

文章编号:1001-9081(2009)04-1171-03

中文文本情感主题句分析与提取研究

樊 娜,蔡皖东,赵 煜,李慧贤

(西北工业大学 计算机学院, 西安 710072)

(fnsea@mail.nwpu.edu.cn)

摘要: 提出一种提取中文文本情感主题句子的方法。首先评估文本中语义概念的概括和归纳能力, 确定文本主题概念。将包含主题概念的句子作为候选主题句子, 计算各个候选句子的重要性, 最终确定文本主题句。然后采用条件随机场模型, 选取情感倾向特征和转移词特征训练模型, 从文本主题句集合中提取情感主题句。实验证明, 以提出的方法为基础进行文本情感分析, 避免了与主题无关的句子对分析结果的影响, 有效地提高了文本情感分析的准确率。

关键词: 情感分析; 主题概念; 条件随机场

中图分类号: TP391 **文献标志码:** A

Extraction of sentiment topic sentences of Chinese texts

FAN Na, CAI Wan-dong, ZHAO Yu, LI Hui-xian

(School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China)

Abstract: This paper proposed a method of extracting sentiment topic sentences. Firstly semantic concepts of a text were evaluated in order to determine which concepts were related to the topic of a text. And the concepts related to the topic were regarded as topic concepts. Sentences including one or more topic concepts were defined as candidate sentences. Significance of every candidate sentences was calculated in order to which ones were topic sentences in the text. Conditional random field model was adopted and two kinds of feature were used in the model training, and one feature was polarity of sentiment and the other feature was transferring words. This approach excluded sentences that were not related to the topic of the text, and eliminated the influence brought by these sentences. Therefore, precision of sentiment analysis is effectively improved.

Key words: sentiment analysis; concept of subject; conditional random field

0 引言

随着互联网的飞速发展, 网上信息急剧增长。如何快速有效地利用网络信息成为人们关注的焦点。目前已相继开展了大量的网络信息挖掘的研究工作, 例如网络信息的主题分类技术等。

网上大量的文本信息不仅包含了主题信息, 而且还包含了个人主观信息。这些个人主观信息的一个显著特点是带有个人主观情感、情绪以及态度。情感是一个非常广泛的概念, 它涉及人们的观点、看法和评价, 包括人类行为相对于社会标准的评价, 产品相对于审美观的评价。相对于情感概念的特征至少包括情感、情绪或态度、情感倾向及其强度等。以往的研究仅着眼于对文本内容的分析, 而忽略了中文文本的情感因素。情感是中文文本的重要组成部分, 仅关注内容不关注情感是很难完整反映作者意图的。

1 相关工作

作为一个新的研究领域, 从 20 世纪 90 年代开始, 文本情感分析在国内外受到了普遍的关注, 并迅速发展起来。现有的情感分析技术主要包括机器学习方法和语义方法两类。基于机器学习的情感分析方法通过大量训练样本对分类模型进

行训练。首先将具有情感色彩的词分成正例和负例, 然后以词频统计信息作为基础, 通过二元分类器进行情感分类。基于语义的情感分析方法是通过分析相关词的语义倾向, 然后计算整个文本的情感倾向。例如文献[1]从一些分散的形容词和动词中学习线索, 对动词进行 N-grams 分析, 识别句子的褒贬性, 从而对文本进行分类, 文献[2]通过建立情感分类器, 进行贝叶斯分类; 文献[3]提出了利用统计信息对词语的语义倾向进行判断的方法。文献[4]采用自定义的 44 个特征进行模式匹配, 识别主客观表达。文献[5]用聚类方法将词语划分成褒义和贬义两类, 以此预测形容词的情感倾向。

由于中英文表达习惯的不同, 同时这些方法大多都是以分析词汇或者句子倾向为基础, 但是并不能保证所有的被分析的词语或句子都是与主题相关的。对与主题无关的词语或句子进行分析, 并将分析结果纳入对整个文本倾向的判断, 会直接影响文本情感分析的准确性。如果在进行文本的情感分析时, 将与主题无关的词语或者句子剔除, 仅仅只对与文本主题相关的句子或词语进行分析, 有助于提高情感分析的准确率。

基于以上的分析, 本文以句子为粒, 主要研究如何有效提取文本中的主题句子, 在此基础上提取情感主题句, 直接对情感主题句进行情感倾向判断, 获得文本的情感倾向, 排除了与主题无关的情感句对判断准确率的影响。

