文章编号:1001 - 9081(2010)11 - 2937 - 04

网络评论倾向性分析

丁建立[1,2,3],慈　祥[1,2,3],黄剑雄[4]
(1. 中国民航大学 计算机科学与技术学院,天津 300300;　2. 中国民航信息技术科研基地,天津 300300;
3. 中国民航大学 智能信号与图像处理天津市重点实验室,天津 300300;
4. 中国国际航空股份有限公司 信息管理部,北京 100071)
(jianliding@yahoo.com.cn; cixiang31415926@126.com;)

摘　要:Web2.0 的兴起使得包括新闻评论、产品评论在内的各种网络评论大量涌现,针对评论信息的监管和利用中的问题多种多样,重点研究其中的网络评论倾向性分析。以知网为基本的语义字典,提出一种改进的词汇相似度计算方法,在此基础上融合同义词词林对词汇的倾向性计算做出改进,进而利用相关语言学知识实现了从细粒度的词汇到粗粒度的评论的倾向性判断。实验表明,该方法对于真实网络环境下的网络评论倾向性分析具有较高的准确率。

关键词:知网;同义词词林;网络评论;倾向性分析
中图分类号: TP311.13;TP391　　**文献标志码**:A

Orientation analysis of Web reviews

DING Jian-li[1,2,3], CI Xiang[1,2,3], HUANG Jian-xiong[4]
(1. *College of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*;
2. *Information Technology Research Base, Civil Aviation Administration of China, Tianjin* 300300, *China*;
3. *Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin* 300300, *China*;
4. *Information Management Department, Air China, Beijing* 100071, *China*)

Abstract: The rise of Web 2.0 leads a lot of Web reviews including news review and product review. Problems during monitoring and using these review information are various, and the orientation analysis of Web reviews was emphasized. HowNet was used as the basic semantic dictionary and an improved method to calculate the word similarity was proposed, then the calculation of word orientation was improved by combining Tongyici Cilin, after that the orientation discrimination was realized from low granularity' word to high granularity' sentence by using language knowledge. The experimental results show that the method has high accuracy to the orientation analysis of Web reviews in the real Internet environment.

Key words: HowNet; Tongyici Cilin; Web review; orientation analysis

0　引言

信息的爆炸式增长给计算机进行信息的自动化处理带来了诸多问题。网络评论的倾向性分析就是其中亟待解决的问题之一,它主要的目标是利用自然语言处理技术对具有主观倾向的网络评论做出正确的倾向判断。网络评论的倾向性分析在舆情分析、内容安全、产品在线跟踪与质量评价、事件分析、企业情报系统等方面有着广泛的应用前景,因此越来越多的学者开始关注这一研究方向。

网络评论的倾向性分析实质上是一个分类问题,即将文档分成正、负两类。从研究的角度来看,主要有基于机器学习和基于统计测度的方法。机器学习方面,Wang 等人[1]对网络上的产品评论进行语义倾向识别,采用了特征选择加朴素贝叶斯的方法,证明该方法比单独使用朴素贝叶斯要好。Kim 和 Hovy[2]在特征的基础上使用最大熵的学习方法来对网络上的产品评论进行语义倾向识别。在基于统计测度的研究方面,刘群和李素建[3]以知网为基本的语义字典提出了一种词汇相似度的计算方法。柳位平等人[4]提出了一种中文情感词语的情感权值的计算方法,并以知网情感词语集为基准,构建了中文情感词典。Turney[5]根据褒贬含义的倾向信息对评论性文章进行分类。

本文利用统计测度的方法,以知网作为基础的语义字典,实现了从词汇到网络评论不同粒度层次的语义倾向性计算。

1　基于知网和同义词词林的词汇倾向性分析

1.1　基于知网的词汇相似度计算的改进

词虽然不是汉语中的最小语素,但它却是分析整个评论倾向的基础,因此在分析整体倾向性之前必须借助某种手段来对这些词的倾向做出正确的判断。

词汇的相似度在文本检索、信息抽取、文本分类等方面有着广泛的应用,但是词汇相似度本身就是一个很模糊的概念。文献[3]中认为词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。在实际的应用过程中,可以利用词汇之间的距离来计算相似度,距离越近,相似度越大。

