中文博客主题情感句自动抽取研究

孙宏纲,陆余良
SUN Hong-gang,LU Yu-liang

合肥电子工程学院 604 实验室,合肥 230037
No.604 Lab,Hefei Electronic Engineering Institute,Hefei 230037,China
E-mail:navysun1977@163.com

SUN Hong-gang,LU Yu-liang.Study of topic sentiment sentences auto-extraction in Chinese blogs.Computer Engineering and Applications,2008,44(20):165-168.

Abstract: In the field of Chinese blog sentiment analysis,previous researchers put most energy on the polarity analysis of word,but not all the word analyzed is relative with the topic,and word-level granularity for sentiment analysis is too small.We try to use sentiment sentences,a sentence-level model,for sentiment analysis.In this paper,it only focuses on topic sentiment sentences auto-extraction.In order to extracting topic sentiment sentences,it designs a novel Bi-segment method to extract the main topic words,and uses TFIDF to extract more topic words.With these words,it recombines original sentences,which contain the topic words.So as long as topic sentiment sentences exist,they must in the set of recombined sentences.Then,based on the analysis of Chinese blogs,it converts the problem of extraction into Chinese chunking by CRFs and has a good performance in extraction experiment.

Key words: Chinese blogs;sentiment analysis;Conditional Random Fields(CRFs)

摘 要:博客作为一种大众化的信息及文化载体被越来越多的人所接受,博客信息的情感分析也逐渐成为了信息挖掘领域的热点。目前,在研究情感分析时,多是通过计算词汇的倾向性来完成的。由于并不是所有的带有情感色彩的词汇都是主题相关的,因此,以词为粒度的情感分析存在一定的缺陷。为了解决这一问题,试图从句子层面进行分析,主要研究了与之相关的主题情感句的自动提取问题。为了有效地提取主题相关情感句,设计了一个新颖的基于二元切分的提取算法来获取主题词,然后利用 TFIDF 算法获取更多的次要主题词,并利用这些主题词重组了那些包含主题词的原始句。因此,如果主题情感句存在的话,那么它一定在这些重组的主题句集合中,只要对该重组句集合进行分析、提取,便能得到主题情感句。最后,利用 CRFs 将主题句提取问题有效转化为了中文 chunking 问题,并在抽取实验中取得了很好的结果。

关键词:中文博客;情感分析;CRFs
DOI:10.3778/j.issn.1002-8331.2008.20.050 **文章编号:**1002-8331(2008)20-0165-04 **文献标识码:**A **中图分类号:**TP391

1 引言

自从 1997 年 Jorn Barger 第一次使用“Weblog”以来,博客(blog)借助互联网的日益普及以及自身的优势而迅速发展壮大。由于博客使用简单、个性化强、实时性好,越来越多的民众开始创建、使用自己的博客。截至 2005 年全球博客用户已经突破 1 亿用户,中国用户超过 1 600 万,这标志着博客正式从精英走向了大众,开启了互联网发展到真正个人化时代的帷幕,而互联网也从商业化进入社会化阶段。由这种互联网应用所创造的动态的网络空间也逐渐成为了一种崭新的草根文化载体[1]。据统计,在博客内容方面,37%的博客用户会写一些与自身生活和经历相关的信息;11%的博客作者倾向于对公共热点问题发表观点[2]。由于博客空间的信息丰富、更新速度快且源于广大民众等特点,开始受到许多政府部门和社会团体越来越多的关注。其中,如何获得博客中广大民众的情感倾向便是一个值得关注的问题。

2004 年 AAAI(Association for the Advancement of Artificial Intelligence)组织了一次相关的研讨会 EAAT(Exploring Attitude and Affect in Text)。会议主要目的就是探讨文本中情感分类、文本主观性、态度等相关问题[4]。TREC 2006 中也加入了一项新的测评内容 blog-track,新内容包括两部分,其中重要的一部分就是博客观点搜索(opinion retrieval),也就是搜寻对某一特定主题阐述观点的博客搜索准确度很多程度上依赖于博客内容情感分析的准确度。