收稿日期:2008-10-15;修回日期:2008-12-01。

基金项目:西安电子科技大学计算机网络与信息安全教育部重点实验室开放基金资助项目(2008CNIS-07)。

作者简介:樊娜(1978-),女,陕西渭南人,博士研究生,主要研究方向:自然语言处理; 蔡皖东(1955-),男,山东威海人,教授,博士生导师,主要研究方向:网络信息安全; 赵煜(1981-),男,陕西西安人,博士研究生,主要研究方向:自然语言处理; 李慧贤(1977-),女,内蒙古乌兰浩特人,副教授,主要研究方向:信息安全。

2 主题句提取

文本的主题句是指包含文本主题概念的句子,它既是文本中心思想的重要载体,同时也是文本内容的集中体现。本文的研究首先定义主题概念评估指标,通过指标对文本的语义概念进行评估,确定一个文本的主题概念。然后将所有包含主题概念的句子作为候选主题句子,通过计算各个候选主题句子的重要性,最终确定文本的主题句,建立主题句集合。

2.1 主题概念提取

句子是构成语篇的基本单位。主题句是与文本主题相关的句子。但是并不是与文本主题相关的句子都能称之为主题句子。这是因为相对于文本中其他句子,主题句子不仅与主题相关,而且应该具有更强的概括性和归纳性。

文本的内容是通过一定的语义概念来表达的。这里提到的语义概念,是指与作者在文本中表达的中心思想相关的基本语义单元,既可以对应文中的一个词语,也可以对应文中的有语义联系的多个词语。

作为文本的主题概念,首先必须要和文本的内容有一定的联系,同时要具备一定的语义归纳能力。本文定义两个参数评估概念,综合考查语义概念的重要性、归纳能力等方面。根据参数评估的结果,确定一个语义概念是否能够成为主题概念。

定义 1 语义概念 C 在文本中出现的次数定义为概念 C 的重要度。一个概念的重要与否,与它在文本中出现的次数有直接的关系。因为如果一个概念在文本中出现的次数比较多,说明它是作者在文本中反复提到的描述对象,因而很有可能就是文本的主题或者与文本主题有较强的相关性。设文本中语义概念为 C 的词语集合为 $\{w_1, w_2, \dots, w_N\}$, C 的重要度计算为:

$$F(C) = \sum_{i=1}^n f(w_i) \quad (1)$$

其中, $f(w_i)$ 是词语 w_i 在文本中出现的频率。这个指标与传统方法中的词语频率是类似的,反映的是表达同一个语义概念的词语出现的次数。

定义 2 语义概念 C 的词语集合中各个词语在文本段落中的分布疏密程度称为语义概念的分布广度。一个语义概念在文本中的分布越广,越有可能与主题有关。因为主题概念通常是贯穿性的出现在文本中的。设文本的段落数目为 N , 概念 C 的分布度计算为:

$$S(C) = \frac{1}{n} \sum_{i=1}^n \frac{d(w_i)}{N} \quad (2)$$

其中 $d(w_i)$ 是文本中含有词语 w_i 的段落的数目。这个指标通过统计语义概念的词语集合中各个词语在文本段落中的分布情况,衡量该语义概念的概括能力。

综合考查上述参数,根据式(3)计算概念的选取度,以确定是否选取该概念作为主题概念。概念的选取度是评估一个概念成为主题概念的可能性的度量。

$$Select(C) = \alpha \log F(C) + \beta \log S(C) \quad (3)$$

其中 $F(C)$ 和 $S(C)$ 分别是概念 C 的重要度和分布广度。 α 和 β 为加权系数,作用是调整参数之间的权重。在本文的实验中,根据经验并结合实验结果调整选取 $\alpha = 1$ 和 $\beta = 0.26$ 。当 $Select(C)$ 越大,概念 C 越有可能被提取成为文章的主题概念。完成概念的选取度计算后,需要设置选取度阈值 $Threshold$,当 $Select(C)$ 大于这个阈值时,则认为概念 C 为主

题概念,并将其归入主题概念集合 T 中。在本文的实验中,设置选取度阈值 $Threshold = 1$ 。

2.2 句子重要度评估

确定文本的主题概念后,将文本中所有包含主题概念的句子都作为候选主题句子。为了从候选集合中最终确定文本的主题句子,需要对句子重要程度进行评估计算。

对所有待处理的候选主题句子 S , 将句子包含的每个词语归入到对应的主题概念上,建立起对应向量 $S(T_1, W_1; T_2, W_2; \dots; T_n, W_n)$, 其中 T_i 为句子所含的各主题概念, W_i 为 T_i 对应的频度。建立空间向量模型后,对各个句子的重要性进行计算,将和主题最相关的重要句子提取出来。句子 S 的重要度计算公式为:

$$I(S) = P_s \frac{\sum_{i=1}^n W_i}{n} \quad (4)$$

其中 P_s 为句子 S 的位置加权系数。不同位置的句子加权系数不同。根据作者通常的表达习惯,位于段首和段尾的句子与其他位置的句子相比,往往更为重要,因此这两个位置的句子加权系数比较高。本文实验中设置段首和段尾位置句子的加权系数值为 1.6, 其他位置句子的加权系数为 1。