根据知网的作者董振东先生的说法[3]:《知网》是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常

收稿日期:2010 - 05 - 21;**修回日期**:2010 - 07 - 21。
基金项目:国家自然科学基金资助项目(60879015);国家 863 计划项目(2006AA12A106)。
作者简介:丁建立(1963 -),男,河南洛阳人,教授,博士,主要研究方向:网络安全;　慈祥(1986 -),男,安徽巢湖人,硕士研究生,主要研究方向:网络安全;　黄剑雄(1970 -),男,北京人,高级工程师,主要研究方向:民航信息系统安全。

识知识库。文献[3]根据知网的结构特点将两个词汇的相似度计算分成了 4 个部分,因此总的相似度由 4 部分加权平均而成。即:

$$Sim(S_1,S_2) = \sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \tag{1}$$

每个部分的具体计算公式如下:

$$Sim_i(S_1,S_2) = \frac{\alpha}{Dis_i(S_1,S_2)+\alpha} \tag{2}$$

其中:$Dis_i(S_1,S_2)$ 表示词汇 S_1、S_2 第 i 部分的相对应义原在知网义原树上的距离,α 是一个可调的参数。但是这种计算方式只是单纯的计算了两个义原在形式上的距离,并没有考虑到义原树本身所包含的一些信息。

一般来说,知网义原树上节点层次越深,节点之间距离越近;节点所在区域密度越大,节点之间的距离越近。文献[6-7]从不同的角度对义原之间距离的计算方式做出了改进,大致的出发点也都是考虑了义原树中节点的层次和密度信息。但是这些计算方式稍显复杂,为了能够适应本文的倾向性计算,对于义原之间距离的计算方式做出如下的简单改进:

设义原树中的根节点为第 0 层,义原 i 所处的层次为 $level_i$,其所拥有的兄弟节点数为 $brother_i$,整个义原树的深度为 h,则义原树中除根节点外的任意义原 i 和其父节 $father(i)$ 的距离为:

$$distance(i,father(i)) = \frac{h}{level_i}\times\frac{1}{brother_i+1} \tag{3}$$

式(3)综合考虑了节点层次和节点密度信息,同时较文献[6-7]的计算方式更为简便。此时词汇相似度公式变为:

$$Improved\_Sim(S_1,S_2) = \sum_{i=1}^{4}\beta_i Improved\_Sim_i(S_1,S_2) \tag{4}$$

其中:

$$Improved\_Sim_i(S_1,S_2) = \frac{\alpha}{Improved\_Dis_i(S_1,S_2)+\alpha} \tag{5}$$

$Improved\_Dis_i(S_1,S_2)$ 的距离计算方式和原来的一致,但是义原树中父义原和子义原之间的距离不再是 1 而是按式(3)计算得到。具体的实验结果及分析参见实验 1。

1.2 基于知网和同义词词林融合的词汇倾向性计算

词汇中常常包含着使用者的主观情感,这些具有主观情感的词往往最能够代表使用者的情感,因此必须借助某种手段对一些主观性的词语的情感倾向做出判断。目前基于语义的方法一般做如下处理:

1)选定一组正面/负面情感词表,句子中出现这些词语时直接赋予其相应的倾向性。

2)对于未出现在情感词表中的词则需要进行判断。通常做法是选定一些基准词,包括褒义基准词和贬义基准词,然后利用某种手段来计算未知词和这两组基准词的相似度,通过二者差值来确定未知词和哪组词更加相似,进而确定未知词的倾向性。具体来说就是:

$$orientation(word) = pos(word) - neg(word) \tag{6}$$

其中:$pos(word)$ 表示未知词和所有褒义基准词的相似度之和,$neg(word)$ 表示未知词和所有贬义基准词的相似度之和。如果 $orientation(word) > 0$ 则表示该词一般具有正面倾向,否则具有负面倾向。