文本情感分析所采用的方法主要有两种。第一种方法,首先将具有情感色彩的词分成是正例和负例,然后以词频统计基础,建立一个二元的分类器,从而进行简单的情感分类;另一种

作者简介:孙宏纲(1977-),男,博士生,主要研究领域为计算机应用,Web 信息挖掘;陆余良(1964-),男,教授,博士生导师,主要研究领域为计算机应用,信息挖掘。

收稿日期:2007-09-26 **修回日期:**2007-12-21

方法是相关词的语义倾向分析,然后计算整个文本的情感倾向指标[5],这也是目前较为有效的一种方法。

这些方法多是单纯从词的倾向性分析入手[6-8],但是以词为粒度的情感分析,不能保证所有被分析的词汇都是主题相关的,这种不确定性将直接影响文本情感分析准确性。解决词汇主题相关不确定性的一个很好的途径就是首先提取主题情感句(Topic Sentiment Sentences,TSS),以句子为分析粒度进行情感分析。本文主要研究了在中文博客中,如何有效提取主题情感句。为了提取准确获得相关的主题情感句,针对博客的形式特点,设计了一个新颖的,基于二元切分的关键主题词提取方法,同时采用 TFIDF 算法获得次要主题词。在此基础上对包含主题词的原始句进行重组和简化,并且认为,只要主题相关情感句存在,那么它一定在这些有主题词构成的主题句集合中。最后,结合汉语语法知识对中文博客的结构和内容进行了分析,将主题情感句提取问题转化为中文 chunking 问题,并通过 CRFs(Conditional Random Fields)算法进行了实验,取得了很好的效果。

2 中文博客主题词提取

2.1 相关分析

在创作博客时,博客内容通常是作者的自我表达,通过博客这个平台,博客作者们可以自由地、广泛地抒发自己的感情,讨论热点事件等等。因此,可以将博客视为一个私人的新闻中心,每天博客作者都会在上面发布新的内容。除了更新速度,博客在内容形式上也接近新闻。博客的内容往往都很短小精炼,同时为了吸引更多的读者,博客的标题也像新闻标题一样是内容的精炼。因此,总可以在标题中找到与博客内容相关的主题词。

在现代汉语中,40%的词汇是单字词汇,但大多数在单独出现时都没有实际的意义;60%的词汇是双字词,这一部分占据了汉语词汇的主体[9]。通过对博客内容的统计,得到了一些有用的特征:每一篇博客通常只有一个主题;标题中的主题词通常只有两个;主题词在博客内容中至少出现一次。

从语言学角度,可以得到:如果在标题和内容中连续词串 A 和 B,且满足 A 包含 B,那么词串 A 的特指性要高于词串 B。

2.2 基于二元切分的主题词提取

目前主题词的提取往往都依赖于字典和大规模的词频统计。对于中文来讲,主题词的提取尤为困难。因为不同于西方语言,汉语在形式上是一个连续的词串,词与词之间没有天然的断开。如果利用专业字典来进行中文博客的主题词提取,那么就需要一个非常完善的字典,否则就可能有一些主题词由于是未登录词,而被忽略掉;另一方面,单纯利用词频统计,也会丢掉一些低频的主题词。

为了简单有效地获得中文博客中的关键主题词,设计了一个基于二元切分的主题词提取方法,也就是将博客中的汉语词串全部划分成二元词汇,并建立倒排索引。例如:输入词串 S="*abcde*",经过二元切分,将输出二元词串 O="*ab bc cd de*"。因此,无论查找词串 O 中的任何二元词汇,都可以很快定位到词串 S。

具体实施时,首先,把标题和内容中的每一个字按顺序编号,然后根据标点符号把标题和内容划分成子串,并将子串保存在对应的向量 $\boldsymbol{T}$ 和 $\boldsymbol{C}$ 中。然后按照下面步骤进行提取:

(1)把所有标题和内容当中的子串进行二元切分,并将二元词汇保存在相应队列 Tq 和 Cq 中;

