句和非观点句。

在观点词对抽取方面，章剑锋等[5]把同一句子中共现的评价词与评价对象作为候选集，应用最大熵模型，并结合词、词性、语义和位置等特征进行关系抽取。

通过分析发现，以上方法都是针对传统的 Web 文本信息进行的。但是，微博这一新型媒体与传统的 Web 文本信息存在较大的差别，因为微博中的用语一般偏向口语化，具有单一性、碎片化等特点。因此，已有的针对传统 Web 文本进行观点挖掘的方法并不能完全适用于微博中的观点挖掘。

为此，本文提出了一种适用于微博的观点挖掘方法。该方法的基本思想如下。

对于某一热门话题来说，微博用户的观点和看法往往是通过微博中的观点句来体现的。如果能够找到每条微博中的观点句，并在找到的观点句中进一步挖掘出能够清晰表达观点的以范式表示的观点词对，再经过汇总与合并处理，即可获得针对该热门话题的微博观点。本文具体做法是：首先在人工标注好的观点句中提取可以识别观点句的句法依存关系模板，通过模板对符合常规表达方式的观点句进行识别。对于模版无法匹配的潜在观点句，再通过特征提取，然后采用机器学习的方法获取。对于已获得的观点句，还需要进一步挖掘评价对象和评价词，组成<评价对象，评价词>形式的观点词对，并按出现频度排列。对于观点相同或相近的观点词对，还需要进行合并。那么，最后出现频率较高的观点词对即为主流的观点。

3　基于模板的微博观点句识别

3.1　微博观点句的判别准则

张博[6]对观点句给出的判别准则如下。

1) 只要句子中表达了对事物的某种评价、看法，或流露出个人的情感倾向，无论是以第一人称发表，还是以第三人称发表，均判定为观点句。

2) 当有情感倾向词出现时，若句子整体上的意图是为了描述某个客观事实，则判定为非观点句；否则，判定为观点句。

3) 若句子表达的是对未来事件的一些预测或期许，则判定为观点句。

微博作为一种短文本，同样蕴含着观点句，上述观点句的判别准则同样适用于微博。因此，本文就采用张博[6]对观点句给出的判别标准。

3.2　句法依存关系模板的建立

句法依存关系用来揭示语言单位之间的句法结构，即句子中“词对”的二元关系。其中一个词为核心词，另一个为依存词，依存关系用来反映核心词和依存词之间的依赖关系。表 1 为部分典型的依存关系及其对应的标记。

表 1　　部分依存关系标记

关系类型	标记
主谓关系	SBV
动宾关系	VOB
状中结构	ADV
动补结构	CMP
核心词	HED

以观点句“条件差得很”为例，其句法依存关系分析结果如图 1 所示。

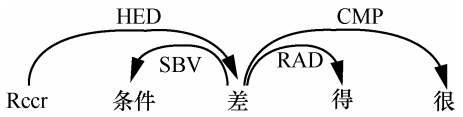


图 1　句法依存关系示例

从图中可以看出，核心词为“差”，依赖于“差”的词为“条件”、“得”、“很”，其依存关系分别为“SBV”、“RAD”、“CMP”。“RAD”为“得”字结构的右附加关系。那么，可以提取“SBV+差+CMP”作为一个可以识别的观点句模板。该观点句的依存关系以 XML 文档形式存储，如下所示。

<sent id=“0” cont=“条件差得很”>

<word id=“0” cont=“条件” pos=“n” ne=“O” parent=“1” relate=“SBV”>

<word id=“1” cont=“差” pos=“a” ne=“O” parent=“−1” relate=“HED”>

<word id=“2” cont=“得” pos=“u” ne=“O”

parent=“1”relate=“RAD”>

<word id=“3”cont=“很”pos=“d”ne=“O”parent=“1”relate=“CMP”>

</word>

</sent>

本文使用哈尔滨工业大学信息检索研究室语言技术平台（LTP, language technology platform）中的句法分析工具对此类文档进行解析并提取可以识别观点句的依存关系模板。提取句法依存关系模板的算法如下。

