文章编号:1000-5862(2015)06-0642-05

面向新闻的情感关键句抽取与判定

罗文兵,徐雄飞,王明文*,左家莉
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要: 情感倾向的分析已经成为当前研究的热点. 面向新闻的情感关键句抽取与判定主要运用的技术有对文本进行预处理、计算文本中词项权重、提取情感关键句、用 SVM 分类器对情感关键句进行情感倾向性分析. 实验结果表明: JXNUIIP 系统对情感关键句提取问题有良好的效果.

关键词: 新闻; 情感关键句; 情感倾向性分析; 支持向量机
中图分类号: TP 311　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2015.06.18

0 引言

随着互联网的高速发展,信息技术已经运用在人们生活中的各个方面. 在这个信息爆炸时代,网络的信息量已经达到人们无法想象的程度. 因此,面对海量的新闻,自动提取新闻的关键句[1-3],以及判断这些句子的情感[4-7],有助于人们快速地找到自己感兴趣的新闻,同时也是舆情热点自动发现、新闻文档自动分类、个性化检索等领域的基础.

在(COAE2014)测评的任务 1 中提出一种新的提取新闻情感关键句的方法,首先需要对文本语义概念的归纳能力进行评估,确定词项权重. 提交结果中词项的权重由用 NLPIR 汉语分词系统中的功能确定,再对系统进行改进,采用 TW-IDF[1] 的方法计算词项权重. 通过词项权重来计算这些候选关键句的权重,选出最终的关键句. 考察情感词典来评估关键句的情感强弱,计算出关键句的情感强度,经过综合考虑关键句词项权重和情感权重选取出最终的情感关键句. 采用 LIBSVM 分类器对 NLP&CC2013 微博情感识别训练语料来分析新闻情感关键句和微博情感倾向.

1 情感关键句的提取

情感关键句既要包含文本的主题,又要有比较强的情感倾向. 因此,在提取情感关键句时首先需要计算文本的词项权重,根据句子中的词项权重选取出排名靠前的关键句作为候选关键句,再从中选取情感关键句.

1.1 关键句的提取

关键句涵盖了整个文本的中心思想,是阐明文本观点的句子,也是文本内容的集中体现. 而判断一个句子是否是关键句,要由句子中的词项重要程度决定. 因此,提取文本关键句首先需要计算词项的权重. 根据句子中出现的词项权重之和确定句子的权重,最终确定文本的关键句. 文本提取关键句的具体流程图如图 1 所示.

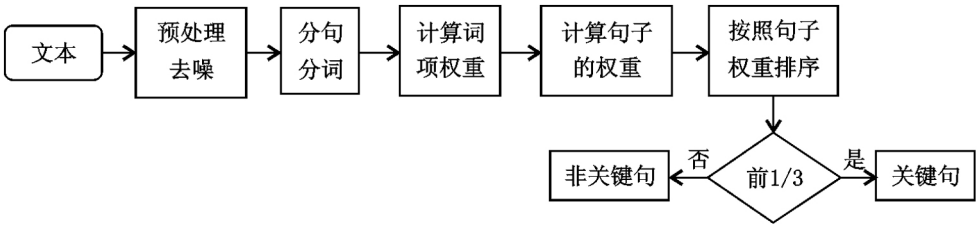

图 1　关键句提取流程图

收稿日期: 2015-07-09
基金项目: 国家自然科学基金(61272212, 61163006, 61203313, 61365002, 61462045)资助项目.
通信作者: 王明文(1964-),男,江西南康人,教授,博士生导师,主要从事信息检索、数据挖掘和并行计算的研究.

COAE2014 测评任务 1 提供的 10 000 篇语料中有包括 2 767 篇文本来自博客,2 589 篇文本来自论坛,4 641 篇文本来自新闻,还有 3 篇格式不正确的文本. 这些文本已采用规整的标签进行标注. 因此,首先需要对文本进行预处理,提取所需要的正文信息(在计算词项权重的时候,将标题中包含的词也考虑进来,这样可以加大标题中出现词的权重,更符合实际情况) . 预处理之后,对结果进行分词. 分采用 NLPIR 汉语分词系统(又名 ICTCLAS2013) 作为词工具,主要功能包括中文分词、词性标注、命名实体识别、用户词典等功能,支持 GBK 编码、UTF8 编码、BIG5 编码. JXNUIIP 系统采用了分词,去除停用词功能. 实验中,句子词项的权重由 NLPIR 系统中的关键词提取功能确定,此时再采用 TW-IDF 的方式改进了句子词项的权重的计算方法. 具体的计算方法为

$$TW-IDF(t,d) = (1+\ln(1+\ln(tw(t,d))))/(1-b+b\times|d|/avdl)\times\log(N+1)/(df(t)), \quad (1)$$

其中 b 为经验参数,默认值为 0.04; $tw(t,d)$ 为词项图中的词项结点的入度数; $df(t)$ 为词项 t 出现的文档数; N 为全部数据集中包括的文档数.

通过(1) 式可以得到数据集的文档词项矩阵. 根据文档词项矩阵,可以得到每篇文本出现的词项以及词项的权重,并以此计算文本中句子的权重. 目前已提出了多种有关计算句子权重的方法. 本文在此基础上做了一定的改进. 定义所有候选的句子 S,可表示为 $S(T_1:W_1,T_2:W_2,\cdots,T_n:W_n)$,其中 T_i 表示某个词, W_i 表示该词对应的权重. 建立出文本中句子的模型,并以此计算出句子的权重,从而提取出排名靠前的句子作为关键句. 句子 S 的权重的计算公式为

$$I(s) = \lambda(n\sum_{i=1}^{n}W_i)/N, \quad (2)$$

其中 λ 为句子 S 的位置加权系数,不同位置的句子加权系数不同,根据作者的写作习惯,通常会在篇章开题阐明主题或在文章结束总结中心思想. 位于首尾的句子通常比较重要,因此这两个位置的句子加权系数也比较高. 本系统将文本中前两句或后两句的加权系数设置为 3,其他句子加权系数为 1. N 为句子中所有词的个数, n 为句子中包含词项的个数. 其中, n/N 是对句子长度进行归一化. 为了和句子的情感权重统一量纲,并和句子的情感权重进行求和运算,对句子的关键程度得分进一步归一化,具体为

$$W(s) = I(s)/I(s)_{\max}, \quad (3)$$

其中 $I(s)_{\max}$ 是文本所有句子中最大的 $I(s)$ 权重. 通过(3) 式得出句子的权重,按 $W(s)$ 大小排序,提取出排名在前 1/3 的句子作为文本的关键句,得到关键句集合 H.

1.2　情感关键句的提取

根据上面的步骤,得到关键句集合 H,接着从 H 中选取情感较强的句子,作为情感关键句. 具体过程包括: (i) 对 H 中的句子去除停用词、分词; (ii) 计算句子中所有情感词的权重之和,即为关键句的情感权重; (iii) 结合关键句的词项权重和情感权重作为句子的综合权重,按照综合权重排序,选取前两句作为文本的情感关键句(有些文本只有一句关键句,则直接选定为情感关键句) . 具体的流程图如图 2 所示.

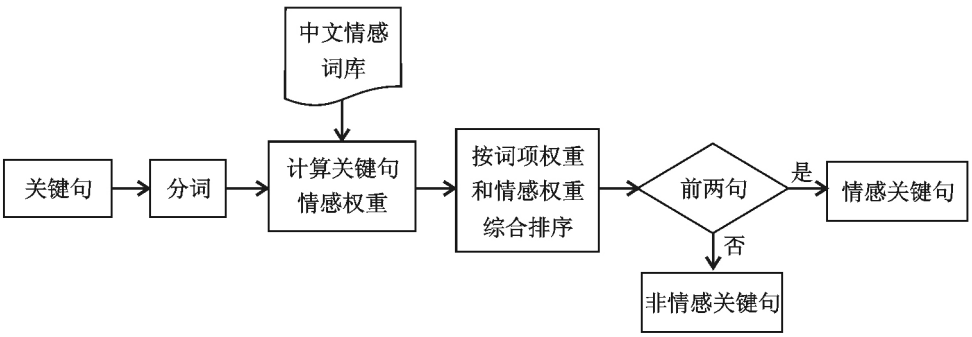

图 2　情感关键句提取流程图

本文用到的情感词典为大连理工大学信息检索研究室提供的情感词汇本体库[4]. 该词典提供了 27 467个情感词,其中包括词的极性,每种极性均分为 1、3、5、7、9 共 5 种强度等级,其中 1 代表极性最弱 9 代表极性最强. 国内外有较多关于情感关键句的研究,本文对该工作进行改进,假设句子的情感倾向性强度与情感词的情感倾向性强度、情感词数目、句子向量维数有关. 此时句子的情感倾向性强度仍

依赖于情感倾向性强度最大的情感词,但句子向量维数影响其依赖性. 经过大量实验,置句子中情感倾向性强度最大的情感词和句中所有情感词的权重为 $\mu_1 = \mu_2 = 1/2$,句子的情感倾向程度较大一部分取决于词语的情感倾向性强度,若句子向量维数一样,则包含的情感词个数越多其情感倾向强度就越大;若句子向量维数不一样,包含的情感词个数一样,则句子的情感倾向强度不一样. 因此,需要对句子长度进行 n/N 的归一化,其中 n 为句子的倾向性词的个数,N 为句子向量的维数. 则句子的情感倾向性强度 bel 为

$$bel = \frac{n}{N}\left|\mu_1 \max_{i=1,\cdots,N}\{w_i\} + \mu_2 \sum_{i=1}^{N} w_i\right|, \quad (4)$$

其中 w_i 为句子中词语的倾向性强度,即倾向性得分. 计算得到的 bel 值,即为句子的情感强度得分.

$$con = \begin{cases} 1, & bel > \pi^2, \\ \sin\dfrac{\sqrt{bel}}{2}, & bel \leqslant \pi^2. \end{cases} \quad (5)$$

为了与(3)式处于同一个量纲,将(5)式对(4)式进行归一化,可以对结果进行求和操作. 情感关键句的权重既要考虑词项权重也要考虑情感权重的得分. 具体计算公式为

$$Select(s) = \alpha \times W(s) + \beta \times con, \quad (6)$$

其中 α 和 β 是两个经验参数,实验中,分别取值 0.7 和 0.3 时结果最优. 最后,按照 $Select(s)$ 的得分对候选关键句排序,选取前两句(有些文章只有一句)作为情感关键句.

2 基于 SVM 分类器的情感关键句倾向分析

完成以上工作后,从 100 000 篇语料中选取出 19 666个情感关键句,除了有些较短的文本只提取一句情感关键句外,大部分文本都保留了两句情感关键句. 接下来,分析所提取出情感关键句的情感倾向. 本文采用的是 LIBSVM 分类器来实现情感倾向性分析.

2.1 情感词典特征提取

在进行 SVM 分类之前,需要提取文本的情感特征向量,这里提取的情感特征包括:(i)所有词语的正向得分的乘积;(ii)所有词语的负向得分的乘积;(iii)所有词语的正向得分的最大值;(iv)所有词语的负向得分的最大值. 情感词典特征的最大值考虑了最具情感倾向的词语产生的影响,而乘积考虑了所有词的整体影响.

2.2 构造 SVM 情感分类器

本文使用 2 级二分类策略,第 1 级为正向—非正向分类,第 2 级为负向—非负向分类: 蒋飞等[15]在 COAE2013 评测报告论文集提到这种方法,具体如表 1 所示,本文实验中进行了相应的调整.

表 1 判断情感分类方法

句子情感	第 1 级分类结果	第 2 级分类结果
正向	正向	非负向
负向	非正向	负向
中性	非正、非负、正向、负向	非正、非负、正向、负向

表 1 给出了情感关键句情感倾向的判断依据,将语料情感分为正向、非正向; 负向、非负向. 用 NLP&CC 多情绪分类任务提供的数据映射到正、负、中共 3 种类别后进行训练,利用训练得到的模型对任务 1 和任务 4 的语料进行分类. 使用 LIBSVM,作为 SVM 工具包,该工具包支持概率估计.

3 测评结果与分析

3.1 测评结果

依据文中给定的方法,实现 JXNUIIP 系统,并参加了 COAE2014 任务 1. 具体测评结果如表 2 所示(任务 1 的原始结果小数点后保留了 7 位有效数字,本文仅保留 3 位有效数字).

表 2 中显示了 10 个参赛队伍的结果,呈现了各系统提取正向、负向情感的 P、R、F1 值,Accuracy(正确率)及微平均的 P、R、F1 值. 从数据中可以看出 IIP 系统的 PosR 值和 JXNUIIP 的一致,PosP、PosF1 值高于 JXNUIIP,说明 IIP 系统在正向情感关键句提取方面比较有效; DUTIR1 系统的 NegP 值比 JXNUIIP 高一些,BUPT 系统在 MicroP 值上高于 JXNUIIP. 其它的值均是 JXNUIIP 系统最高. 综上所述,本文实现的 JXNUIIP 系统在面向新闻的情感关键句抽取与判定方面取得了良好的结果.

任务 1 在提取情感关键句实验中,调整句子位置加权系数 λ 对测评结果影响比较大,具体系数 λ 调整对系统结果影响在下一节中详细讨论.

表 2　任务 1 评测结果

系统	PosR	PosP	PosF1	NegR	NegP	NegF1	Accuracy	MicroR	MicroP	MicroF1
JXNUIIP	0.189	0.049	0.078	0.177	0.074	0.104	0.060	0.322	0.088	0.138
AHU	0.003	0.010	0.004	0.032	0.040	0.036	0.034	0.050	0.067	0.057
XXX	0.088	0.022	0.035	0.003	0.015	0.005	0.021	0.111	0.059	0.077
IIP	0.189	0.052	0.082	0.139	0.067	0.090	0.058	0.304	0.081	0.128
PRIS	0.125	0.028	0.045	0.046	0.031	0.037	0.029	0.241	0.064	0.102
BUPT	0.158	0.038	0.062	0.036	0.052	0.043	0.041	0.188	0.100	0.131
DUTIR1	0.077	0.039	0.052	0.074	0.084	0.078	0.056	0.154	0.082	0.107
IRLab	0.042	0.022	0.030	0.008	0.018	0.011	0.021	0.129	0.069	0.090
zutnlp	0.025	0.017	0.021	0.061	0.024	0.034	0.022	0.092	0.052	0.067
hut	0.077	0.030	0.043	0.059	0.044	0.050	0.036	0.092	0.049	0.064

3.2　句子位置加权系数 λ 对测评结果的影响

在计算句子词项权重时，考虑不同位置的句子权重应该不一样. 位于文本前两句和后两句的加权系数 λ 调整如图 3 所示.

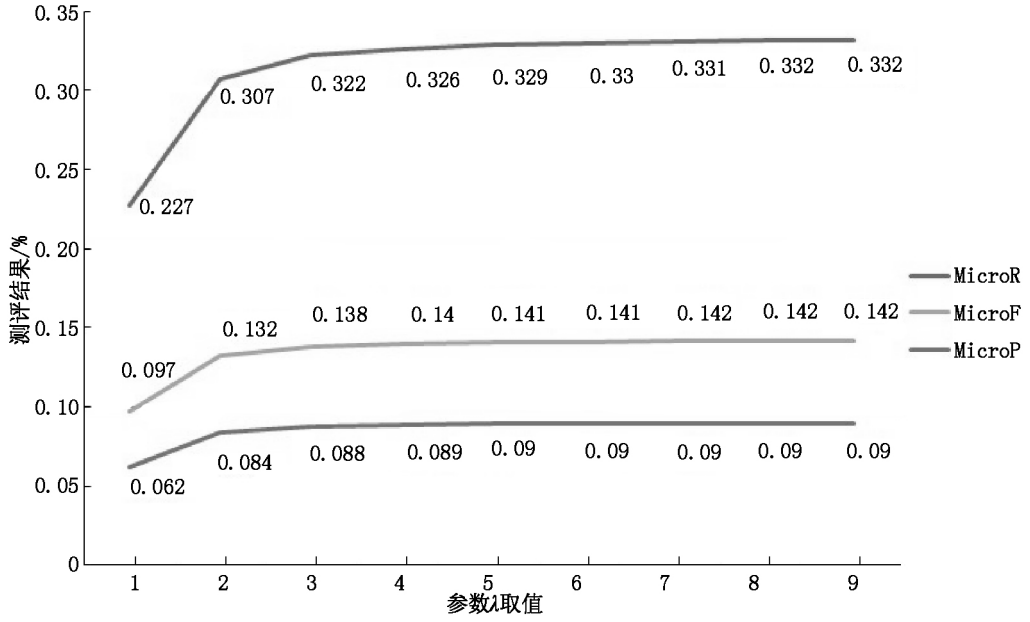

图 3　参数 λ 对结果的影响

从图 3 中发现，随着参数 λ 的增长，MicroR、MicroP、MicroF1 先增长，后趋向稳定. 说明适当的增加首尾句的加权系数，可以获得比较好的结果. 本次改进的测评系统 JXNUIIP 中 λ 的取值为 3(并没有选取最好结果的参数值，这样比较符合实际情况) .

4　总结与展望

本文针对 COAE2014 中任务 1(面向新闻的情感关键句抽取与判定) 的实现过程给出了较为详细的解决方案. 从提取关键句、提取情感关键句、情感倾向分类等方面展开. 从参赛结果数据来看，JXNUIIP 还是比较理想的，但还存在一些不足. 今后主要可以从以下几方面来改进: (i) 构建更加符合任务的情感词典; (ii) 优化判断句子情感倾向性的方法; (iii) 收集并建立适合新闻情感关键句倾向判定的语料等.

5　参考文献

[1] Rousseau F，Vazirgiannis M. Graph-of-word and TW-IDF: new approach to ad hoc IR [C] . New York: ACM，2013: 59-68.

[2] 樊娜，蔡皖东，赵煜，等. 中文文本情感主题句分析与提取研究 [J] . 计算机应用，2009，29(4) : 1171-1173.

[3] 周文，张书卿，欧阳纯萍，等. 基于情感依存元组的新闻文本主题情感分析 [J] . 山东大学学报: 理学版，2014，49(12) : 1-6.

[4] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造 [J] . 情报学报，2008，27(2) : 180-185.

[5] 李艺红，蒋秀凤. 中文句子倾向性分析 [J] . 福州大学

学报:自然科学版,2010(4):504-508.
[6] Kherwa P,Sachdeva A,Mahajan D,et al. An approach towards comprehensive sentimental data analysis and opinion mining [EB/OL]. [2014-10-16]. 10. 1109/IAdCC. 2014. 6779394.
[7] Pang Bo,Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts [EB/OL]. [2014-10-23]. 10. 3115/1218955. 1218990.
[8] 黄萱菁,张奇,吴苑斌. 文本情感倾向分析 [J]. 中文信息学报,2012,25(6):118-126.
[9] 杜振雷,张仰森,李文坤,等. 基于多特征融合的中文微博情感分类方法研究 [C]. 第五届中文倾向性分析评测研讨会,2013:44-49.
[10] 罗凌,陈毅东,曹茂元. 微博观点句识别的话题影响研究 [J]. 电脑知识与技术:学术交流,2014,10(1):123-127.
[11] 朱艳辉,杜锐,鲁琳,等. 中文文本情感分析与比较句的识别研究 [C]. 第五届中文倾向性分析评测研讨会,2013:34-43.
[12] 周胜臣,瞿文婷,石英子,等. 中文微博情感分析研究综述 [J]. 计算机应用与软件,2013,30(3):161-164.
[13] 刘志广,董喜双,关毅. 中文微博情感倾向性研究 [C]. 第五届中文倾向性分析评测研讨会,2013:81-87
[14] 朱艳辉,杜锐,鲁琳,等 中文文本情感分析与比较句的识别研究 [C]. 第五届中文倾向性分析评测研讨会,2013:34-43.
[15] 蒋飞,刘奕群,张敏,等. THUIR-SENTI:COAE2013 测评报告 [EB/OL]. [2013-10-17]. http://wenku. 55. la/P-93139. html.
[16] 万韩永,左家莉,万剑怡,等. 基于样本重要性原理的 KNN 文本分类算法 [J]. 江西师范大学学报:自然科学版,2015,39(3):297-302.
[17] 徐雄飞,徐凡,王明文,等. 中文微博句子倾向性分类中特征抽取研究 [J]. 江西师范大学学报:自然科学版,2015,39(3):290-296.

The Sentiment Key Sentence Extraction and Identification for News

LUO Wenbing,XU Xiongfei,WANG Mingwen*,ZUO Jiali
(School of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Sentimental opinion analysis has become a research focus currently. Sentiment key sentence extraction and identification for news,firstly preprocessed of the text,calculated weight of the terms,extracted the sentiment key sentences,and then used the SVM classifier to analysis the sentimental opinion of the sentence. Experimental results validated that JXNUIIP system has preferable performance on extracting sentiment key sentence.

Key words: news; sentiment key sentences; sentimental opinion analysis; SVM

(责任编辑:冉小晓)