(2)从 Tq 中逐个取出切分后的二元词汇,然后在 Cq 中检索该词。如果 Cq 中包含该二元词,则把该词加入到倒排索引 G=<二元词,(pos_i,…,pos_k,…),频率>;

(3)如果不同词汇在倒排索引 G 中的位置标记 pos 是连续的,则认为这些 pos 连续的二元词构成了一个复合词。不断重复步骤(2)和(3),直到找到所有位置标记连续的词汇;

(4)对于每一个由二元词构成的复合词,采用一组策略来判断该复合词是否是博客的主题词。

根据 2.1 节中所分析的博客特征,定义了如下策略用于判断主题词:

如果复合词串 S 是由多个复合词串构成,且在博客中出现的次数多于 1;

如果复合词串 s 包含于复合词串 S 中;

如果复合词串 s 在博客中的词频多于 X;

如果复合词串的结尾是名词;

如果复合词串的开头是名词;

……

若经过第一轮匹配,标题中没有发现显著的关键主题词,而只是出现了多个长度接近,且词长较短的次关键主题词,如果这些次关键主题词不超过两个,则视为关键主题词;如果多于两个,则首先根据词串在标题中所处的位置,进行判断:破折号引出的、处在双引号、书名号中的词串重要性高;其次,词串长的重要性高;特殊情况下,如果仍无法选出关键主题词,则将重要性接近的前 n 个次关键主题词,同时作为关键主题词。

对该方法我们进行了抽取测试,测试数据从网站 http://blog.sina.com.cn/下载。测试结果表明,该方法的抽取准确度达到 92%。

3 主题句重组

3.1 博客信息分析

任何一篇文档可以用一个由关键词构成的向量 $\boldsymbol{D}$=<$w_1 Word_1$,……,$w_n Word_n$>表示[10]。理论上,不同文档的文档向量是不同的,这一特点对于句子重组非常有用。

假设一个博客标题和博客内容是一一对应的,则可以在标题和内容之间建立映射关系 B=<*title*,*Doc*>。在 2.1 节中曾讨论过,博客在形式上具有新闻的特性,因此,可以把标题概括为相关的主题词,同时在建立映射 T=<*title*,*topic words*>。利用映射 B 和 T 可以建立主题词和文档之间的映射 TD=<*topic words*,*Doc*>,并且计算主题词和文档向量的相关性。因此可以得到结论,博客内容和主题词是一一对应的。

实验发现,文档向量中参与相似度计算的有效词主要包括名词、动词、形容词和副词。由这些关键元素组成的子向量可以很好地表示博客内容。根据这些词汇在博客中的位置信息,可以将它们还原成"句子",当然这些"句子"已不是原来完整的句子,它只包含子向量中的关键元素。

通过以上分析,可以认为由子向量中关键元素重组的句子都是主题相关的,因此,如果博客中存在主题相关情感句,那么它们一定存在于这些重组的主题相关句中。

3.2 情感句重组

为了有效获得主题相关情感句,需要将子向量中的关键元

素(名、动词、形容词和副词)进行重组,获得主题相关句。子向量中的每一个元素都有相同的结构<$word$,$weight$,$Sentence_i$,$Position_{ij}$>,其中$Sentence_i$和$Position_{ij}$表示某一关键词的具体位置是句子i,绝对位置j。通过这种索引结构,很容易就将主题句重组。接下来就可以计算句子和主题之间的相关性了。句子与主题的相关度定义如下:

$$Cor(T,S_j)=\frac{\sum_i \frac{\sum W_i}{M_i}}{I}=\sum_i \frac{\sum W_i}{M_i*I}=\sum_i \sum \frac{W_i}{M_i*I}, i\in I(a,ab,v,n)$$

其中I表示同一句子中不同词性标记的数量,$\sum W_i$表示具有相同词性标记i的所有单词的权重和,M_i表示具有相同词性标记i的单词的数量。在计算句子的主题相关度时,主要利用了词的权重,如果一个句子都是由高权重的词组成,那么该句子的主题相关性就高,反之则低。按相关度$Cor(T,S_j)$的大小对句子进行排序,这样就可以根据需要获得不同相关度的主题句。