Input: 已标注好的语料，*OpSn*(*n*=1,2,3…)为观点句的集合，*NopSn*(*n*=1,2,3…)为非观点句的集合，W_i(*i*=1,2,3…)表示句子中的每个词语。

Output:可以识别观点句的句法依存关系模板集合 Opinion Template

Step1 对已标注好的观点句集合 *OpSn* 进行句法依存关系分析，将 XML 格式的分析结果 *syntactic*(*OpSn*)存入集合 *S* 中。

Step2 检索集合 *S* 中 relate = HED 的核心词，将检索出的核心词记为 HedNode。

Step3 查找依赖于核心词 HedNode 的所有词，并将其依存关系记为 DependenceNode。

Step4 在 DependenceNode 集合中取“DependenceNodei−1(核心词前一个词与核心词的依存关系)+ HedNode(核心词)+ DependenceNodei+1(核心词后一个词与核心词的依存关系)”作为候选模板存入 Candidate Template 集合中。

Step5 各个模板在观点句中出现的次数 GetCount(Candidate Template, OpSn)$\rightarrow$*Op*，在非观点句中出现的次数 GetCount(Candidate Template, NopSn)$\rightarrow$*Nop*。

Step6 如果 *Op*>10 并且 *Nop*<5，则将这个候选模板 Candidate Template 判定为观点句模板，并将其置入观点句模板集合 Opinion Template 中。

表 2 为使用该算法在 300 句人工标注好的数据中提取的出现次数排名前 5 的依存关系模板。

表 2 依存关系模板抽取部分结果

模板	观点句中出现次数	非观点句中出现次数
SBV 认为 VOB	48	1
SBV 表示 VOB	43	4
ADV 就是 VOB	42	6
ADV 认为 VOB	33	2
SBV 觉得 VOB	30	1

3.3 基于句法依存关系模板的观点句识别算法

具体识别算法如下。

Input:未经人工标注的微博集合 S_n(*n*=1, 2, 3…)

Output: 观点句的集合 Op_n(*n*=1,2,3…)

Step1 把未经人工标注的观点句集合 S_n 进行句法依存关系分析，将 XML 格式的分析结果 *syntactic*(S_n)存入集合 *R* 中。

Step2 在集合 *R* 中检索关系类型为 HED 的核心词，结果记为 GetRelation(R,‘HED’)，并记录其 *id*(值为 *i*)，将 *GetRelation*(*R*,‘*HED*’)$\rightarrow$*W*。

Step3 检索表 2 中的 Opinion Template，如果存在核心词 *W*，跳至 Step4，如果不存在则结束。

Step4 在集合 *R* 中检索 *id*=*i*−1 和 *id*=*i*+1 与 *W* 的依存关系，其结果分别记为 *GetRelation*(*R*,*i*−1)和 *GetRelation* (*R*,*i*+1)，组成“*GetRelation*(*R*,*i*−1)(核心词前一个词与核心词的依存关系)+ *W*(核心词) + *GetRelation* (*R*,*i*+1)(核心词后一个词与核心词的依赖关系)”的组合，并将 *GetRelation*(*R*,*i*−1)+*W*+*GetRelation*(*R*,*i*+1)$\rightarrow$*Template*。

Step5 在 Opinion Template 中检索是否存在观点句模板 Template，如果存在则判定该句为观点句，并置入 Op_n 中，否则结束。

仍以句子“条件差得很”为例。通过句法依存关系分析得到关系类型为“HED”的词为“差”，该词 *id* 为 1。假设在依存关系模板中存在“差”这个词，提取 *id*=0 和 *id*=2 的词与“差”的依存关系，组成“SBV+差+RAD”，RAD 为“得”字结构的右附加关系，该关系无法表达观点，所以取 *id*=3 的词与“差”的依存关系，组成“SBV+差+CMP”，如果该结构在句法依存关系模板中可以查找到，即判定为观点句。

4 基于机器学习的观点句识别

观点句的识别结果可以分为观点句和非观点句。由此看来，观点句识别问题可以看作一个二分类问题。由于支持向量机(SVM, support vector machine)在二分类问题中表现最为出众，因此本文使用 SVM 对微博进行观点句识别。

