基于聚类引擎的话题褒贬度计算

李斌[1] 卢俊之[1] 章成志[2] 陈小荷[1]

（1. 南京师范大学文学院，南京 210097； 2. 南京大学信息管理系，南京 210093）

Email: gothere@126.com

摘 要：互联网是人们表达各种观点的重要媒介，自动获取网络上对话题的褒贬态度是自然语言处理的一项新兴课题。本文提出了利用两种搜索引擎进行话题褒贬态度计算的方法。首先，利用聚类引擎近似地得到话题的若干子话题。然后，使用 PMI 算法利用关键词检索的搜索引擎计算出子话题的褒贬度，进而利用多语搜索引擎和地区搜索观察同一话题的跨语言分布和地区分布情况。该方法可用于搜索结果优化、话题分析、产品跟踪等领域。

关键词：中文信息处理，搜索引擎，搜索结果聚类，褒贬态度

Semantic Orientation of Topics Based on Clustering Engines

LI Bin[1], LU Junzhi[1], ZHANG Chengzhi[2], CHEN Xiaohe[1]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097)

(2. Department of Information Management, Nanjing University, Nanjing 210093)

Email: gothere@126.com

Abstract: The Internet is an important media for people to express their standpoints on specific topics. Thus, getting the semantic orientation (SO) of people's comments became a new task in NLP research filed. This paper proposed a new method to compute the SO value by 2 kinds of search engines. First, the clustering search engine is used to get the subtopics of the given topic represented by the user's search query. Then, Keyword in Context (KWIC) search engine is used to get the SO values of subtopics by PMI algorithm. The method can be applied to search result optimization, topic analysis and products tracking, etc.

Key words: Chinese Language Processing, Search Engine, Search Results Clustering, Semantic Orientation

1 引言

随着互联网的发展，网络已经从单一的信息传递模式转向了丰富多彩的网络生活。人们可以在网络上针对不同话题发表自己的言论和种种看法。社区、博客、论坛、评测网站、在线的新闻评论系统，都成为表达观点的平台。话题的褒贬态度计算，就是计算网络上人们对于特定话题的褒贬态度，即，对于某人、某事、某物的态度，是赞扬的还是批评的，是赞成的还是反对的。而话题的褒贬度，是指对于给定话题褒贬程度的一个度量值。态度计算可以服务于政府决策、舆论监测、民意调查、企业产品跟踪等重要领域。

目前，褒贬态度计算是自然语言处理的一个新兴热点，已有的研究主要集中在三个层级，即词语、句子和语篇的褒贬度。词语和句子的研究主要是为计算语篇的褒贬态度服务的。词语的褒

基金资助：南京师范大学学生科学基金（面向因特网的中文文本褒贬态度自动分类研究）

2006 年江苏省研究生创新工作项目（主题聚类及其应用）

作者简介：李斌（1981-），男，博士生，主要研究方向为计算语言学.

贬度是基础，句子是褒贬度的具体阐发者，语篇则是评论者——话题——褒贬态度的综合体。同时，褒贬态度总是与话题联系在一起的。比如用户在谈论一部电影的好坏时，语篇中往往会出现不同评论者对于电影的若干方面（即子话题）的评价，如人物、剧情、画面等。因此，在篇章中，态度与评论者、话题和子话题密不可分。面向语篇的态度计算，应该是一个复杂的过程，涉及话题提取、未登录词识别、词义消歧、句法语义分析、指代消解等问题。大量文本计算的结果就可以得出话题的态度分布、评论者分布和子话题的分布，可以应用于许多领域。然而，已有的研究和技术大都还只能处理短小的、态度鲜明的单话题语篇，大多只限于计算其整体的褒贬态度。

本文考虑以一种简便快捷的方式计算话题及子话题的褒贬度。首先，获得话题的若干重要的子话题，然后利用计算词语褒贬度的方法，把话题和子话题分别进行褒贬度的计算，最后得到话题的褒贬度。为了得到话题的子话题，我们使用了聚类引擎来得到话题的前 N 个聚类标签，把这些标签近似地看作子话题。由于互联网的数据是不断更新的，利用聚类引擎，就可以知道一个话题的最热门的子话题，大致地看出网络媒体对这些子话题的评价。进一步地，利用不同语种的搜索引擎和地区搜索功能，还可以观察到不同语种和不同地区对于同一个话题的褒贬态度差异。利用搜索引擎进行词语的褒贬度计算的好处是，系统可以自动地计算用户输入的任意查询串，能够较好地解决未登录词问题。

