基于PMI-IR算法的Blog情感分类研究*

段秀婷　何婷婷　宋乐

华中师范大学 计算机科学系，湖北 武汉 430079

E-mail：abc381858424@163.com; tthe@mail.ccnu.edu.cn;

摘　要：Blog信息源和信息量的广泛增长，给中文文本分类带来了新的挑战。本文提出了一种基于PMI-IR算法的四种情感分类方法来对Blog文本进行情感分类。该方法以情感词语为中心，通过搜索引擎返回的结果来计算文本中的情感要素和背景情感词之间的点互信息值，从而对文本进行情感分类。该方法在国家语言资源监测与研究中心网络媒体语言分中心2008年度的Blog语料和COAE2008的语料上分别进行了测试。与传统方法相比，准确率和召回率都有了较大的提高。

关键词：中文信息处理；情感分类；互信息；PMI-IR算法

Research on Sentiment Classification of Blog Based on PMI-IR

Xiuting Duan, Tingting He, Le Song

Department of Computer Science, Huazhong Normal University, Wuhan, 430079

E-mail：abc381858424@163.com; tthe@mail.ccnu.edu.cn;

Abstract: Development of Blog texts information on the internet has brought new challenge to Chinese text classification. Aim to solving the semantics deficiency problem in traditional methods for Chinese text classification, this paper implements a text classification method on classifying a blog as *joy*, *angry*, *sad* or *fear* using a simple unsupervised learning algorithm. The classification of a blog text is predicted by the max semantic orientation (SO) of the phrases in the blog text that contains adjectives or adverbs. In this paper, the SO of a phrase is calculated as the mutual information between the given phrase and the polar words. Then the SO of the given blog text is determined by the max mutual information value. A blog text is classified as *joy* if the SO of its phrases is *joy*. Two different corpora are adopted to test our method, one is the Blog corpus collected by Monitor and Research Center for National Language Resource Network Multimedia Sub-branch Center, and the other is Chinese dataset provided by COAE2008 task. Based on the two datasets, the method respectively achieves a high improvement compared to the traditional methods.

Keywords: Chinese information processing, semantic classification; mutual information; PMI-IR algorithm

1 引言

随着Web应用的不断发展，越来越多的人们通过博客、空间等网络形式来进行个人情感的表达和生活需求的交流。于是如何快速地、自动地从海量信息中对Blog文本所表达的情感等主观内容进行分类就变得十分重要。如果采用传统的文本分类方法进行分类，则会忽略文本中包含的情感语义信息，造成语义缺失的现象。这对专门针对情感内容进行分类的Blog文本来说，显然是远远不够的。本文引入了情感计算技术对网络信息进行有效的分析和挖掘，识别出其中的情感要素并对包含了这些信息的文本集合做出情感分类，有效的解决了这个问题。这也是当前情感计算领域的一个重要研究课题。

情感分类基本上包括三个方面内容[1]：主观性(Subjectivity)判断，即如何区分文章中的事

*基金项目：国家自然科学基金重大研究计划(90920005)；国家自然科学基金(60773167)；国家十一五科技支撑计划课题(2006BAK11B03)；973 国家重点基础研究发展计划(2007CB310804)；教育部/国家外国专家局高等学校学科创新引智计划(B07042)；湖北省自然科学基金计划项目资助(2009CDB145)；武汉市晨光计划项目资助(201050231067)。

实和观点,对文章、句子、以及词语的主观性和客观性进行判断,区分哪些是主观的,哪些是客观的;倾向性(Orientation, Polarity)判断,也称语义倾向性判断,主要的任务是判定文本中的某些内容(整篇文章、句子或词语)是肯定的,否定的,还是中立的;等级强度(Gradability)判断是对其主观性和倾向性的强度进行计算,判定肯定或否定的等级强度。本文主要介绍了我们在情感分类第二个方面和第三个方面的工作。