4.1 观点句特征提取

在微博中，存在语言表达不规范、俚语和网络用语较多等问题，并且微博比较简短，因此，单纯地使用文本特征（词频逆文档频率、信息增益、χ^2 检验等）会造成数据稀疏性问题。鉴于此，本文在

传统的 5 个非文本特征基础上，增加了句法依存关系个数、连词个数、表情符号 3 个特征，共计 8 个非文本特征，如表 3 所示。

表 3　　八维非文本特征

特征编号	特征	取值
1	是否含有情感词	0,1
2	是否含有主张词	0,1
3	是否含有程度副词	0,1
4	是否含有问号	0,1
5	是否含有感叹号	0,1
6	句法依存关系的个数	*num*
7	连词个数	*num*
8	是否含有表情符号	0,1

下面对新增加的 3 个特征介绍如下。

1) 基于句法依存关系个数的特征

通过对数据集进行分析发现，在观点句中，“主谓结构”（SBV）和 “动宾结构”（VOB）在观点句中出现次数较多。例如：在“苹果的分辨率很高”这条微博中，“分辨率高”即为 SBV 结构。在“我喜欢三星的屏幕”这条微博中，“喜欢屏幕”即为 VOB 结构。故此，本文统计这 2 种结构在一条微博中出现的次数 *num*，并将其作为特征值存放在特征向量中。

2) 基于连词个数的特征

通过分析发现，使用连词的句子也具有较强的观点倾向，尤其是转折性连词出现时更是如此。例如，“虽然学习很辛苦，但是我相信只要肯付出努力就一定会有所收获的!”这句话中的连词为“虽然”、“但是”，连词个数为 2。因此，可将该特征作为观点句特征之一。

3) 基于表情符号的特征

在微博这种新型媒体中，以文本形式夹杂在微博中的表情符号同样也能代表用户对一个事件或人物的喜与憎，有明显的主观倾向。例如：表情  所对应的文本为“/哈哈”。因此，本文将该特征也作为观点句的特征之一，且该特征很容易匹配。

4.2　基于 SVM 的观点句识别

观点句分类为线性可分问题，SVM 对于线性可分问题的分类原理[7]如下。

训练集 $T=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2)(\vec{x}_i,y_i)\}$，其中，$\vec{x}_i\in R^n$，$i$=1, 2,…, l, $\vec{x}_i$ 表示第 i 个样本的特征向量，$y_i\in\{-1,1\},i=1,2,\cdots,l$，表示第 i 个样本对应的类别。若存在 $\vec{w}\in R$、b 和正数 ε，当 y_i=1 时，有 $(\vec{w},\vec{x}_i)+b\geqslant\varepsilon$；当 y_i=−1 时，有 $(\vec{w},\vec{x}_i)+b\leqslant-\varepsilon$，称训练集线性可分。

本文中，$\vec{x}_i$ 代表第 i 条微博的特征向量，y_i 代表第 i 条微博对应的类别，y_i=1 时，该条微博即为观点句，y_i=−1 时，该条微博即为非观点句。

对于线性可分问题可以使用最大间隔方法求解。首先需要根据向量 $\vec{w}$ 找到两条临界直线，记为 l_1 和 l_2，并使这两条直线间距离最大。然后设 l_1 的直线方程为 $(\vec{w},\vec{x}_i)+b=1$，l_2 直线方程为 $(\vec{w},\vec{x}_i)+b=-1$，l_1 与 l_2 间的距离为 $\frac{2}{\|w\|}$，那么，$(\vec{w},\vec{x}_i)+b=0$ 即为本文所需要的分类直线。本文通过对微博进行特征提取，构造的特征向量为 $\vec{x}_i$，并人工标注 y_i，通过 $\vec{x}_i$ 构造训练集 T，通过对 SVM 模型进行训练之后即可用于识别观点句。

5　观点词对的抽取

前面获取的观点句所表达的观点虽然相同，但观点句的表达形式可能不同。因此，需要将每个观点句加工成一个范式，即<评价对象，评价词>形式的观点词对。这样就便于将反映相同观点的观点句进行合并，获取无重复度的观点。而且，观点词对可以更加清晰、简洁的表达观点。

