基于论坛主题的网页褒贬倾向性识别

王 爽,熊德兰,赵会洋
(许昌学院 计算机科学与技术学院,河南 许昌 461000)

摘 要:褒贬倾向性识别在信息过滤、自动文摘、文本分类等领域有良好的应用前景。针对褒贬倾向性较为集中的论坛网页,提出了基于特定论坛主题的网页文本褒贬倾向性计算方法。结合句法分析和词语相似度计算方法,提取反映主题倾向的特征词,根据每个信息块的倾向性计算页面的褒贬倾向,实现了论坛网页句子级别、信息块级别和网页级别等三个层次的褒贬倾向性计算,在部分语料范围内的实验结果良好,对于此类网页的分析评价有一定的意义。

关键词:论坛主题分析;倾向性识别;网页评价;词语相似度
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2009)09-0111-04

Appraisial Orientation Identification in WebPages Based on Forums Theme

WANG Shuang, XIONG De-lan, ZHAO Hui-yang
(Department of Computer Science and Technology, Xuchang University, Xuchang 461000, China)

Abstract: Orientation identification has good application in some fields such as information filtering, automatic summarizations, text classification and so on. Aiming at Web forums which mass appraisial orientation, present a method to calculate orientation of Webpages based on certain forums. Combining with syntax analysis and words similarity, extract trait words to reflect theme orientation, and compute the orientation of whole forums pages basing on each information block. And the calculating orientation of Web forums sentences rank, information block rank and page rank has been implemented. The computing of results of experiments at range of considerable Web data are good. It has some significance to the analysis and evaluation of this kind Webpage.

Key words: forums theme analysis; orientation identification; Webpages evaluation; words similarity

0 引 言

互联网的发展与普及深刻地改变了人们的生活和思维方式,极大地推动了全球信息化进程,网络已经成为当今人们获取知识、发布信息、交流沟通的主要工具。然而,面对纷繁复杂的网络世界,要准确快速地获取自己所需的信息也并非易事,正如 Rutherford D. Roger 所说的"我们被信息淹没,但却缺乏知识"。信息检索、信息过滤、自动文摘等信息处理技术成为解决该矛盾的主要方法。

情感态度可以认为是信息表达的要素之一,因为信息在传递过程中会或多或少带有信息创造者的情感倾向。根据信息表达方式的不同,情感态度的识别和研究方法有很大的不同,具有的应用领域也有很大的差别。目前研究较多的是图形图像和语音信号中情感特征的提取及识别,文本中词汇倾向性识别等[1~3]。文本倾向性识别是通过对文本格式的信息资料中提取具有一定感情色彩的词汇、短语、常用结构等,来判定文本的情感态度倾向。因而是较深层次的信息处理技术之一,可以在信息检索、信息过滤、信息安全、自动文摘、主动推送等诸多领域都有广泛的应用[4]。

网络论坛是通过计算机网络来传播和获取消息,是人们在网络环境下进行自由交流的良好平台。网络论坛反映了不同地域、不同背景的网民对某一主题的思想和看法,从而能够集中大多数人的智慧来鉴别某一人、物、事件或观念等的是非曲直。文中在对论坛格式网页文本分析的基础上,介绍了褒贬倾向性词汇的提取和识别方法,进而判定论坛主题的褒贬倾向性。

收稿日期:2009-01-05;**修回日期**:2009-03-26
基金项目:河南省自然科学研究计划项目(2008B520031);许昌学院2009年校内科研项目(2009056)
作者简介:王 爽(1982-),女,河南驻马店人,讲师,研究方向为网页分类。

1 论坛网页的分析与处理

1.1 网页预处理

网页与普通的文本不同,Web 页面呈现一种半结

构化和动态性的复杂形式,因而,网页文本倾向性分析需要进行一定预处理过程。

论坛网页是一种特定的网页形式,一般来说论坛网页中包含三个主要部分:

(1)网页格式信息、网站信息、管理者和用户操作链接、广告链接等;

(2)主题发贴者、主题回复者以及他们的名称、登陆或发布信息的时间、上网的地址等相关信息;

(3)关于发贴者的主题内容和回复者的会复内容。

根据文中研究需要,在处理过程中只关注第(3)部分,进行网页预处理时过滤无用标记和信息,只提取出当页面中论坛主题(即帖子主题)和每个回复者的回复内容对应的文本部分。

从语法形式上看,论坛 Web 页面包含多个信息单元,它们排列紧凑、风格相似,可以将每个回复内容看作为一个信息块,这样,就可以根据 HTML 文件标记建立页面分析树,确定信息块。

通过对页面分析树的观察发现论坛网页文件具有如下特征[5]:

1)Web 文件有多个信息块,这些信息块组成一个统一的整体,这个整体处于 Web 页面的某个离散区域之中,即某一子树中;

2)主题使用<TITLE>和</TITLE>标记,每个信息块则使用相似标记来描述,且多个信息块之间处于分析树的同一层次上。

根据上述分析可以对所建立的页面分析树进行修剪,提取特定 HTML 标记的文本作为论坛主题和回复内容。

1.2 褒贬倾向性度量思路

对网页的褒贬倾向性进行判定是在篇章级别上对单个文档的倾向性研究,根据自然语言理解处理方法,篇章的理解是建立在对段落的理解上,段落的理解归结于句子或句群的理解,句子的理解归结于词语的理解[6]。同样篇章的倾向性研究也是以词语倾向性为基础的,同时又综合考虑到句子内部语法结构、句子间关系和段落间联系。

根据论坛网页的特殊性,可以认为上述预处理后得到网页文本分割为若干个信息块,论坛主题称为基本信息块,一个信息块可以分解为若干个句子,一个句子由若干个褒贬特征词来表示。这样,文本褒贬倾向性就可以在以下几个层次上进行计算度量(如图 1 所示)。

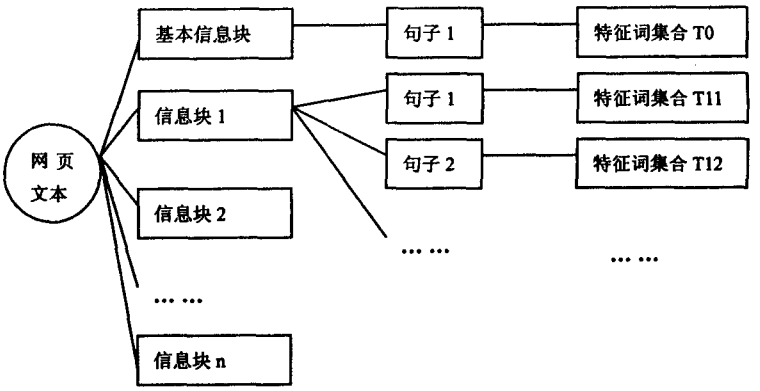

图 1 网页文本褒贬倾向性度量方法

2 褒贬倾向性计算

2.1 独立词语的倾向性

词语的语义倾向是指该词语在使用时所表现感情色彩或反映出说话者的情感态度,一般用正面褒扬和反面贬斥两种相对立的态度来表示,其褒扬或贬斥的程度可以用一定的数值来度量。目前比较常用的做法是将度量值规定为位于[-1,+1]之间的实数,其中正数表示为褒扬倾向,负数表示为贬斥倾向,数值的大小表示褒贬程度的强烈。

词语倾向性通常采用 HowNet 词语语义相似度计算机方法来衡量待定词语的倾向性。《知网》(HowNet)是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。HowNet 中词语相似度定义为两个词语在不同上下文中可以相互替换使用而不改变文本的句法语义结构的程度,其值为[0,1]之间的实数,数值的大小反映了可以替换的可能性大小[7]。

在文本倾向性研究中,可以先选取一定量具有强烈褒贬色彩的词语作为基准词,然后按照词语相似度计算方法计算待定词语的倾向性。

某个待定词语 W 的褒贬倾向性 Orientation(W)计算方法如下:

$$\text{Orientation}(W)=\sum_{i=1}^{kp}\text{Sim}(WP_i,W)-\sum_{j=1}^{kn}\text{Sim}(WN_j,W) \quad (1)$$

其中 WP_i 表示褒义基准词,WN_j 表示贬义基准词,kp 和 kn 分别表示选取的褒义和贬义基准词个数。

2.2 句法分析后词语倾向性的计算方法

《知网》中词语相似度侧重于计算两个概念之间的相似度,对于考察两个孤立的词语比较有效,但对于处于一定上下文中的两个词语就不够准确。因而同一句

中的两个词语的褒贬倾向性还要考虑词语所处的位置,以及它们在句中所作的成分。一般认为,词性相同、语法成分相同、语法分析树中层次相同的两个词语的相似性更大一些。这样就可以从句法分析结果和词语自身相似性两个方面来衡量词语的褒贬倾向性,具体工作流程如下:

首先从网页文本中选取少量的具有强烈褒贬色彩的词语作为基准评价词语。基准词通常根据知网中标注为"良"和"莠"的反义词对选取,但这样选取的词语与网页主题有一定的偏差,所计算的词语在该主题意境下的相似度也有失偏颇,因此粗略浏览网页,根据当前网页主题和风格,人工选取从网页中出现的具有褒贬倾向的少量词语做为基准词。

其次,对每个句子进行句法分析和词性标注,得到相应的句法分析树。考虑到具有褒贬色彩的词语通常是作为句子主干部分的名词、动词、形容词等,所选取的待定词为切分标注为普通名词、动词、形容词的词语。

然后,根据句法分析树和词语相似度来判定待定词语的褒贬倾向性,即待定词 W 与褒义基准词集的相似度 $\mathrm{Sim}(W, WP)$ 为:

$$\mathrm{Sim}(W, WP) = \sum_{i=1}^{kp} \frac{\mathrm{Sim}(WP_i, W)}{|D_i - D_w| + \alpha} \tag{2}$$

其中 D_i, D_w 分别为词语 WP_i 和待定词在句法分析树中的深度,α 为调解参数,防止分母为零时计算出错。

同样计算待定词 W 与贬义基准词集的相似度 $\mathrm{Sim}(W, WN)$,取二者的最大者作为该待定词语的褒贬倾向值,并将该词语添加到相应的基准词集合中。更新原有的褒贬基准词集合,重新计算其余待定词与新褒贬基准词集的相似度。

可以证明该算法在有限次迭代后收敛,且最后得到褒义和贬义词集合就是文本相应的褒贬特征词语集合。

2.3　网页文本倾向性

在经过上述处理后,得到网页文本的褒贬特征词语集,再次扫描每个句子,可以得到每个句子对应的褒贬特征词集,及其褒贬倾向性度量值。则句子的褒贬倾向性就可以根据它所出现的褒贬词语的频率和褒贬倾向性度量值的加权求和的方法来衡量。假定句子 S_i 经分析处理得到的褒贬特征词有 n 个,记为 $\{W_1, W_2, \cdots, W_n\}$,其褒贬倾向性分别为 $\{V_1, V_2, \cdots, V_n\}$,每个特征词出现的频率为 TF_i,则该句子的褒贬倾向性 $\mathrm{Orientation}(S_i)$ 为:

$$\mathrm{Orientation}(S_i) = \sum_{i=1}^{n} TF_i * V_i \tag{3}$$

一个信息块的褒贬倾向性为每个句子褒贬倾向性的累加,而文本的倾向性不仅要考虑每个信息块的褒贬倾向性,还要考虑信息块与基本信息块之间的相关性。因为网络论坛的自由性决定了回复内容可能会与主题无关、甚至是无用的广告链接,与主题无关或相关性比较低的回复应该对于主题的倾向性影响较小。与主题相关性可以采用一种简单的做法来实现,就是依据信息块与基本信息块中特征词集的相似性来确定。两个特征词语集合的相似性根据知网中集合的相似性计算方法来计算[8]。

于是,单个网页文本的褒贬倾向性就可以计算如下:

$$\mathrm{Orientation}(H) = \sum_{i=1}^{m} \mathrm{Sim}(B_0, B_i) * \mathrm{Orientation}(B_i) \tag{4}$$

3　实验与结果分析

采用 Visual C++ +6.0 编程环境对上述研究进行初步实验。实验选用的语料是从百度贴吧《贞观长歌》的部分影视评论中下载的原始网页,论坛主题共有 260 篇,其中帖子平均回复数约为 40 个。

实验程序分别对褒贬特征词语提取、句子褒贬倾向性、信息块褒贬倾向性和网页文本倾向性几个层次上进行褒贬倾向性的评定,采用人工检查评定的方法验证结果的有效性。

具体的实验过程下:

Step1:搜集论坛上一定量网页,进行预处理,得到主题和回复内容部分文本,形成初始语料;

Step2:对于每一个文本,做以下处理:

Step2.1:人工提取少量褒贬基准词,分别保存为集合 WP 和 WN;

Step2.2:对于每个句子,利用哈工大中文信息处理平台[9]实现分词标注和句法分析,提取标注为名词、形容词、动词的待定词,根据公式(2)计算待定词语的褒贬倾向性,并添加到 WP 或 WN 集合中;

Stip2.3:根据公式(3)和公式(4)分别计算每个句子和文本的褒贬倾向性;

Step3:输出提取后褒贬特征词语集合,给出当前网页文本的褒贬评价倾向性度量值。

从网页中提取的褒贬特征词,可以采用人工校对的方法检查所提取出来的词语及词语的褒贬极性。实验时,人工选取基准褒贬词语按照三种方法选取:第一组选取 5 个褒义词;第二组选取 5 个贬义词;第三组选取 5 个褒义词和 5 个贬义词,将三次选择的基准词分别作为三组输入数据进行实验验证。

将三次运行程序后得到特征词语集合进行人工校对,结果如表 1 所示。

表 1 三次实验提取的褒贬特征词语

测试方法	正确提取的词语数	错误提取的词语数	未提取的词语数
第一组	130	25	32
第二组	127	31	35
第三组	114	55	18

定义所提取褒贬词语的正确率=正确提取出来的褒贬义词语/总词语数,遗漏率=未提取出来的褒贬义词语/应有褒贬词语的总数,则上述三种方法提取褒贬特征评价词语的正确率分别为 0.839、0.799、0.675,漏检率分别为 0.171、0.181 和 0.096。

可见,只用一种倾向的褒义或贬义词语提取与其相同倾向的褒贬词语正确率较高,但漏检率较大,而使用褒义和贬义基准词能提高漏检率,但正确率有一定的下降。

在此基础上,使用褒贬倾向性计算方法对单个文本的倾向性进行计算,将计算结果与人工评价结果的数值进行比较,二者数值之差的绝对值为评价偏差。二次褒贬倾向性评价结果均为一个[-1,+1]之间的一个实数,所以评价偏差应为[0,2]之间的正数,数值越大,说明评价偏差越大。

在单个句子、信息块、网页文本三个层次上分别计算评价偏差,实验选取 100 个句子、200 个信息块和 250 个网页文本分别检查,得到评价偏差的算术平均值,结果如表 2 所示。

表 2 不同层次上的评价偏差

评价对象	句子	信息块	网页文本
评价偏差	0.275	0.450	0.322

由于褒贬词语在提取和相似度计算时较多利用了句法分析结果,因此使用所提取出的褒贬特征词语进行句子褒贬倾向性的度量较准确,评价偏差较小;信息块中多个句子间形成的递进、转折关系或讽刺、反诘等语气效果都有可能影响整个信息块的倾向性效果,所以信息块的评价偏差较单个句子要大些。而网页文本平衡了多个信息块的评价偏差,只考虑对于主题的综合褒贬倾向性,与人工评价结果差别较小。

4 结束语

网页褒贬倾向性识别是对自然语言理解、信息处理、Web 数据挖掘等研究领域的结合与拓展,它给出了从大量网络资源获取人们主观见解和思想认识的好方法,在一定层次上理解文本内容,获取文字背后的深层含义有着重要的启发意义。针对褒贬态度比较明确、争论观点比较集中的论坛型网页进行了褒贬倾向性的分析和识别研究,提出了人工选取基准评价特征词和机器提取特征词相结合的方法。在汉语句法分析树的基础上,根据词性标注和分析树的层次关系来提取和计算词语的褒贬相似性,利用网页信息块分割方法,实现了论坛网页句子级别、信息块级别和网页级别等三个层次的褒贬倾向性计算,对于此类网页的分析评价有一定的意义。由于论坛中语言的多样性、自由性,这给词语和句法分析都带来了多种可能的误差,因而目前的实验结果还不够理想。希望在进一步的工作中,对此类网页格式和用词模式做深入分析,进一步改进实验结果,目标是开发出具有评价或摘要功能的可用系统。

参考文献:

[1] Kamps J, Marx M, Mokken R J, et al. Using WordNet to measure semantic orientation of adjectives[C]//In: Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation. Lisbon: [s.n.], 2004: 1115-1118.

[2] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL. Association for Computational Linguistics. New Brunswick, N J: [s.n.], 1997: 174-181.

[3] Turney P, Littman M. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactionson Information Systems, 2003, 21(4): 315-346.

[4] 王治敏,朱学锋,俞士汶.基于现代汉语语法信息词典的词语情感评价研究[C]//In: Computational Linguistics and Chinese Language Processing. 台湾新竹:中华民国计算语言学学会, 2005: 581-592.

[5] 瞿有利,于 浩,徐国伟,等. Web 页面信息块的自动分割[J]. 中文信息学报, 2003(18): 6-13.

[6] 杨学兵,钱 蓉.语义检索系统中的查询语句扩展算法改进[J]. 计算机技术与发展, 2008, 18(12): 1-3.

[7] 刘 群,李素建.基于《知网》的词汇语义相似度计算[EB/OL]. 2003-03. http://www.keenage.com.

[8] 刘克非,王 红,王卫玲.基于语义相似度的 Web 服务发现研究[J]. 计算机技术与发展, 2007, 17(2): 4-6.

[9] The XTAG Research Group. 中文自然语言处理开放平台[EB/OL]. 2003-07. http://www.nlp.org.cn.