我们的工作主要侧重于对整篇文本的情感的倾向性做出分析。我们的基本思想之一是Harris提出的分布式假设,即一个词语如果和一些人工标注的属于肯定倾向的“种子”词共同出现的频率越高,那么这一词语属于肯定的语义倾向性则越高。接下来,我们选取了符合这个基本思想的PMI-IR(PMI, Point-wise Mutual Information, 点式互信息; IR, Information Retrieval, 信息检索)算法进行实验,并提出了一种基于该算法的情感分类方法。该方法以情感词语为中心,通过搜索引擎返回的结果来计算文本中的情感要素和背景情感词之间的点互信息值,从而判定出该篇文本所属的情感倾向类别。利用PMI-IR算法,我们既可以从Blog文本中的情感语义信息出发进行文本分类,又可以省去建立大型背景语义知识库或语义模型的繁重工作,同时保证了该实验结果的有效性;而PMI-IR算法也具有其自身的优越性,它很好的解决了传统PMI算法中难以确定词语之间语义距离的问题,为我们的研究和实验带来了很大的便利。我们采用了华中师范大学国家语言资源监测与研究中心网络媒体语言分中心2008年度的Blog语料进行了测试,并在第一届中文倾向性分析评测研讨会(COAE2008)的语料上展开了应用,提出了建立“用户——服务”情感分类系统的设想。与传统方法相比,该方法比传统的文本分类方法更具现实意义。

本文余下的内容做如下安排:第二部分:介绍在这个领域目前的相关工作;第三部分:系统描述该方法,介绍其处理过程;第四部分:展示实验结果,并对结果进行分析;第五部分:总结,简述下一步工作。

2 相关研究

PMI-IR算法是由Peter D. Turney提出的,最初被用于英文的同义词识别,即测定词对之间的相似性[2]。它综合了点式互信息和信息检索两项技术来测定两个词语之间的相似性,即通过某一词语和一个参考肯定词 *pword*(此处为“excellent”)的相似性,以及此词语与一个参考否定词 *nword*(此处为“poor”)的相似性,来测定词对之间相似性的大小。

接下来,Peter D. Turney将PMI-IR算法用于测定词汇的褒贬倾向性[3],进而判定一篇评论中的评论对象是被推荐的还是不被推荐的。该算法在Epinions上的410篇评论上进行了测试,获得了74%的平均精度。而在特定领域中,实验精度分别是:汽车评论84%,电影评论66%。

进一步地,香港大学语言信息科学研究中心提出了基于 Turney 改进的一种词汇语义褒贬倾向性判定方法[4]。他们将参考肯定词和参考否定词扩充为参考肯定词词集(*pwords*)和参考否定词词集(*nwords*)的形式,而词语的语义倾向性计算公式也做了相应的改进,扩展为统计的形式。该实验采用了中文繁体语料,测试集为包含了34,000,000个词的语料中的1249个词语(604个肯定词和645个否定词)。实验精度提高到81.5%,但召回率降低为45.56%。

另外,国内还有不少其他的情感分类方法。比如根据文本中词语具有褒贬倾向性概率的大小,建立相应的最大熵特征模型来识别文本中所涵盖的情感词,再用SVM方法对BBS 文本进行情感分类[5]; 另外,还有通过已有的General Inquirer(GI)词典、《学生褒贬义词典》、知网、《褒义词词典》、《贬义词词典》五种资源,构建出中文情感词词表,并采用加权线性组合的句子情感分类方法对句子进行情感类别判断[6]。但以上方法都只是判断了分类文本是属于肯定的、否定的还是中立的。显然,这与人们表述观点、抒发情感的复杂性和多样性相比,是远远不够的。

本文结合第二届中文倾向性分析评测研讨会(COAE2009)的任务1,提出了一种基于PMI-IR算法的情感分类方法,对人们表达最为丰富的喜、怒、哀、惧四种情感进行了情感分类的研究,并通

过进一步的实验改进来提高实验精度。经实验证明,该方法是可行的。

3 基于 PMI-IR 算法的情感分类

3.1 情感分类算法

互信息(Mutual Information)是信息论中一种广泛使用的信息度量。它主要运用概率论与数理统计的方法进行研究,用来衡量变量之间的依赖程度。互信息是指两个事件集合之间的相关性。比如,两个事件 X 和 Y 的互信息可定义为:

$$I(X,Y)=H(X)\cdot H(Y)-H(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)\geq 0 \quad (1)$$

其中 H(X|Y)是条件熵(conditional entropy),H(X,Y)是联合熵(Joint Entropy)。H(X,Y)可定义为:

$$H(X,Y)=\sum p(x,y)\log\ p(x,y) \quad (2)$$

我们根据互信息的理论思想,采取如下步骤进行实验:

第一步,统计出能够代表待分类 Blog 文本情感特征的最大相关属性集合,即情感要素词集;

第二步,将其中的每一个情感词都看作一个可统计互信息的点,然后通过点式互信息的计算公式计算出各个情感词和参考词之间的互信息值。点式互信息的计算公式如下所示:

$$PMI\ (word_1, word_2)=\log_2\left(\frac{p(word_1\ \&\ word_2)}{p(word_1)p(word_2)}\right) \quad (3)$$