2 相关工作

目前，词语的倾向性计算已经有了一些成果。Hatzivassiloglou & McKeown（1997）使用了形容词的词缀和连用时所使用的连词来计算其褒贬度，精确率较高，但局限于形容词的计算。Turney（2003）利用基准褒贬词对的 PMI 算法，可以基于本地语料库或网络搜索引擎进行褒贬度计算。该方法可以处理各种类型的词语，在包含形容词、副词、名词、动词的测试集上准确率达到 82.8%。Turney（2003）还使用了极性语义分析（LSA）的方法，但是计算量过大，只适用于小规模语料库。Yuen et al.（2004）在处理中文时，对 PMI 算法进行了改进，指出在分词语料库上使用褒贬语素得到的结果更好。朱嫣岚 等（2006）利用知网（HowNet）收录词语的概念定义项，通过计算和基准褒贬词语的语义相似度和相关场的方法得出词语的褒贬度，其缺陷是无法处理未登录词。一个话题往往表示为词语形式，因此计算词语褒贬度的方法可以直接应用于话题褒贬度的计算。

在话题及子话题态度计算方面，Yi et al.（2003）改变了过去只计算一篇评论的整体褒贬度的做法，以单篇评论为单位，基于评价句模板获取文本中话题的各个子属性及评论，从而使每个评论都与话题相关联。在话题褒贬度计算的结果表示方面，Fujii et al.（2006）提出了一种可视化方法，在标注好话题、子话题和褒贬度的语料库基础上，可以把话题的关注点排列在重要性和褒贬度两个维度构成的平面中，形成直观的可视化图形界面，但并没有对褒贬度计算本身提出解决方案。姚天昉 等（2006）、苏祺（2006）都使用了领域知识本体的方法，建立有关汽车的知识本体，然后利用句法分析或者句子模板方法，把对汽车各种属性的评论映射到相应的知识本体结点上，从而给出用户对于汽车各方面的褒贬态度。这样做的好处是只处理文本中与知识本体相关的内容，缺点是过于依赖知识本体和映射规则，需要大量的人力投入，对于语料的处理速度过慢。

在应用系统方面，微软美国研究院 Gamon（2005）的 Pulse 系统，可以自动计算出网络用户发表的关于汽车评论的褒贬态度。美国伊利诺斯大学的 Liu（2005）开发了 Opinion Observer 系统，处理客户对产品的网络评价，对产品（如数码相机等）的各种子特征的优缺点进行统计，以可视化方式显示出来。IBM Almanden 研究中心 Yi（2005）的 WebFountain 系统中的意见挖掘器，可以对数码相机和音乐的评论进行分析计算。这些系统大都依托于特定的网站和特定种类的产品评论，需要大量的人工标注样本，还处在实验阶段。

3 算法

3.1 系统流程

本文提出了一个简单高效的算法，利用聚类引擎近似地得到话题的子话题，再进行褒贬度计算。首先，使用聚类引擎得到话题的前 N 个聚类标签，再把这些标签近似地看作子话题进行褒贬度计算。系统流程相应设计如图 3-1 所示。

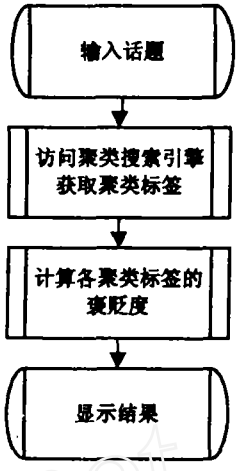

图 3-1：系统流程图

3.2 词语褒贬度的 PMI 算法

本文使用了 Turney(2003)提出的利用搜索引擎计算词语褒贬度的 PMI 算法。首先，定义两个词语之间的点互信息（Pointwise Mutual Information, PMI）为：

$$\mathrm{PMI}(word_1, word_2) = \log_2\left(\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)}\right) \quad (1)$$