5.1　基于句法依存关系的观点词对抽取

本文首先通过情感词典对已识别出的观点句查找评价词，并将找到的评价词记为 $Word_e$，然后通过句法依存关系得到 $Word_e$ 对应的评价对象 $Word_o$，即可组成观点词对<$Word_o$, $Word_e$>。本文通过三种句法依存关系提取评价词对应的评价对象，即 SBV、VOB、ATT(定中关系)。下面分别说明每个句法依存关系对应的观点词对提取方法。

1) SBV 关系观点词对提取方法

例如：“屏幕太小了”。通过情感词典过滤得到 $Word_e$ 为“小”，在这句话中“屏幕”通过 SBV 关系依赖于“小”，可以得到 $Word_o$ 为“屏幕”，所以提取到的观点词对为：<屏幕，小>。

2) SBV 和 VOB 组合关系观点词对提取方法

例如：“调研多是走过场”。通过情感词典过滤得到 $Word_e$ 为“走过场”，“调研”和“走过场”分别通过 SBV 关系和 VOB 关系依赖于“是”，提取的观点词对为：<调研，走过场>。

3) ATT 关系观点词对提取方法

例如：“省长水平之高”。通过情感词典过滤得

到 $Word_e$ 为“高”,“水平”通过 ATT 关系依赖于“高”,提取的观点词对为：<水平，高>。

5.2 基于词法关系的观点词对抽取

所谓词法关系是指句子中词性和词性的组合关系。根据词法关系，本文同样可以找到观点词对。本文认为形容词和名词的组合最容易表达观点，但是形容词和名词的组合方式会在观点句中以不同的形式出现。表 4 为本文使用 3 种词法关系提取观点词对的例子。表中第一列代表词法关系的组合。其中，n 代表名词，ws 代表命名实体，adj 代表形容词。名词作为评价对象，形容词作为评价对象对应的评价词。

表 4　　词法关系观点词对提取

词法关系	提取方式	示例
n/ws+程度副词+adj	<adj，n>	例如：图像/n　非常/d　细腻/a 提取：<图像，细腻>
n/ws+adj	<adj，n>	例如：方舟子/n　无聊/a 提取：<方舟子，无聊>
adj+的+n/ws	<adj，n>	例如：很好/adj 的学生/n 提取：<学生，很好>

分别使用句法依存关系和词法关系提取评价词对应的评价对象，并组合成<评价对象，评价词>形式的集合，将句法依存关系和词法关系提取的结果集合分别记为 R_1，R_2，取 $R_1 \cup R_2$。

5.3 观点词对的合并

由于所获得的观点词对集合 R 中有些观点词对内的评价对象是相同的或相似的（同义词或近义词），为此本文需要首先对 R 中每个观点词对内的评价对象计算与其他观点词对内的评价对象的相似度，对于相似度较高的评价对象来说，要考虑对这样的观点词对进行合并。但是，对于要合并的这些观点词对来说，其内的评价对象又可能对应相同的或不同的评价词。对于具有相同评价词的相似评价对象来说，可以直接按评价对象进行合并。但是对于具有不同评价词的相似评价对象来说，还需要对评价词进行相似度计算，只有对于评价词相似（情感色彩是相当的）的情况，才可以对这样的观点词对进行合并。对观点词对的合并，本文针对不同情况给出了 2 种不同的方法，即：基于 HowNet（知网）进行观点词对合并的方法和基于词语覆盖度进行观点词对合并的方法。

5.3.1 基于 HowNet 的观点词对合并

对于不同评价对象之间和不同评价词之间的相似度计算，本文基于知网 HowNet 进行计算。

在知网中每个词汇有多个概念，用来解释概念所用到的词汇叫做义原，义原是用来表示概念的最小单位。一个实词概念的语义表达式分为 4 部分：第一独立义原描述式的相似度，记为 sim_1，其他独立义原描述式的相似度，记为 sim_2，关系义原描述式的相似度，记为 sim_3，符号义原描述式的相似度记为 sim_4。本文使用刘群等人[8]的方法计算词语之间的相似度，计算方法如下