4 基于 CRFs 的主题情感句提取

4.1 条件随机场(Conditional Random Fields,CRFs)

CRFs(Conditional Random Fields)是一种用于词性标记、命名实体识别的较为有效的概率模型[11]。它在形式上类似于 HMM 模型,不仅具有 MEMM 模型的优点,同时有效解决了 label bias 问题。CRFs 可以看作是一个无向图模型,设$G=(V,E)$是一个无向图,其中V是无向图的顶点,E是无向图的边。X是一组被观察的随机变量,Y是一组由V确定的需要预测的输出变量。当以X为输入,且Y遵循马尔科夫性质,则(X,Y)是由X决定的无向图G,称为条件随机场(CRFs)。

将 CRFs 模型用于自然语言理解时,根据语言的特性可以简化为一个线形模型,线性模型是 CRFs 的一个很重要的应用。假设$X=(X_1,X_2,\cdots,X_n)$是自然语言的一个随机的观察序列,$Y=(Y_1,Y_2,\cdots,Y_n)$是需要标记的状态序列,那么X条件下Y的概率是[11]

$$P(Y|X)=\frac{\phi(Y_i,X_i)}{Z(X)} \quad (1)$$

其中$\phi(Y_i,X_i)$是势函数,$Z(X)$是归一化因子:

$$Z(X)=\sum_Y \phi(Y_i,X_i) \quad (2)$$

通常势函数$\phi(Y_i,X_i)$由形如$f_k(Y_i,X_i)$的二选特征构成:

$$\phi(Y_i,X_i)=\exp(\sum_i^T \sum_k \gamma_k f_k(Y_i,X_i)) \quad (3)$$

γ_k是权重系数。

如果从条件分布定义的角度理解线性 CRFs,首先考察给定X、Y时,它们的联合分布:

$$P(Y,X)=\phi(Y,X)=\exp(\sum_i^T \sum_k \gamma_k f_k(Y_i,X_i)) \quad (4)$$

根据条件概率的定义,给定X情况下,Y的条件概率,可以表示为:

$$P(Y|X)=\frac{p(Y,X)}{\sum_Y p(Y,X)}=\frac{\exp(\sum_i^T \sum_k \gamma_k f_k(Y_i,X_i))}{\sum_Y \exp(\sum_i^T \sum_k \gamma_k f_k(Y_i,X_i))} \quad (5)$$

其中,分母$\sum_Y p(Y,X)$是X的边缘分布。

根据势函数定义,可以将$\phi(Y_i,X_i)$分解为

$$\phi(Y_i,X_i)=\exp(\sum_k \lambda_k t_k(Y_{i-1},Y_i,X,i))+\sum_j \mu_j s_j(Y_i,X,i) \quad (6)$$

其中$t_k(Y_{i-1},Y_i,X,i)$是在输入X情况下,标记状态在位置i和i-1时的二选特征值;$s_j(Y_i,X,i)$是标记状态在位置i时,相对于X的二选特征值。λ_k和μ_j分别是它们的权重

每一个t和s都是一个二选的特征函数,s可以视为是状态特征。

$$s(X,i)=\begin{cases}1 & \text{如果在 } X \text{ 中,位置 } i \text{ 处是一个名词}\\ 0 & \text{其它情况}\end{cases}$$

而t则是传递特征,所有的状态和传递特征函数都具有类似的定义形式。

$$t(Y_{i-1},Y_i,X,i)=\begin{cases}1 & \text{如果 } Y_{i-1} \text{ 和 } Y_i \text{ 存在某种关系}\\ 0 & \text{其它情况}\end{cases}$$

4.2 参数估计

给定训练集$T=<O_i,S_i>,0\leqslant i\leqslant n$,参数估计就是寻找适当的参数向量$\boldsymbol{\lambda}$和$\boldsymbol{\mu}$使得对数似然(7)取得最大值。

$$L_\Lambda=\sum_{i=1}^n \log P_\Lambda(S_i|O_i)-\sum_{k=1}^K \frac{\gamma_k^2}{2\sigma^2}=\sum_{i=1}^n (\sum_k \gamma_k f_k(S_i,O_i)-\log Z(O_i))-\sum_{k=1}^K \frac{\gamma_k^2}{2\sigma^2} \quad (7)$$