第三步,统计所有的互信息值,得到该篇文本分别与四个目标情感类的最大依赖值,将其中最大的所在的目标情感类作为该篇文本的情感分类。

公式中,对数内分子 $p(word_1\&word_2)$,即待分类文本的各个情感词与参考词的共现概率如何计算?这正是计算互信息值的关键所在。

PMI-IR 算法凭借自身的优越性很好的解决了这一问题。该算法利用了点式互信息和信息检索来测定两个词语的相似性,不需要考虑不同的情感要素词之间的共现距离,而是通过搜索引擎返回的结果来计算某一情感要素词和一个参考肯定词(*pword*)的相似性,以及该词语与一个参考否定词的相似性(*nword*),最后计算二者相差来测定这一词语的语义倾向性(Semantic Orientation, SO)。算法的计算公式如下所示:

$$SO(phrase)=PMI(phrase, pword)-PMI(phrase, nword) \quad (4)$$

同样的,我们把参考词 *pword* 和 *nword* 扩充成为参考词集 *pwords* 和 *nwords*,计算公式如下所示:

$$SO-A(word)=\sum_{pword\in Pwords}A(word, pword)-\sum_{nword\in Nwords}A(word, nword) \quad (5)$$

通过参考情感词集的扩充,我们可以更加全面的把握待分类文本的情感信息,并对文本的情感分类做出更加准确的定位。这对提高实验的准确率和召回率都是大有帮助的。

3.2 实验过程

对 Blog 文本中情感内容的识别可以从“作者”和“读者”两个角度来理解[7]。例如,Blog 文本中的正面情感可能会诱发读者的负面情感响应。通常情况下,作者会运用各种表达手法和表达方式来激发读者产生各种各样的情感状态。因此,为了避免混淆,本文把 Blog 文本中的情感内容统一定义为作者所表达的情感内容和所呈现出情感状态的片断。这样,一篇 Blog 文本的情感内容就是由这些情感片段组成的集合。

另外,我们根据人的主观情感体验,把“喜”、“怒”、“哀”、“惧”这四种情感的情感表达倾向性和强度定义为:“喜”为正面情感,强度较强;“怒”、“哀”、“惧”为负向情感,“怒”

的强度较强，“哀”、“惧”较弱。我们的实验过程如下：

第一步，下载华中师范大学国家语言资源监测与研究中心网络媒体语言分中心 2008 年度的 Blog 语料作为实验语料，并人工标注主要表达了“喜”、“怒”、“哀”、“惧”情感的语料文本各 100 篇共 400 篇作为测试集，其中每一种情感代表一个目标类。

第二步，对测试集中的文本进行带词性标注的分词(例如，“/a”表示形容词，“/n”表示名词)。分词工具为 EasyNLP2009 工具集中的分词工具。

第三步，抽取测试集文本中的情感要素词作为该篇文本的情感特征集。由于人们的情感表达一般是通过形容词和副词来体现的，所以我们的情感要素词主要包括标注了“/a”(形容词)、“/ad”(副词)以及“/ag”和“/an”(形容词性要素)的词语。而单音节性质形容词描述事物的属性是以事物实体为基础、媒介形成概念意义的[8]，所以我们对其不作考虑。

第四步，参考人工标注情感语料库 Ren-CECps 1.0 中带有特定情感倾向性(Joy, Angry, Sorrow, Fear)且人工标注为 1.0 的词语，以及 COAE2009 比赛中所用的情感“喜”、“怒”、“哀”、“惧”种子词集，我们对应每类情感分别人工抽取了 10 个带有对应情感且极性很强的词语作为参考词集，如下所示：

jword = {愉快，高兴，快乐，喜极，自豪，开心，喜悦，得意，幸福，兴奋}；

aword = {不悦，咒骂，恼火，愤怒，抨击，抱怨，气冲冲，气愤，谴责，愤恨}；

sword = {伤心，哀伤，哀愁，哭泣，心痛，悲伤，悲剧，难过，噩耗，委屈}；

fword = {不安，受惊，唯恐，余悸，害怕，恐怖，战战兢兢，胆战心惊，骇然，畏惧}。

最后，根据词语的语义倾向性计算公式计算出测试文本对应每种情感类别的倾向性值 SO_j、SO_a、SO_s、SO_f，而这些值都代表了这篇文本与各个类之间的依赖程度。选取其中的最大值，其对应的情感类别即判定为该篇文本的情感类别。这样我们就得到了这篇测试文本所属的类别。