根据式(4),计算确定所有句子的重要性,按照重要度的大小对句子进行排序。

通过上述步骤,提取出了文本中的主题句子,并将所有提取的主题句归入主题句集合 Q 中。

3 基于条件随机场模型的情感主题句提取

为了从文本的主题句集合 Q 中提取情感主题句,本文采用了条件随机场模型。在模型的训练中,根据汉语情感句子的特点,选取两类特征:情感倾向特征和转移词特征。

3.1 条件随机场模型

条件随机场(Conditional Random Field, CRF)是基于输入节点计算输出节点配置(Configuration)的条件概率模型,也是基于马尔科夫性的模型^[6]。它与最大熵模型有相同的特征指数加权形式,但训练和推理过程采用了完全的、非贪婪的搜索算法,非常有效^[7]。近年来 CRF 模型受到广泛关注,在自然语言处理、信息抽取等领域都有应用。

对于一个观察序列 y , 标签序列 x , 定义一个线性的 CRF 模型, 形式如下:

$$P_A(y|x) = \frac{1}{Z_A(x)} \exp \left[\sum_{c \in C} \sum_k \lambda_k f_k(c, y_c, x) \right] \quad (5)$$

其中: C 为时序列的簇集合, $A = \{\lambda_k\}$ 为模型的参数集合, $Z(x)$ 为对所有可能状态序列的归一化配分函数, f_k 为某个特征函数, λ_k 为特征函数的权值。

对于条件随机场模型的训练一般都是利用最大似然法则,以迭代训练的方式获得定义特征的权值,在线性链条件随机场模型中常采用前向—后向算法训练特征的权值,本文采用的优化算法是 L-BFGS 优化算法^[8]。

标准线性链条件随机场模型的解码求得概率最高的状态序列,其过程与 Viterbi 类似。基于最大似然函数的学习过程也建立在序列估计基础上。它们的计算复杂度均与特征空间的规模成正比。

3.2 情感主题句提取

目前中文的情感分析大多是基于语义词典资源,以词为粒度进行分析,但是这种分析将与主题无关的词语也纳入分

析中,影响准确性。本文研究以句子为粒度,与英文相比,汉语句式相对比较复杂。根据句子表达情感的复杂度,本文将情感句子划分为两种类型,分别是简单句情感句和复杂情感句。简单情感句包含两种句子类型:单情感词的简单句和多情感词的简单句。单情感词的简单句就是指只包含一个情感词汇的简单句。多情感词的简单句是指包含多个情感词汇,但句子结构还是比较简单,不包含复杂的转折等连词。与前面的简单情感句相比,复杂情感句是指包含多个连词,尤其是转折连词的句子,因为这样的语句通常表达的情感都有一个转折,或者是情感倾向上的转折,或者是情感主体的转折。

考虑到上述各类情感句的特点,本文在CRFs模型训练的特征选择中,主要选取两类特征进行模型训练。一类是情感倾向特征,对于观察 y 上的点特征,选择具有情感倾向的词语作为特征。表示情感倾向的词语主要有形容词、副词以及少量的动词等。

对于简单情感句,选取情感倾向特征就足够了,但是对于复杂情感句,单一的情感倾向特征是不够的。复杂情感句通常都包含情感转折,因此必须要引入第二类转移词特征来描述这种情感转移现象。转移词特征选择具有递进或者转折意义的副词和连词,例如“可是,然而,但是”等词语。本文人工对语料库进行统计,对这类词语建立了一个基本词典。根据汉语表达习惯,这些表达转折或递进意义的词语在句子中所处的位置相对比较固定,因此主要考查句子中前两个词语的位置是否有这些词语。

4 实验

本文的实验的目的是评估提取情感主题句方法的性能及其对文本情感分析准确率的影响。因此将实验设计为两个部分。一是评估本文方法的性能,二是将本文方法应用到文本情感分析中,判断文本的情感倾向,并与传统支持向量机(Support Vector Machine, SVM)分类器的情感分析性能进行比较。