$$SIM = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \tag{1}$$

义原和义原之间的相似度可通过式(2)计算得到

$$sim = (p_1, p_2) = \frac{\alpha}{\alpha + d} \tag{2}$$

其中，d 是 2 个义原在义原树上的距离，α 和 β 为预设参数。

这里选取的参数：α=1.6；β_1=0.5；β_2=0.2；β_3=0.17；β_4=0.13。

该相似度是在 0 到 1 之间的实数，当该数值大于阈值 σ_{HowNet} 时，即判定为相同或相似。

评价对象的合并过程举例如下。

“警察”和“女警察”，通过 HowNet 计算相似度达到 0.9，可以判定“警察”和“女警察”为同一评价对象，然后在热门话题名称中查找是否含有以上 2 个评价对象。如果热门话题为“#最美女警察#”，“警察”和“女警察”则合并为女警察。如果待合并的评价对象在热门话题中不存在，为提高对评价对象的概括程度，合并为义原个数较多的评价对象。

对于评价词的合并，同样使用 HowNet 进行相似度计算。在评价对象可以合并的情况下，如果它们的评价词属于同一情感强度，通过 HowNet 进行相似度计算，大于阈值 σ_{HowNet} 时可以进行合并，合并时取义原个数较多的评价对象。例如“美丽”和“漂亮”，相似度为 1，可以进行合并，它们的义原个数分别为“1”和“2”，合并为义原个数较多的评价对象，最终合并为“漂亮”。

5.3.2 基于词语覆盖度的观点词对合并

在基于 HowNet 进行评价对象合并时发现，有些评价对象并不属于 HowNet 登录词，对此，本文使用词语覆盖度进行观点词合并。

假设两个评价对象 $Word_{o1}$ 和 $Word_{o2}$ 含有重复的词或词语，$count(Word_{o1} \cap Word_{o2})$表示重复字的个数，$count(Word_{o1} \cup Word_{o2})$表示共同含有字的个数。

词语覆盖度的计算如下

$$\text{Coverage}(Word_{o_1},Word_{o_2})=\frac{count(Word_{o_1}\cap Word_{o_2})}{count(Word_{o_1}\cup Word_{o_2})} \quad (3)$$

词语覆盖度是 0 到 1 之间的实数，当该数值大于阈值 σ_{Coverage} 时，即判定为同一评价对象。

对于评价对象的合并过程举例如下：“葱”和“大葱”的词语覆盖度为 1/2=0.5，如果 σ_{Coverage} 为 0.5，即可以判定这 2 个词语为同一评价对象。例如，在热门话题“#疯狂的大葱#”中，“疯狂的大葱”分词结果为“疯狂 的 大葱”，热门话题中含有“大葱”，那么“葱”和“大葱”合并为“大葱”。对于在热门话题中没有查找到的评价对象，为了使结果更加准确，合并为字数较少的评价对象。

对于评价词的合并，同样使用 HowNet 进行相似度计算，方法同 5.3.1 中评价词的合并。

5.3.3 观点词对合并算法

观点词对合并算法如下。

Input: 5.2 节中的观点词对集合 R 及 R 中元素的个数 n。

Output: 合并后的观点词对集合。

Step1 在集合 R 中提取全部评价对象 GetObject(R)和其对应的评价词 GetEvaluation(R)，并将 GetObject(R) → Object，GetEvaluation(R) → Evaluation。

for(i=0;i<n; i++){

　for(j=0; j<n−1;j++){

Step2 如果 Object 中 2 个元素 $Object_i$ 和 $Object_{j+1}$ 都属于 HowNet 登录词，那么通过 HowNet 计算它们的相似度，其结果为 σ_{HowNet}，如果不属于 HowNet 登录词，则通过词语覆盖度计算相似度，其结果为 σ_{Coverage}。