4 实验结果与分析

4.1 实验结果

对 Blog 文本进行分类可以看作是一种机器学习的过程，机器学习中常用的评估标准有准确率(查全率)、召回率(查准率) 与 F1 测度值。本文通过这三个指标来衡量情感分类的质量。

准确率(查全率)=事实属于此类且被分类正确的文档数/属于此类的总文档数

召回率(查准率)=事实属于此类且被分类正确的文档数/被判为此类的文档数

F1 测度值=2 ×查全率×查准率/ (查全率+ 查准率)

经统计，实验结果如下表所示：

表 1

	准确率	召回率	F1 值
喜	0.677	0.511	0.583
怒	0.426	0.370	0.396
哀	0.33	0.317	0.324
惧	0.5	0.446	0.471

从实验结果来看，情感“喜”的准确率、召回率和 F1 值都相对较高，这与情感“喜”本身的正面极性很强是密切相关的。而情感“怒”因其情感表达比较激烈，也得到了相对较好的结果。但由于情感“怒”、“哀”、“惧”都属于负面情感，且彼此之间容易造成情感转移，作者在进行 Blog 写作时很容易将三者混杂在一起进行情感的表达。比如，我们的实验语料中一篇人工标注为“哀”的语料 blog_cps3，在情感表达上情感“哀”是主要的，但又带有情感“惧”的色彩。因此使用机器来判定这篇语料的情感分类时，受到情感“惧”的干扰较大，而受到情感“怒”的干扰较小。同时，从对应的实验结果来看，这篇语料情感“哀”的情感倾向性值 SO_s=28.509，而另外两种

的情感倾向性值为 SO_s=28.185 和 SO_f=28.836。通过 SO_s 和 SO_a 的比较，我们可以得到正确的分类，即该篇语料被机器判定正确为“哀”类；而加入了情感“惧”之后，机器会将这篇语料误判为“惧”类。类似的，实验中其他几种情感的分类也遇到了这样的情况。

总的来说，实验过程中产生的错误的中间结果，会对实验的最终结果造成巨大的影响。这在针对四种情感的多分类问题上显得尤为明显。所以我们将对这四种情感分类的测试集做出更改，以提高实验的精度。

4.2 实验的改进

考虑到四种情感同时进行分类时，各项评测指标测定的结果都不够理想，我们将对实验的方法进行一些改进。考虑到情感“喜”是其中唯一的正向情感，我们对测试集的范围作如下修改：以情感“喜”为正向参考词集，依次取出另外三种情感的参考词集作为负向参考词集，然后对这两种情感的文本集合进行分类，即分别对“喜”和“怒”，“喜”和“哀”和“喜”和“惧”进行分类。同样，我们测试集文本中词语的语义倾向性计算公式就减少到两两进行比较。

经过上述改进实验后得出结果如下图所示：

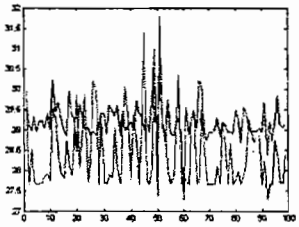

图1 “喜-怒”情感对的实验结果

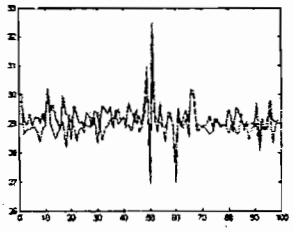

图2 “喜-哀”情感对的实验结果

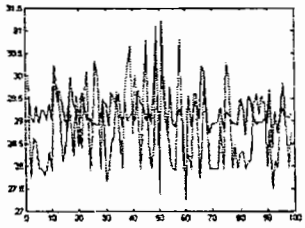

图3 “喜-惧”情感对的实验结果

上述实验仍然采用华中师范大学国家语言资源监测与研究中心网络媒体语言分中心 2008 年度的 Blog 语料作为实验语料。图 1、图 2 和图 3 分别为情感“喜”和“怒”、“喜”和“哀”、以及“喜”和“惧”的各 100 篇语料测试结果曲线图。经统计，可得出以上各个情感对分类的准确率、召回率和 F1 值，如下表所示：

表2

		准确率	召回率	F1 值
喜	怒	0.808	0.816	0.812
	哀	0.758	0.926	0.833
	惧	0.657	0.619	0.637

可以看出，各个情感对的分类结果都得到了较大的改善，准确率、召回率和 F1 值都提高了不少。但“喜”和“哀”这两种情感相对来说分类效果较差。与 Turney 的实验相比，我们的实验在精度和召回率上都得到了较大的提高；与香港大学在中文繁体语料上的实验相比，我们的实验在精度上降低了 0.7%，但在召回率上提高了 16.44%-47.04%。相应的，实验结果也从数值的角度证明了我们对情感等级强度的定义是正确的，这对我们建立情感空间时选定情感极性坐标的指向和测度也是一大启示。