4.1 语料

本文实验中采用的语料为中文手机产品评论文本。首先从手机产品使用评论网站(http://product.it168.com/newpinglun/cSpace_pl.asp?cType_code=0302)搜集整理手机评论文本,并对所有评论认真审查,去除语言不规范以及倾向不明显的文本,最终选出文本600篇。将选取的文本转换为统一的文本格式。然后将所有语料分为两部分,其中350篇作为训练语料,其余250篇作为测试语料集合 T 。手工对测试语料集中的所有文本提取情感主题句,并标注情感倾向(正面或者反面),同时标注每个文本的全文情感倾向。标注结果中132篇为反面倾向文本,118篇为正面倾向的文本。

4.2 实验分析

在实验1中,性能指标的选择方面,选择两套不同指标。一是采用传统的评价指标:查全率(Recall)、查准率(Precision)以及 F_1 值来刻画性能的优劣。 F_1 为查全率和查准率的函数,是评价性能的综合指标。

为了考查CRF标注结果的总正确率,采用宏平均和微平均作为衡量指标。宏平均 P_{mi} 即求两类情感倾向(正面和反面)准确率的平均值,微平均 P_{mi} 则是用所有标注正确的句子除以句子的总数:

$$P_{mi} = \frac{1}{n} \sum_{k=1}^n f(k) \quad (6)$$

其中 n 是文本中所有句子的总数,如果第 k 句标注正确,那么 $f(k)$ 的值为1,否则取0。

随机选取测试语料库中的20%、50%形成两个新的测试集 T_1 和 T_2 ,在这两个测试集以及全部测试语料上采用本文提出的方法分别进行三次实验。表1为各次实验结果。

表1 本文方法实验结果 %

| 实验测试集 | 准确率 | 召回率 | F_1 |
|-------|------|------|-------|
| T_1 | 73.1 | 75.4 | 74.2 |
| T_2 | 74.4 | 76.5 | 75.4 |
| T | 74.3 | 77.8 | 76.0 |
| 合计 | 73.9 | 76.6 | 75.2 |

表1中的结果显示,各次实验的平均准确率达到73.9%,平均召回率达到76.6%,同时 F_1 值达到了75.2%。

表2为三次实验的宏平均和微平均结果。

表2 CRF模型标注结果 %

| 实验测试集 | 宏平均 | 微平均 |
|-------|------|------|
| T_1 | 67.3 | 69.1 |
| T_2 | 67.7 | 70.1 |
| T | 67.5 | 69.8 |

实验2目的是评估本文方法对文本情感分析判断的影响。

首先采用本文方法提取文本的情感主题句,将提取的结果应用到传统SVM分类器,对文本的情感进行分析。同时,采用传统SVM分类器直接对文本进行情感分析,将这两种方法的情感分析结果进行比较。

采用的评价指标依然是查全率、查准率以及 F_1 值。图1为两种方法的实验结果的比较,其中R、P、F分别代表上述三个评价指标。

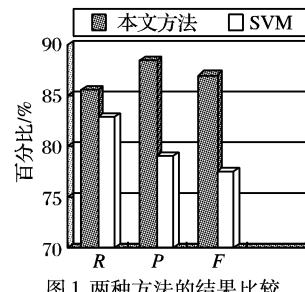


图1 两种方法的结果比较

从图1可以看出,采用本文方法提取情感主题句对文本情感进行分析,有效地提高了判断的准确率和 F_1 。与传统SVM分类器相比,基于本文方法进行情感分析,性能提高明显,其中准确率提高了近10%, F_1 值提高了11%。

由于本文方法首先去除了与主题无关的句子,排除了这些句子对整个文本情感判断的影响,直接对与主题相关的情感句子进行判断,因此提高了情感判断的准确率。

5 结语

本文提出了中文文本情感主题句的提取方法。通过评估参数确定文本主题概念,结合句子重要度计算,确定文本的主题句子。采用条件随机场模型,选取情感倾向和转移词两类特征,从文本主题句子集合中提取文本的情感主题句。情感主题句的提取研究对意见挖掘、情感分析等研究具有重要意义。在后继的研究中,将进一步完善本文的研究,选取语义特征,提高模型的提取性能。

(下转第1176页)

$$\rho(b, b') = \frac{\sum_{j=0}^{l-1} b_j b'_j}{\sqrt{\sum_{i=0}^{l-1} b_i^2} \sqrt{\sum_{i=0}^{l-1} b_i'^2}} \quad (7)$$

其中, ρ 的取值在 0 至 1 之间。 $\rho = 1$ 代表嵌入的水印信息全部被正确检测出来。 ρ 的值越大, 受攻击后 ρ 的变化越小, 代表水印的鲁棒性越强。