其中，$p(word)=hits(word)/N$，N 是搜索引擎索引的页面总数。$P(word_1 \& word_2)$是 $word_1$ 和 $word_2$ 的同现概率，$p(word_1)$和 $p(word_2)$是 $word_1$ 和 $word_2$ 分别出现的概率。$hits(w)$表示对于查询串 w 搜索引擎返回的网页数，N 是搜索引擎数据库中的文档总数。对于 $hits(w)$的零值，采取了简单的加一法。利用 PMI 值可以衡量两个词语之间的相关程度。一个词语的褒贬度（SO_PMI）就定义为该词语与一组褒义词互信息之和与一组贬义词的互信息之和的差。

$$\mathrm{SO\_PMI}(word) = \sum_{pword \in Pwords} \mathrm{PMI}(word, pword) - \sum_{nword \in Nwords} \mathrm{PMI}(word, nword) \quad (2)$$

pword 表示褒义词集合 *Pwords* 中的一个褒义词，*nword* 表示贬义词集合 *Nwords* 中的一个贬义词。公式（3）是一个 log 形式的比值。

$$\mathrm{SO\_PMI}(word) = \log_2\left(\frac{\prod_{pword \in Pwords} hits(word \ \mathrm{AND} \ pword) \cdot \prod_{nword \in Nwords} hits(nword)}{\prod_{pword \in Pwords} hits(pword) \cdot \prod_{nword \in Nwords} hits(word \ \mathrm{AND} \ nword)}\right) \quad (3)$$

Turney（2003）采用的是 NEAR 运算。NEAR 运算是某些搜索引擎（如 Altavista）提供的，可以把两个查询串在 10 个词的窗口内同现作为查询条件。由于大多数搜索引擎不提供该功能，本文只采用 AND 运算。AND 运算是普通的搜索引擎提供的布尔检索式，表示是两个查询串在单篇文档中同现。根据 SO_PMI 的值就可以判定词语的褒贬度，正值为褒义，负值为贬义，绝对值是褒贬程度。Turney（2003）对英文查询串使用的 7 对褒贬词语如下。

Pwords = {good, nice, excellent, positive, fortunate, correct, superior}

Nwords = {bad, nasty, poor, negative, unfortunate, wrong, inferior}

从公式（3）计算一个词语的语义倾向需要向搜索引擎发出 28 个查询串。由于 *hits（nword）* 和 *hits(pword)* 可以通过预计算得到，只需查询一次，所以查询量只有 14 个。Yuen（2004）把该方法用在中文词语的倾向性计算上，人工选择了 10 对反义词作为的基准词对：

Pwords ={诚实,聪明,充足,幸运,正确,优秀,兴盛,善良,英勇,谦虚}

Nwords ={虚伪,愚蠢,短缺,不幸,错误,恶劣,衰落,残暴,懦弱,傲慢}

我们使用 Yuen（2004）的基准词对，利用中文搜索引擎百度进行计算，测试了其对中文褒贬词语的判断效果。测试使用的是《汉语褒贬义词语用法词典》，共收录褒贬词型 1013 个，其中褒义词语 499 个，贬义词语 514 个。测试结果有 823 个词语计算正确，正确率为 81.2%。可见，PMI 方法对于汉语词语也有着不错的判定效果。

3.3 子话题的获取方法

本文利用聚类引擎来获取给定话题的子话题。近年来发展迅速的聚类引擎，是一种采用了聚类算法的元搜索引擎。这种引擎根据 Google、Yahoo 等各大搜索引擎返回的结果进行聚类，给出与查询串最相关聚类标签，使用户可以快速了解搜索结果的整体分布情况。国外代表性的英文聚类引擎有 Vivísimo（www.vivisimo.com）、Carrot2（www.carrot2.org）等，中文聚类引擎则以 BBMao（www.bbmao.com）为代表。聚类引擎所给出的类别标签，往往是与查询词高度相关的专名和事件、查询词的组成要素、不同义项等。如，在聚类引擎 BBMao 中查询“戴尔笔记本”，可以得到“电池”、“起火”、“报价”、“处理器”等聚类标签。查询“熊猫”时，可以得到“熊猫烧香”、“南京熊猫”、“熊猫手机”、“野生”等聚类标签。相对于领域知识本体的方法而言，使用聚类的好处有两点：

（1） 适应话题的多样性。不同的话题，其相关属性、子话题是不固定的。如，汽车产品往往是外观、马力、油耗、刹车等方面，而数码相机产品则是外观、镜头、像素、LCD 等方面，所以基于知识本体的方法需要花费较大人力来建立不同领域的知识和资源。使用聚类引擎则可以迅速得到话题的若干重要方面，或是与话题高度相关的一些方面。