其中,等式右边的第二项是均值为 0,协方差为σ^2牛顿先验值,$Z(O_i)$和式(2)具有相同的形式。

CRFs 具有 MaxEnt 模型的所有特性[11],L_Λ在定义域内是一个凸函数,这就保证了局部最优点便是全局最优点。在此使用了拟牛顿算法和 BFGS 修正对L_Λ进行最优化。同牛顿法相比,拟牛顿法利用函数的一阶信息建立一个近似的 Hessian 矩阵,从而大大提高了拟牛顿法的优化效率。

4.3 情感句分析

目前的文本情感分析研究主要利用 Wordnet、Hownet[7,8]这样的语义词典进行词的极性分析。但是以词为粒度的分析,并不能保证被分析的词汇确实是主题相关情感词。因此,在情感分析时不能单纯依赖词汇的极性分析,这一点在中文情感分析中存在跟多的问题。在汉语中相同的词语在不同的句子形式中表示不同的含义;相同的句子但是标点符合不同表达的意思也截然不同。

在现代汉语中,基本上包含 8 种简单句式和 10 种复合句式。虽然几乎所有的句式都可以用来表达情感,但是在表达习惯上,却往往只用其中的几种句式。经过分析发现,大概只有一半数量的句式被经常用来表达情感。在简单句中形容词词组、动词词组、一些成语都是表达感情的重要元素;在复句中除了上述词语以外,某些连词也是必不可少的。

总之,对于有效的情感表达来说词汇,尤其是形容词、动词以及某些具体的句型都是不可缺少的。因此,要对博客进行情感分析,提取出博客中包含有上述元素组成的主题情感句是必要的。通过第 3 章的句子重组,可以准确地获得博客地主题相关句,只要从这些主题相关句中找到主题相关情感句,便可进行博客的情感分析。

4.4 基于 CRFs 的主题情感句提取

之前许多学者利用 CRFs 和其它方法对汉语的浅层分析进行了研究[12-14],这里把主题情感句的提取也看作是一个浅层分析的过程。具体来说它是一个中文 Chunking 的过程,由于划分的对象是重组的主题句,而不是原始的句子,因此,并不是涉及所有的组块划分,重点放在动词、形容词、副词以及连词组块的划分。根据 Abney 在它的文章中对英文 Chunk 的定义[15],结合主题情感句提取的特点,定义了 4 个不同的中文情感词组块,每一个组块都采用了 IOB2[14]的标记形式。表 1 列出了情感词组块。

表 1 情感词组块

Chunks	解 释	例 子
VP	动词词组	成功,支持,反对
ADJP	形容词词组	正义,美丽
ADVP	副词词组	非常,特别,值得
CP	连词	虽然,但是

在应用CRFs 进行主题情感句提取时,就是最优化似然对数 L_Λ,找到它的解向量 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\mu}$ 使得似然对数最大。L_Λ 中的二选特征函数 $f_k(Y_i, X_i)$ 通过定义重组主题句中情感词组块之间的关系得到。

5 实验

5.1 实验设计

实验所用的博客集合是从网站 http://blog.sina.com.cn/中下载的。由于不是所有的博客都包含主题情感句,因此在训练时,从博客中选择了一些典型的主题情感句作为训练集,然后用整个博客集合作为测试集合。在整个博客集合中,包括两个子集,一个子集由 500 个包含主题情感句的博客组成,称之为正例集合;另一个子集由 500 个普通的博客组成,称之为负例集合。

在训练集合中之所以只包含句子,主要因为在一篇博客中只包含少量的主题情感句,甚至不包括,而大量的其它信息对于训练是没有帮助的。通过简化训练集合,可以节省大量的训练时间。测试集合由博客组成,是因为主题情感句提取的最终目的是要分析博客的情感倾向,因此用博客作为最终的测试单元是必要的。