接下来，我们对上述改进的实验方法做了进一步的推广，选取了第一届中文倾向性分析评测研讨会(COAE2008)比赛的语料进行实验。该语料集共有 40000 篇语料，可分为主观性文本和客观性文. 我们根据该比赛给出的主客观分类结果，对这些语料中所包含的 3091 篇主观性文本进行了测试。为避免引入不必要的噪声，我们首先对这些主观性文本进行了筛选，筛选标准为：(1) 专业词汇过多的；(2) 人名和地名过多的；(3) 仅描述电影剧情的。符合这三条标准的文本都不在我们的考

虑范围之内。

经过筛选后我们得到共215篇测试语料，然后用改进后的实验方法对其进行了测试。测试结果如下图所示：

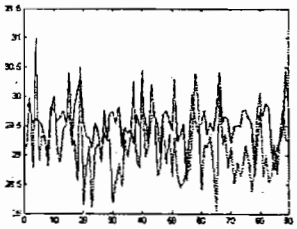

图4 “喜-怒”情感对的实验结果

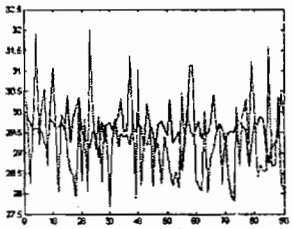

图5 “喜-惧”情感对的实验结果

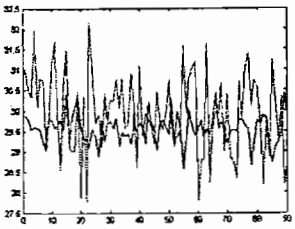

图6 “喜-哀”情感对的实验结果

经统计，可得出以上各个情感对分类的准确率、召回率和F1值如下表所示：

表3

		准确率	召回率	F1值
喜	怒	0.6	0.730	0.659
	哀	0.82	0.3	0.24
	惧	0.667	0.659	0.663

从测试的结果来看，多数情感对的情感分类达到了较为理想的效果，其中情感“喜”和“哀”和分类效果准确率最高，但召回率和F1值相对较差；而另外两个情感对的分类效果准确率相对较差，但召回率和F1值较为理想。由此可见，选取极性非常强的词作为情感参考词集是非常必要和有效的。另外，对于不同类型的文本，作者在情感的表达上有强有弱，所表达的情感内容和表达方式之间的映射关系的也具有多样性。这些对所描写的内容涉及到不同领域的文本来说，在情感分类上都有一定的影响。

5 结论及展望

本文提出了一种基于PMI-IR算法的四种情感（“喜”、“怒”、“哀”、“惧”）分类方法来对网络中的Blog文本进行情感分类。实验表明，该算法在中文简体语料的词汇语义倾向性判定上是有效的，并且在进一步的Blog文本的情感分类上达到了良好的效果。与传统方法想比，实验的准确率和召回率都得到了较大的提高。

在下一步的工作中，我们将在这个方向展开进一步的研究，包括如何选取更为丰富、有效的情感要素词集，降低文本情感特征的维度，以及情感分类系统的建立。

参 考 文 献

[1] 张智雄，吴振新，赵琦，洪娜徐健，刘建华. 非结构化文本中内容对象抽取的技术方法综述. 数字图书馆论坛，2008，第9期:2

[2] Peter D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning, 2001:2

[3] Peter D. Turney. Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002:2

[4] Raymond W.M.Yuen, Tom B.Y.Lai, O.Y.Kwrong, Benjamin K.Y.Tsou. Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words. Proceedings of Language Information Sciences Research Centre, the City of Hong Kong:2

[5] 陈锦禾, 范新, 沈闻, 沈洁. 基于情感词识别的 BBS 情感分类研究. 计算机技术与发展, 2009, 7:2
[6] 王素格, 杨安娜, 李德玉. 基于汉语情感词表的句子情感倾向分类研究. 计算机工程与应用, 2009, 24:2
[7] 孙凯. 面向观众的电影情感内容表示与识别方法研究. 华中科技大学: 计算机科学与技术学院, 2009:3
[8] 张伯江. 性质形容词的范围和层次. 第十四次现代汉语语法学术讨论会, 2006 年:4