知网是一个主观性的语义词典,它的构建本身就具有很强的主观性,由于某个词语在知网结构中处于何种层次直接决定了其和基准词的语义距离,但是这种层次的选择有时候不一定是正确的,这就会导致词的倾向性计算错误,进而可能直接导致整个文档倾向性的判断错误,因此直接使用知网计算未知词和基准词的相似度不太合适。对此本文结合同义词词林做出了一些改进。改进的出发点是同义词基本上具有相同的倾向性,计算某个词的倾向性时不单单是计算它和褒贬基准词的相似度,而是计算该词以及所有和它同义的词与褒贬基准词的相似度。这样可用一群词来有效的规避单个词计算时出现的误差,即:

$$new\_orientation(word) = new\_pos(word) - new\_neg(word) \tag{7}$$

其中:

$$new\_pos(word) = pos(word) + pos(synonyms(word)) = pos(word) + \sum_{i=1}^{n} pos(sw_i) \tag{8}$$

$$new\_neg(word) = neg(word) + neg(synonyms(word)) = neg(word) + \sum_{i=1}^{n} neg(sw_i) \tag{9}$$

$synonyms(word) = (sw_1, sw_2, \cdots, sw_n)$ 表示 $word$ 的同义词集合,本文采用的是哈尔滨工业大学信息检索研究室的同义词词林扩展版[8]作为标准同义词集。如果 $new\_orientation(word) > 0$ 则表明该词具有正面倾向,反之具有负面倾向。具体的实验及相关分析见实验 2。

2 网络评论的倾向性分析

一般的网络评论可以看做是一篇较短的文档,下面将分别阐述句子的倾向性判断以及如何对句子倾向性进行合成。

2.1 汉语句子的倾向性分析

汉语句子分为单句和复句,单句的倾向性判断相对容易,对复句的判断则较为复杂。从语义上来看,只有递进和转折关系的复句前后句之间的情感发生了明显的改变,因此在进行倾向性判断时,对这两种句型需特别注意。一般说来复句的每个分句情感强度由其具有情感的词确定,词的倾向性由上文提到的方法确定,而倾向性的权值即情感强度则由修饰这些词的副词确定,表 1 列出了本文用到的一些程度副词[9]。根据需求将其分成了 4 个等级,即低量、中量、高量和极量。分别赋予 0.5,1,1.5 和 2 的权值。这些词和其修饰的情感词共同确定了该情感词在其所属分句中的情感倾向。

表 1 程度副词的感情等级

副词 w_j	感情等级 $deg(w_j)$
稍、稍稍、稍微、稍为、稍许、略、略略、略微、略为、些微、多少、有点、有些	0.5(低量)
较、比较、较比、较为、还、不大、不太、不很、不甚	1(中量)
更、更加、更为、更甚、越、越发、备加、愈、愈加、愈为、愈益、越加、格外、益发、很、挺、怪、老、非常、特别、相当、十分、好、好不、甚、甚为、颇、颇为、异常、深为、满、蛮、够、多、多么、殊、特、大、大为、何等、何其、尤其、无比、尤为、不胜	1.5(高量)
最、最为、太、极、极为、极其、极度、极端、至、至为、顶、过、过于、过分、分外、万分	2(极量)

汉语中还有一些否定词,这些否定词和情感词搭配使用时,词的情感发生逆转,这是在确定分句倾向性时需要特别注意的。综合考虑情感词、否定词、程度副词和复句类型之后就可以确定一个句子的倾向性。设 $sentence = (s_1, s_2, \cdots, s_m)$, $s_i = (w_1, w_2, \cdots, w_n), i = 1, 2, \cdots, m$。其中 $sentence$ 表示一个复句,s_i 是该复句的一个分句,而 $(w_1, w_2, \cdots, w_n)$ 则表示分句 s_i 中可能具有情感的词汇组成的一个集合。此时:

$$orientation(sentence) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i s_i\right) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i \left(\sum_{j=1}^{n} \beta_j w_j\right)\right) \tag{10}$$

$$\beta_j = \begin{cases} deg(w_j), & w_j \text{ 无否定词修饰} \\ -deg(w_j), & w_j \text{ 有否定词修饰} \end{cases} \tag{11}$$

$$\alpha_j = \begin{cases} 2, & s_j \text{ 为递进或转折关系句子的后半句} \\ 1, & \text{其他} \end{cases} \tag{12}$$