5.2 性能指标

性能指标的选择方面,选择了两套不同的指标。中文 Chunking 采用了分类中常用的正确率(*precision*)和召回率(*recall*)。在评估主题情感句提取时,采用了经过调整的正确率($Precision_{adjust}$)和召回率($Recall_{adjust}$):

$$Precision_{adjust}=\frac{\sum f(n)}{N_p} \tag{8}$$

$$Recall_{adjust}=\frac{\sum f(n)}{N_r} \tag{9}$$

其中 $f(n)$ 的定义如式(10):

$$f(n)=\begin{cases}1 & \text{如果博客中包含的主题情感句多于 } n \text{ 个}\\ 0 & \text{其它情况}\end{cases} \tag{10}$$

N_p 输出的包含主题情感句的博客;N_r 是正例集合中博客的数量,N_r=500;通过调整,可以很容易判断主题情感句提取的性能。

5.3 结果分析

首先对中文 Chunking 进行了实验,涉及动词词组、形容词词组、副词词组和连词结构。表 2 列出了实验的结果,同其它系统相比[13],实验结果在数值上略高,主要是因为本实验中所分析的组块数量少且易于划分。高性能的组块划分结果将有利于后面有效的提取主题情感句。

表 2 中文 Chunking 结果

Chunks	precision	recall
VP	0.978 2	0.986 9
ADJP	0.926 6	0.913 5
ADVP	0.992 5	1.000 0
CP	0.980 9	0.951 8

图 1 是主题情感句提取的结果,横坐标是 $Precision_{adjust}$ 和 $Recall_{adjust}$ 中参数 n 的取值,纵坐标分别是 $Precision_{adjust}$ 和 $Recall_{adjust}$。

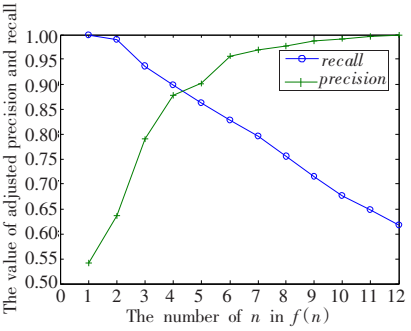

图 1 主题情感句提取结果

在式(10)中,n 是一个判断被分析博客是否包含主题情感句阈值。如图 1,当 n=3 时,如果一个博客包含等于或多于 3 个主题情感句,则认为该博客属于正例集合。从图中可以发现,当 n=1 时,虽然 $Precision_{adjust}$=0.55,但是 $Recall_{adjust}$=1,这是一个非常好的结果。由于可用于分析的博客信息相对较少,因此高的召回率,可以尽可能减少有用信息的损失,而后期的处理可以补偿低精度值带来的损失。相反,当 n=1 时,如果召回率低,而精度高,则会直接丢掉许多有用的博客信息,这种损失在后期的处理过程中是无法弥补的。

很幸运的是,当 n=4 时,召回率和精度都取得了一个不错的数值,$Recall_{adjust}$=0.88,$Precision_{adjust}$=0.9。

图 1 中的结果是严格按照实验设计得到的,只有当输出的满足阈值的结果在正例集合中才认为该博客包含主题情感句,但是在对正例和负例集合进行分类时,由于人为的疏忽,会漏掉某些正例。这一点在实验中得到了证实。在实际中这种错误也是很那避免的,对情感内容进行分类本身就是一种主观的行为,不可能得到一个标准的结果。尽管如此,实验的结果还是很令人满意的,对下一步进行有效的情感分析产生很大帮助。

6 下一步的工作

中文博客的情感分析是建立在 Web 信息挖掘、信息抽取基础上的一项非常有意义的研究。本文所研究的主要涉及主题情感句进行博客情感分析的前期准备工作,下一步的工作将进一步完善本文的研究,并在此基础上进行基于句子粒度和词汇粒度的中文博客情感分析。

(下转 221 页)