（2） 适应话题的动态性。知识本体方法不易于表现话题新增要素的动态变化，同时，对新话题的分类也较为困难。如，电子产品往往会出现一些新的品牌、型号和功能，这些都需要人工地对知识本体进行修改和补充，较为繁杂。使用聚类引擎则可以及时反映互联网上有关某一话题的最受关注的方面，避免这种问题的出现。

4 实验结果及分析

4.1 实验方法

针对一个话题，我们采用先聚类再计算 PMI 值的方法。使用中英文聚类引擎得到聚类结果，对于每个聚类标签使用 PMI 算法计算其的褒贬度，中英文使用的资源如表 4-1 所示。

语种	褒贬词对集合		计算 PMI 的搜索引擎	聚类引擎
英文	褒义词	good, nice, excellent, positive, fortunate, correct, superior	Google	Vivísimo
	贬义词	bad, nasty, poor, negative, unfortunate, wrong, inferior		
中文	褒义词	诚实,聪明,充足,幸运,正确,优秀,兴盛,善良,英勇,谦虚	百度	BBMao
	贬义词	虚伪,愚蠢,短缺,不幸,错误,恶劣,衰落,残暴,懦弱,傲慢		

表 4-1：中英文褒贬度计算使用资源

我们以百度风云榜（http://top.baidu.com）等搜索热榜计算了人们关心的不同话题的褒贬度。由于数据过多，下面仅给出聚类计算、跨语言聚类和地区分布的个别计算结果。

4.2 实验结果

4.2.1 聚类结果

对于一般的话题，可以通过聚类引擎和 PMI 算法，得到话题及其子话题的褒贬度。子话题的个数限制为前 10 个（下同）。保健类产品一直是消费者关心的话题，我们选择了较有名气且饱经争议的品牌“脑白金”为例（表 4-2，2007 年 4 月 2 日计算）。从表中可以看出，中文网页对“脑白金”的总体评价是呈弱负性的，分十个子方面来看，则褒贬不一。看来，网络上虽然充满了对脑白金的各种正面的商业宣传，依然无法抵抗广大消费者的批评，无论是整体评价还是在子话题的评价上都有负面评价。

话题+聚类标签	褒贬度
脑白金	-0.455390
+品牌	0.039239
+收礼	1.675817
+分泌	0.839708
+白金营销	6.618273
+营销策略	6.887032
+黄金搭档	-0.371230
+白金产品	-3.499800
+白金策划	6.190640
+传播	0.575927
+改善睡眠	6.294158

表 4-2：话题聚类结果

4.2.2 跨语言聚类

对于同一个话题，我们可以使用中英文引擎分别计算褒贬度，以观察国内外的网络评价。系统调用了 Google 提供的在线翻译服务，可以自动地将中英文词语互译。如，“陈水扁”（表 4-3，2007 年 4 月 15 日计算）。通过该表可以看出中英文聚类引擎对“陈水扁”的聚类结果是不同的。中文网页在总体上和十个子方面对陈水扁是非常否定的，而英语网页对其也呈批评态度，只有在“民进党”方面呈较弱肯定。

（中）话题+聚类标签	褒贬度	（英）话题+聚类标签	褒贬度
陈水扁	-5.90192	Chen Shui-bian	-6.31256
+台独	-2.49193	+Taipei	-6.25145
+两岸关系	-5.09006	+Asia	-9.84001
+民进党	-1.15946	+Politics	-9.72348
+国民党	-1.47531	+Republic, China	-2.02246
+罢免	-3.71243	+Independence	-4.08236
+美国	-1.96608	+Taiwan leader Chen Shui-bian	-6.67944
+丑闻	-0.95946	+Government	-9.45096
+李登辉	-4.58772	+Remarks	-5.54497
+台海	-5.27436	+Democratic Progressive Party	0.68043
+真面目	-3.51746	+Inauguration	-5.01526

表 4-3：跨语言聚类结果

4.2.3 地区分布差异

使用百度提供的高级搜索功能——按地区搜索，我们选择了网站数量最集中而且观点较具有代表性的三个地区，分别是北京、上海和港澳台地区。增加了地区限制以后，就可以观察对同一话题的不同地区的网络评价态度。以萨达姆为例（表 4-4，2007 年 4 月 2 日计算），萨达姆在中文网页上的褒贬度差异明显，我们从有关的媒体报道就可以得到解释。整体上看，网络媒介对萨达姆是持批评态度的，北京地区对萨达姆较为同情，上海地区则有所下降， 而港澳台地区批评较多。