Step3 如果 σ_{HowNet} 大于设定阈值，或者 σ_{Coverage} 大于设定阈值，则继续执行；否则，不合并评价对象，跳出当前循环。

Step4 在热门话题中检索是否含有 $Object_i$ 或 $Object_{j+1}$，如果含有,则合并为在热门话题中出现的评价对象，否则继续。

Step5 如果 $Object_i$ 和 $Object_{j+1}$ 都属于 HowNet 登录词，那么通过 HowNet 获取它们的义原个数 GetCount($Object_i$) 和 GetCount($Object_{j+1}$)，并将 GetCount($Object_i$) → $Count_i$，GetCount($Object_{j+1}$) → $Count_{j+1}$，比较 $Count_i$ 和 $Count_{j+1}$ 的大小，如果 $Count_i$>$Count_{j+1}$，则评价对象合并为 $Object_i$；如果 $Count_{j+1}$<$Count_i$,则评价对象合并为 $Object_{j+1}$；如果 $Count_i$=$Count_{j+1}$，不合并评价对象。如果 $Object_i$ 和 $Object_{j+1}$ 都不属于 HowNet 登录词，则合并为字数较少的评价对象。

Step6 在 R 中获取 $Object_i$ 和 $Object_{j+1}$ 对应的评价词，分别记为 $Evaluation_i$ 和 $Evaluation_{j+1}$

Step7 在情感词典中检索 $Evaluation_i$ 和 $Evaluation_{j+1}$，如果存在，则继续执行；否则，不合并评价词，跳出内循环。

Step8 如果 $Evaluation_i$ 和 $Evaluation_{j+1}$ 属于同一情感强度，并且 $Evaluation_i$ 和 $Evaluation_{j+1}$ 都属于 HowNet 登录词，那么通过 HowNet 获取它们的义原个数 GetCount($Evaluation_i$)和 GetCount($Evaluation_{j+1}$)，并将 GetCount($Evaluation_i$) → Count_i，GetCount($Evaluation_{j+1}$) → $Count_{j+1}$。比较 $Count_i$ 和 $Count_{j+1}$ 的大小，如果 $Count_i$>$Count_{j+1}$，则评价词合并为 $Evaluation_i$，且该观点词对的出现频率 f 增加 1；如果 $Count_{j+1}$<$Count_i$，则评价词合并为 $Evaluation_{j+1}$，且该观点词对的出现频率 f 增加 1；如果 $Count_i$=$Count_{j+1}$，不合并评价词。如果 $Evaluation_i$ 和 $Evaluation_{j+1}$ 都不属于 HowNet 登录词，则合并为字数较少的评价词。否则，跳出当前循环。

　}

}

6 实验

6.1 数据集及数据预处理

本文实验数据采用 NLP&CC2012(自然语言处理与中文计算会议)提供的《中文微博情感分析测评数据》[9]。数据集包括 20 个话题，每个话题采集大约 1 000 条微博，共约 20 000 条微博，数据采用 XML 格式，已经预先切分好句子，共 31 675 句。首先使用哈尔滨工业大学社会计算与信息检索研究中心提供的 LTP（语言技术平台）工具进行分词，在用户自定义词典中加入情感词典，然后使用正则表达式去除了带有网址链接和 QQ 号码的广告信息，剩余微博 16 736 条，共 25 104 句。

情感词典包含 3 个部分：褒义词典、贬义词典和表情词典。本文将知网[10](HowNet)中的情感词典和台湾大学整理的中文情感词典(NTUSD)[11]相合并，但去除了其中重复的情感词，并添加了人工收集的网络常用词和微博中的表情符号。对这些词典中的褒义词、贬义词由人工进行分类，分别存入褒义词典和贬义词

典中，并将褒义词和贬义词的情感强度分为 3 级（按照褒贬的程度分为 3 级，且逐级增强）。

6.2 评价指标

本文评价指标使用微博情感分析测评大纲[9]中的评价标准作为评价指标。观点句识别(观点词对抽取)评价指标计算方法如下

准确率： $P=\dfrac{system\_correct}{system\_proposed}$

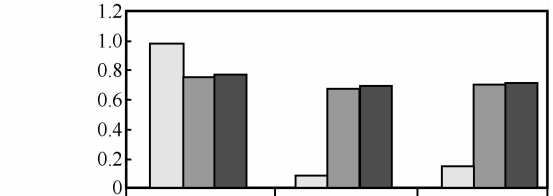
召回率： $R=\dfrac{system\_correct}{gold}$

F 值： $F=\dfrac{2PR}{P+R}$

其中，观点词对抽取评价标准采用微博情感分析测评大纲[9]中的严格标准，抽取结果与标准答案完全对应即为正确。*gold* 是人工标注结果的数目，*system_correct* 是提交结果中与人工标注匹配的数目，*system_proposed* 是提交结果的数目。