表 1 水印前后平均码率变化 kbps

| QP | 水印前视频码率 | 水印后视频码率 | 码率相对变化率/% |
|----|---------|---------|-----------|
| 28 | 417.45 | 420.68 | 0.8 |
| 36 | 142.54 | 142.84 | 0.2 |

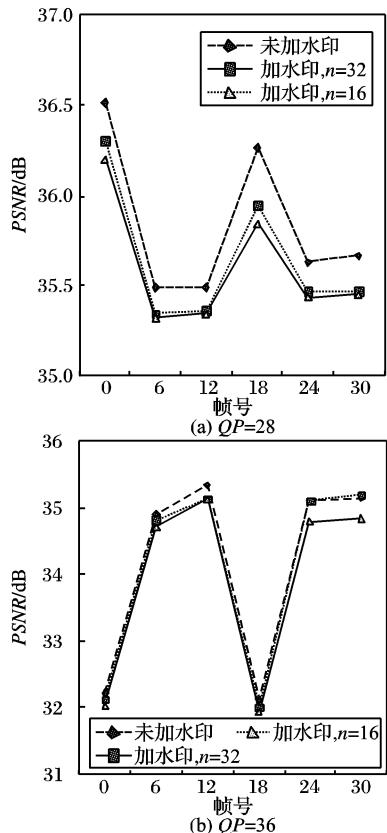


图 2 $QP = 28, 36$ 时嵌入水印前后视频质量 ($R_D = 0.002$)

(上接第 1173 页)

参考文献:

表 2 为量化参数 $QP = 28$ 和 36 的 foreman 水印序列重编码提取水印相似度。由表 2 可见, 经重量化编码后, 从 I 帧提取的水印相似度很大, 可以说得到了较完整的水印提取, 即说明水印对重量化攻击具有一定的鲁棒性。

表 2 foreman 水印序列重编码提取水印相似度 (ρ)

| QP | I 帧帧号 | | | | | | | |
|----|-------|---|------|----|----|----|------|------|
| | 0 | 4 | 8 | 12 | 16 | 20 | 24 | 28 |
| 28 | 1 | 1 | 1.00 | 1 | 1 | 1 | 1.00 | 1.00 |
| 36 | 1 | 1 | 0.98 | 1 | 1 | 1 | 0.96 | 0.99 |

4 结语

本文将 DEW 算法引入 H.264 视频内容认证和版权保护的研究中, 在文献 [5] 所提出的能量差比率方法基础上对其进行改进, 对图像能量分布是否满足水印位的嵌入做出判别, 提高了嵌入的有效性和水印的鲁棒性。同时给出了视觉失真、视频码率变化和水印鲁棒性等方面的结果。实验结果表明本文所提出的水印算法具有较高的保真度与鲁棒性。

参考文献:

- [1] ALATTAR A M, LIN E T, CELIK M U. Digital watermarking of low bit-rate advanced simple profile MPEG-4 compressed video[J]. IEEE Transactions Circuits Systems Video Technology, 2003, 13(8): 787–800.
- [2] ZHANG J, HO A T S. Robust digital image-in-video watermarking for the emerging H.264/AVC standard[C]// Proceedings of the 2005 IEEE Workshop on Signal Processing Systems Design and Implementation: SIPS 2005. [S. l.]: IEEE Computer Society, 2005: 657–662.
- [3] YIN P, YU H H. Semi-fragile watermarking system for MPEG video authentication[C]// Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing: ICASSP'02. Orlando: IEEE Computer Society, 2002: 3461–3464.
- [4] LANGELAAR G C, LAGENDIJK R L. Optimal differential energy watermarking of DCT encoded images and video[J]. IEEE Transactions on Image processing, 2001, 10(1): 148–158.
- [5] 凌贺飞, 卢正鼎, 邹复好. 基于 MPEG 的实时视频水印技术[J]. 小型微型计算机系统, 2005, 26(12): 2181–2185.

- [5] HATZIVASSILOGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives[C]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics: ACL-97. Geneva, Switzerland: [s. n.], 1997: 174–181.
- [6] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 21st International Conference on Machine Learning. Tokyo: [s. n.], 2001: 282–289.
- [7] SHA F, PEREIRA F. Shallow parsing with conditional random fields[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown, NJ, USA: Association for Computational Linguistics, 2003: 213–220.
- [8] PHAN X H, NGUYEN L M, NGUYEN C T. FlexCRFs: Flexible conditional random field toolkit[EB/OL]. [2008-08-20]. <http://flexCRF.sourceforge.net>.