其中:sign 是符号函数,$deg(w_j)$ 表示修饰情感词 w_j 的副词的感情等级。

2.2　句子倾向性的合成

所有单句和复句的倾向行确定之后,需要以一定的标准将其合成为整个评论的倾向性。在对网络评论的特点进行分析之后我们确定了如下的基本原则用于句子权重选择:

1)正常情况下所有句子的权重一致,即所有句子对于评论倾向性的贡献是一致的。

2)对于具有强烈感情的句子赋予更高的权值,这里主要是指疑问句和感叹句。

3)评论中如果是总结性的句子,则直接以该句的倾向性代替整句的倾向性。

根据上面的权重选择原则,最终确定一个评论的倾向性计算方法如下:

$$document = \begin{cases} orientation(sentence_i), \\ \quad sentence_i \text{ 是总结句} \\ \sum_i orientation(sentence_i) + 2\sum_j orientation(sentence_j), \\ \quad sentence_j \text{ 为疑问句或感叹句} \end{cases} \tag{13}$$

其中:$i + j = K$,K 表示文档中总的句子数。当 $document > 0$,表明该评论具有正面倾向;否则具有负面倾向。

3　系统基本架构设计

图 1 是系统的基本设计架构。从图中不难看出,整个系统的核心部分是词语倾向性的计算,具体的处理步骤如下:

1)对网上采集到的评论进行分词,评论被保存在文本文件中,分词利用的是中国科学院的 ICTCLAS4J。通常对于分词后的结果需要去除停用词,但是系统的后续分析需要利用到标点符号和大量的副词,因此保留停用词。

2)分词后的结果首先需要进行词性的分类。一般来说,在文本倾向性的分析中,可能产生倾向性的只有形容词、动词和名词。其中最有用的是形容词,其次是动词和名词。副词对于情感等级判断至关重要,所以也将它单独分出。对于一般不予考虑的标点符号,本文认为问号和感叹号常常代表作者的某种强烈语气,予以保留并将其分至标点符号类,辅助副词来对整个句子的情感等级做出判断。除了这些之外,其余的词性不予考虑。

3)形容词、动词和名词是本文判断倾向性的核心词类,但是这三者的处理方式并不完全相同。对于形容词,首先判断其是否出现在由知网系统提供的正负情感词表中,若在则直接赋予其相应的情感。若发现形容词未出现在正负情感词表中,则需要利用上文中提出的 new_orientation 算法进行计算。在基准褒贬义词的选择上,出于系统效率的考虑,经过反复测试最终选定只使用两个基准褒贬义词:“好”和“坏”。由于基准的正负面情感词表是精心挑选出来的具有强烈情感的词语,通过 new_orientation 算法计算出来的词的情感强度不会超过这些基准情感词,因此在计算后对于计算得到倾向性的词语统一赋予 0.5 的加权,即将其情感强度降至低量级。相比较而言,大部分的动词和名词都不具有倾向性,因此对于分词后的动词和名词只有当其出现在正负情感词表中时才认为其具有相应的倾向性,否则并不利用 new_orientation 算法进行计算,这是有别于形容词的地方。