6.3 实验结果及其分析

在本次实验中，人工标注的训练集共 4 000 句，覆盖了数据集中的全部热门话题，测试集共 4 000 句，按照句法依存关系模版进行观点句识别、基于 SVM 进行观点句识别、以及将句法依存关系模版和 SVM 相结合进行观点句识别的实验结果如图 2 所示。



	准确率	召回率	*F*值
模板	0.987	0.082	0.151
SVM	0.761	0.075	0.715
模板+SVM	0.773	0.682	0.725

图 2 3 种方法识别观点句的结果

通过图 2 可以看出，按照句法依存关系模版进行观点句识别的准确率高达 98%，但是召回率仅有 8%。说明该方法精度较高，可以和其他方法结合使用。采用 SVM 进行观点句识别，识别的准确率为 76.1%，召回率为 67.5%。说明可以在召回率方面弥补使用模板识别观点句方法的不足。采用依存关系模版和 SVM 相结合进行观点句识别，识别的准确率为 77.3%，召回率为 68.2%。通过将这 2 种方法相结合，发现在各个指标上都稍有提高，证明了这 2 种方法结合使用的可行性。

例如，针对#疯狂的大葱#这一热门话题进行观点提取，所获得的观点句如表 5 中第 2 列所示。由表 5 可以看出，该方法不仅可以识别带有表情符号的观点句，而且其识别准确率较高。

由于采用依存关系模版和 SVM 相结合进行观点句识别方法的识别效果明显好于其他 2 种单独的识别方法，为此本文仅保留该方法进行观点句识别的结果。接下来，以此为基础进行观点词对提取，获得的观点词对如表 5 中第 3 列所示。进一步进行观点词对合并后的得到的最终观点词对如表 5 中第 4 列所示。

表 5 观点识别结果

热门话题	观点句	合并前的观点词对	合并后的观点词对	出现频次
#疯狂的大葱#	1.大葱贵！	<大葱，贵>	<大葱，贵>	127
	2.葱贵死了！	<葱，贵>		
	3.季节的因素很关键，关键是工资太低！	<工资，低>	<工资,低>	98
	4.这年头物价一个劲儿涨，就是工资不涨。	<物价，涨>	<物价，涨>	84
	5.物价上涨的厉害啊！	<物价，上涨>		

图 3 为 σ_{HowNet} 和 σ_{Coverage} 取值对准确率的影响，左纵坐标为 σ_{HowNet} 取值对应的准确率，右纵坐标为 σ_{Coverage} 取值对应的准确率，横坐标为 σ_{HowNet} 和 σ_{Coverage} 的取值范围。由图 3 可以看出，σ_{HowNet} 的取值对实验结果的影响并不大，因为数据集中属于 HowNet 的登录词过少。另外，σ_{Coverage} 取 0.54 时，准确率最高。综上所述，σ_{HowNet} 和 σ_{Coverage} 的取值分别为 0.53 和 0.54。

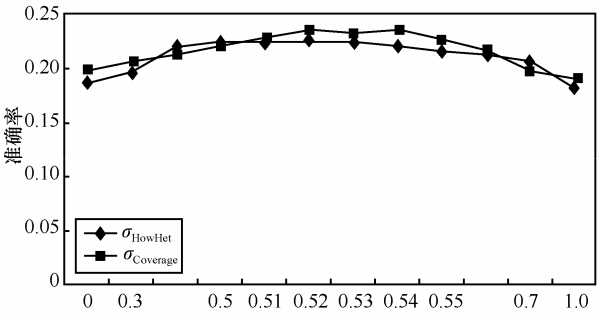


图 3 σ_{HowNet} 和 σ_{Coverage} 取值对实验结果的影响

表 5 中第 5 列为观点词对经过合并后出现频次较高的观点词对。可以看出，这些比较有代表性的

抽取结果基本可以概括大部分用户的观点，可以为舆情分析、新闻媒体以及个人用户提供直观、清晰、无重复的结果。

7　结束语

在对微博热门话题进行观点挖掘过程中，本文充分考虑了微博这一特殊网络媒体的观点表达特点，在观点句识别中加入了表情符号、连词个数、句法依存关系个数这 3 个特征，并首先使用句法依存关系模板进行观点句识别，对于不符合句法依存关系模板的，再进一步利用 SVM 进行观点句识别。在此基础上，将观点句进一步加工成观点词对，这样不仅可以更加清晰地展现微博的观点，而且也便于进行观点句合并。实验给出了采用句法依存关系模板、SVM 以及句法依存关系模板与 SVM 相结合进行观点句识别效果的对比情况，并进一步展示了进行观点词对抽取与合并的过程和结果。在接下来的工作中，将进一步考虑微博用户的 IP 地址、地理位置、发布微博的时间等信息与用户观点的关系，更加准确、有效地挖掘用户观点。

参考文献：

[1] WIEBEM J, *et al*. Development and use of a gold standard data set for subjectivity classifications[A]. Proc 37thAnnual Meeting of the Association, for Computational Linguistics(ACL-99)[C]. NJ: Association for Computational Linguistics, 1999.246-253

[2] 叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法[J]. 信息系统学报, 2007,1(1):79-91.
YE Q, ZHANG Z Q, LUO Z X. Automatically measuring subjectivity of Chinese sentence for sentiment analysis to reviews on Internet[J].China Journal of Information Systems,2007,1(1):79-97.

[3] 蒙新泛, 王厚峰. 主客观识别中的上下文因素的研究[A].中国计算机语言学研究前沿进展(2007-2009)[C]. 山东: 中国中文信息学会, 2009.594-599.
MENG X F, WANG H F. A study on the impact of context information in subjectivity detection[A]. Advances of Computational Linguistics in China(2007-2009)[C]. Shandong: CIPS,2009.594-599.

[4] PANG B, *et al*. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[A]. Proceedings of the association for Computational Linguistics(ACL)[C]. PA: Association for Computational Linguistics, 2004.271-278.

[5] 章剑锋, 张奇, 吴立德等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报,2008,22(2):55-59.
ZHANG J F, ZHANG Q, WU L D, *et al*. Subjective relation extraction in Chinese opinion mining[J]. Journal of Chinese information, 2008, 22(2):55-59.

[6] 张博. 基于 SVM 的中文观点句抽取[D]. 北京: 北京邮电大学, 2011.
ZHANG B. Chinese Opinion Sentence Extraction Based on SVM Classification[D]. Bejing: Beijing University of Posts and Telecommunications, 2011.

[7] 邓乃扬, 田英杰. 支持向量机:理论,算法与拓展[M]. 北京:科学出版社, 2009. 45-50.
DENG N Y, TIAN Y J. Support Vector Machines: Theory, Algorithm and Development[M]. Beijing: Science Press, 2009.45-50.

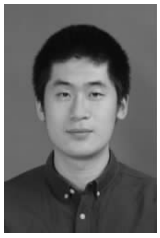
[8] 刘群,李素建.基于《知网》的词汇语义相似度计算[A]. 第 3 届中文词汇语义学研讨会论文集[C]. 2002.59-76.
LIU Q, LI S J. Word similarity computing based on HowNet[A]. 3th ISCID[C]. 2002.59-76.

[9] NLP&CC2012.中文微博情感分析测评[EB/OL]. http://tcci.ccf.org.cn/conference /2012/ pages/page04eva.html, 2012.

[10] DONG Z D, DONG Q. HowNet[EB/OL]. http://www.kenage. com.

[11] KU L W, et *al*. NTUSD[EB/OL]. http://nlg18.csie.ntu.edu.tw:8080/opinion/.

作者简介：



张光磊（1989-），男，北京人，北京信息科技大学研究生，主要研究方向为自然语言处理。

徐雅斌（1962-），男，辽宁锦州人，北京信息科技大学教授，主要研究方向为社交网络、云计算、未来网络。