在地区分布差异计算上，也可以使用聚类引擎，从而观察每个子话题在不同地区的褒贬度。不仅如此，还可以使用搜索引擎的其他高级搜索选项以得到褒贬态度的分布情况，如：限定时间搜索、站内搜索、限定语体搜索（新闻、博客、论坛）等等，此处不再展开。

话题_地区	褒贬度
萨达姆_全部结果	-0.768030
萨达姆_北京地区	4.394206
萨达姆_上海地区	0.745471
萨达姆_港澳台地区	-4.408276

表 4-4： 地区聚类结果

5 结论与未来工作

本文提出了利用两种搜索引擎进行话题褒贬度计算的方法。首先，利用聚类引擎近似地得到

话题的若干子话题。然后，使用PMI算法利用关键词检索的搜索引擎计算出子话题的褒贬度，进而利用不同语言的搜索引擎和地区搜索来观察话题态度的跨语言分布和地区分布。该方法简便快捷，对于一些典型的话题可以得到较好的实验结果，可以应用于相关机构和企业，对于打击各类野广告和不良信息有着重要的作用，也可以使广大消费者及时了解网络上对各种事件尤其是产品信息的评价，帮助人们进行消费决策，具有较大的经济效益和社会效益。

目前存在的问题主要有两个：（1）计算的精度还不够高，对话题褒贬度的计算结果不好解释，尤其是对不太熟知的话题，正确与否很难进行人工判断。（2）无法指出观点态度的持有者是谁。在以后的研究中，我们考虑收集大量的评论性文本，在本地语料库中进行聚类和褒贬度计算。一方面可以较好地利用分词和词性等信息，一方面可以采用先获取文本中的“话题——态度——持有者”这样的三元组，再对这些三元组进行自动聚类，使得褒贬度的计算不再脱离话题和持有者。

参 考 文 献

[1] Atsushi Fujii and Tetsuya Ishikawa. A System for Summarizing and Visualizing Arguments in Subjective Documents: Toward Supporting Decision Making [A]. In: *Proceedings of COLING-ACL Workshop on Sentiment and Subjectivity in Text*, 2006:15-22.

[2] B. Liu, M. Hu, J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In: *Proceedings of WWW'05, the 14th International Conference on World Wide Web*, Chiba, Japan, 2005:342-351.

[3] J. Yi and W. Niblack. Sentiment Mining in WebFountain [A]. In: *Proceedings of ICDE-05, the 21st International Conference on Data Engineering*, 2005:1073-1083.

[4] M. Gamon, A. Aue, S. Corston-Oliver, E. Ringger. Pluse: Mining Customer Opinions from Free Text [A]. In: *Proceedings of IDA-05, The 6th International Symposium on Intelligent Data Analysis*. Lecture Notes in Computer Science, Springer-Verlag. Madrid, Spain. 2005.

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

[6] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives [A]. In: *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997:174–181.

[7] Wiebe. Learning Subjective Adjectives from Corpora [A]. In: *Proceedings of AAAI-00, 17th Conference of the American Association for Artificial Intelligence*, 2000:735-740.

[8] YE Qiang(叶强), LI Yijun(李一军), ZHANG Yiwen(张奕文). Semantic-Oriented Reviews: Sentiment Classification for Chinese Product an Experimental Study of Book and Cell Phone Reviews [J]. *Tsinghua Science and Technology*. Vol.10(1), 2005:797-802.

[9] Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques [A]. In: *Proceedings of ICDM 2003*, 2003:427-434.

[10] Yuen, Raymond WM. Chan, Terence YW. Lai, Tom BY. Kwong, OY. Tsou, Benjamin KY. Morpheme-based Derivation of Bipolar. Semantic Orientation of Chinese Words[A]. In: *Proceedings of Coling 2004*. 2004:1008-1014.

[11] 苏棋. 问答系统中的情感倾向性问题回答策略[D]. 北京大学博士论文, 2006.

[12] 王国璋. 汉语褒贬义词语用法词典[M]. 北京：华语教学出版社, 2001.

[13] 姚天昉,聂青阳,李建超,等. 一个用于汉语汽车评论的意见挖掘系统[A]. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议. 曹佑琦, 孙茂松 主编. 清华大学出版社, 2006:260-281.

[14] 朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, Vol.20(1), 2006:14-20.