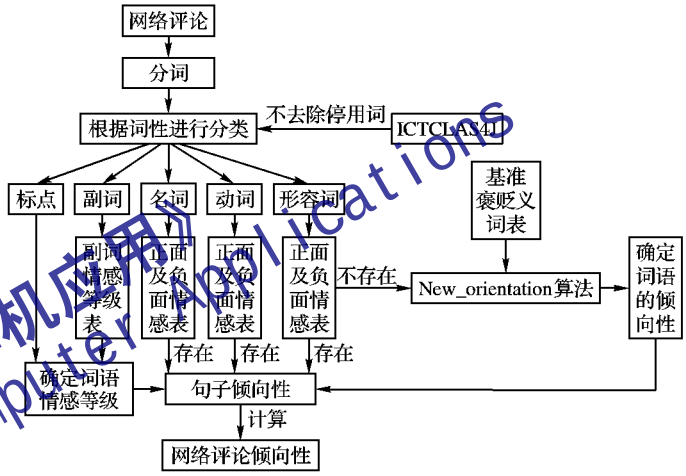

图 1　系统基本设计架构

4)计算完上述 3 类词语的倾向性后,结合程度副词和否定词判断出每个形容词、动词和名词所属单句的倾向性。

5)将单句的倾向性合成为复句的倾向性,其中特别需要注意转折和递进关系的单句。

6)将所有的句子倾向性合成为整个评论的倾向性,输出结果,其中特别需要注意疑问句、感叹句和总结句。

4　实验及结果分析

为了验证本文提出的两个算法及最终评论倾向性计算系统的可行性,共设计了 3 组实验分别进行验证。

实验 1 主要是用来验证本文的词汇相似度改进算法的有效性。为了比较实验结果,本文选取了文献[3]中的一些比较典型的词语并加入了一些新的词语进行比较。

表 2 是采用两种不同方法计算词汇相似度结果的比较。为了便于比较,主要实验参数仍沿用文献[3]中的设置,即 $\alpha = 1.6, \beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13$。目前判断词汇相似度并没有一个通用的标准,因此本文利用人工进行判断。从表 2 中可以看出,两种算法的计算结果彼此相差不是很大,基本都在一个较小的范围,但是本文算法相比文献[3]算法更趋于合理。例如:“男人”和“收音机”、“男人”和“鲤鱼”、“男人”和“篮球”、“男人”和“图书”这些从日常经验来看都不应当有较大相似度的词汇,在本文算法中其相似度值较文献[3]算法均有所下降,比较合理。另外在文献[3]中

指出,“中国”和“联合国”以及“中国”和“安理会”的相似度计算都偏小,而这两组词利用本文提出的本文算法所计算出的相似度值较文献[3]算法均有所上升,这进一步证明了本文方法的有效性。

表 2 词汇相似度实验结果对比

词语 1	词语 2	文献[3]算法	本文算法
男人	女人	0.861 111	0.689 004
男人	父亲	1.000 000	1.000 000
男人	收音机	0.111 628	0.068 602
男人	鲤鱼	0.208 696	0.151 766
男人	苹果	0.171 429	0.122 195
男人	坏	0.119 601	0.105 371
男人	篮球	0.132 554	0.040 406
男人	图书	0.122 997	0.045 533
工人	教师	0.722 222	0.792 306
工人	科学家	0.575 926	0.686 567
教师	运动员	0.722 222	0.886 478
科学家	农民	0.575 926	0.686 567
中国	美国	1.000 000	1.000 000
中国	联合国	0.136 434	0.394 037
中国	安理会	0.113 580	0.259 503
中国	欧洲	1.000 000	1.000 000
思考	考虑	1.000 000	1.000 000
跑	跳	0.444 444	0.282 360
跑	跳舞	0.126 984	0.069 298
鲤鱼	苹果	0.242 424	0.111 329

实验 2 是用来验证本文提出的 new_orientation 算法的有效性。本文的思路是:先选定某个情感词语集,然后利用 orientation 和 new_orientation 算法逐一计算并比对效果。这里选用的词语集是知网的正面情感词语集,因为这些词都是经过精心挑选的具有确定倾向的词语,褒贬义基准词仍选用“好”和“坏”。经过实验发现大部分词语的倾向性判断两种算法都是一致的,但是也有一些词在计算二者的结果时出现了偏差,表 3 列出了部分出现偏差的词语。

表 3 词汇倾向性实验结果

词语	orientation	new_orientation
称羡	-0.048 379	0.015 991 0
伸张	-0.048 379	0.591 009 0
叹羡	-0.048 379	0.015 991 0
希罕	0	0.607 771 0
犒赏	-0.041 545	0.608 484 0
交口称誉	-0.048 379	0.803 142 0
甜滋滋	无	1.386 328 0
情有独钟	无	0.600 704 0
爱上	无	0.600 703 4
豁朗	无	1.311 148 0

表 3 中词语很明显的具有正面倾向,但是有些词利用 orientation 算法计算出的结果不为正值,也就是说具有负面倾向,这显然是错误的。而对这些词语利用本文提出的融合同义词词林的 new_orientation 算法计算值均为正值,也就是说具有正面倾向,这是符合实际情况的。另外如“甜滋滋”、“情有独钟”、“爱上”、“豁朗”等词在义原中是没有的,如果利用原来的方法显然无法判断其倾向性,但是本文提出的算法却准确的利用其同义词计算出其具有正面倾向。总的来说,本文的 new_orientation 较原始的 orientation 算法无论是在计算的准确性还是对于义原未收录词的处理方面都有所改进。

实验 3 是用来验证本文提出的网络倾向性判断方法的有效性。目前倾向性判断并没有一个标准的测试集,本文选用中国科学院谭松波博士收集的一个关于酒店服务的网络评论集[10]。这个集合名为 ChnSentiCorp_htl_ba_2000,其中共包含了 2 000 条评论,正负倾向各 1 000 条。但是经过测试发现这个数据集存在着一些问题,除了有不少评论数据重复之外还有部分评论的倾向性分类错误,实际上正倾向性评论有 964 个,负面评论 1 036 个。

实验评价指标仍选用准确率 P、召回率 R 和 F 值 3 个指标。最终的实验结果如表 4 所示。

表 4 评论倾向性实验结果

类别	准确率 P/%	召回率 R/%	F 值/%
正面倾向评论	75.30	69.20	72.10
负面倾向评论	73.30	78.60	75.90

从表 4 来看,结果还是比较好的,各项指标基本都接近或超过了 70%,这说明系统在真实网络环境下的效果测试效果还是不错的,具有一定的实用价值。

5 结语

机器学习和统计测度在倾向性分析上各有优点,但是统计测度的倾向性分析可以灵活地控制分析内容的粒度层次。下一步的工作将是如何从混杂有多个观点的文本中准确提取主要观点,尽可能地消除无关话题的干扰,进一步提高系统的准确率。

参考文献:

[1] WANG CHAO, LU JIE, ZHANG GUANGQUAN. A semantic classification approach for online product reviews [C]// Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC: IEEE, 2005: 276 - 279.

[2] KIM S M, HOVY E. Automatic identification of pro and con reasons in online reviews [C]// Proceedings of the COLING/ACL on Main conference poster sessions. Morristown, NJ: Association for Computational Linguistics, 2006: 483 - 490.

[3] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[C]// 第三届汉语词汇语义学研讨会. 台北:[s. n.],2002.

[4] 柳位平,朱艳辉,栗春亮,等. 中文基础情感词词典的构建方法研究[J]. 计算机应用,2009,29(10):2875 - 2877.

[5] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2002: 417 - 424.

[6] 蒋溢,丁优,熊安萍,等. 一种基于知网的词汇语义相似度改进计算方法[J]. 重庆邮电大学学报:自然科学版,2009,21(4):533 - 537.

[7] 李峰,李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007,21(3):99 - 105.

[8] 哈尔滨工业大学信息检索研究室. 同义词词林扩展版[EB/OL]. [2010 - 04 - 05]. http://ir.hit.edu.cn/.

[9] 张锦明. 中文语义倾向识别的关键算法研究[D]. 北京:北京邮电大学,2008.

[10] 谭松波. 中文情感挖掘语料——ChnSentiCorp[EB/OL]. [2010 - 05 - 01]. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm.